ISEIS

Journal of Environmental Informatics 26(1) 71-79 (2015)

Journal of Environmental Informatics

www.iseis.org/jei

Spatiotemporal Classification Analysis of Long-Term Environmental Monitoring Data in the Northern Part of Lake Taihu, China by Using a Self-Organizing Map

W. Li[1,*], H. T. Zhang[1,*], Y. Zhu[2], Z. W. Liang[2], B. He[3], M. Z. Hashmi[2], Z. L. Chen[1], and Y. S. Wang[4]

[1]*School of Environment, Tsinghua University, Beijing 100084, China*
[2]*College of Environmental and Resources Sciences, Zhejiang University, Hangzhou 310058, China*
[3]*Tourism College, Hainan University, Haikou 570228, China*
[4]*Beijing Guohuan Tsinghua Environmental Engineering Design & Research Institute, Beijing 100084, China*

Received 16 January 2013; revised 21 August 2014; accepted 15 September 2014; published online 14 August 2015

ABSTRACT. Characterizing the spatiotemporal patterns of water bodies is an important environmental issue in the management and protection of water resources. The primary objective of this study was to assess the spatiotemporal characteristics of environmental monitoring data from Lake Taihu to improve water pollution control practices. A methodologically systematic application of a self-organizing map (SOM) was utilized for data mining in the northern part of Lake Taihu, China. The monitoring data set contained 14 variables from eight monitoring stations during the period 2000-2006. The SOM classified the data set into 10 clusters displaying a markedly different pattern. We determined the spatiotemporal distribution of water quality based on the data frequency at each station monitored monthly in the study area. Based on the SOM analysis results, we suggest that the government should increase the number of monitoring points in the region. Given the relatively poor water quality in the region, unnecessary points should be decreased and different control measures should be implemented during different seasons. The results of this study could assist lake managers in developing suitable strategies and determining priorities for water pollution control and effective water resource management.

Keywords: spatiotemporal classification, self-organizing map, water quality

1. Introduction

Human activities in basins have increased the amount of pollutants discharged into rivers and eventually imported into lakes and reservoirs worldwide, thereby resulting in a substantial deterioration of water quality and degradation of water environments (Jin et al., 2011). Hence, preventing and controlling water pollution and regularly implementing monitoring programs, which provide water managers with the necessary information for water resource management in general and water quality management in particular, are essential (Zhou et al., 2007a; Khalil and Ouarda, 2009). One critical step to effectively control water pollution is the development of water quality monitoring programs that can adapt to environmental conditions and spatio-temporal patterns (Su et al., 2011). To control water pollution and protect water resources, the Chinese government has spared no effort to establish a number of environmental monitoring systems and implement various water quality monitoring programs for prevention policy making. However, such monitoring systems obtain a large amount of water quality data, including physical properties and nutrient, inorganic, and biological parameters, which are difficult to analyze and interpret because of the latent inter-relationships between parameters and monitoring stations (Shin and Fong, 19 99; Zhou et al., 2007b; Zhang et al., 2009). Therefore, extracting meaningful information from these data by using advanced mathematical methods is a fundamental requirement to interpret spatiotemporal patterns and mining useful information for water quality management.

Generally, the application of multivariable statistical methods is a valuable tool to obtain a better understanding and interpretation of complicated data sets. Canonical correlation analysis, cluster analysis (CA), discriminant analysis (DA), principal component analysis (PCA), factor analysis, absolute principle component score-multiple linear regression, and factor analysis-multiple regression analysis are the commonly accepted traditional multivariate methods used to evaluate spatiotemporal variations in environmental research (Lovchinov and Tsakovski, 2006; Zhou et al., 2007a; Omo-Irabor et al., 2008; Noori et al., 2012). In recent years, efforts have been made to involve more sophisticated approaches, such as self-organizing maps (SOM) (Tsakovski et al., 2010a, b; Jin et al., 2011; Oyana, 2009), in spatiotemporal classification, pollution pattern recognition, and modeling studies with surface water quality data sets or to compare SOM classification with more traditional multivariate statistical classification methods (Astel et al., 2007). These methods have already been used for

* Corresponding author. Tel.: +86 10 62792625; fax: +86 10 62792625.
E-mail address: leewei0329@126.com (W. Li); Zhanght@mail.tsinghua.edu.cn (H. T. Zhang).

ISSN: 1726-2135 print/1684-8799 online
 doi:10.3808/jei.201500297

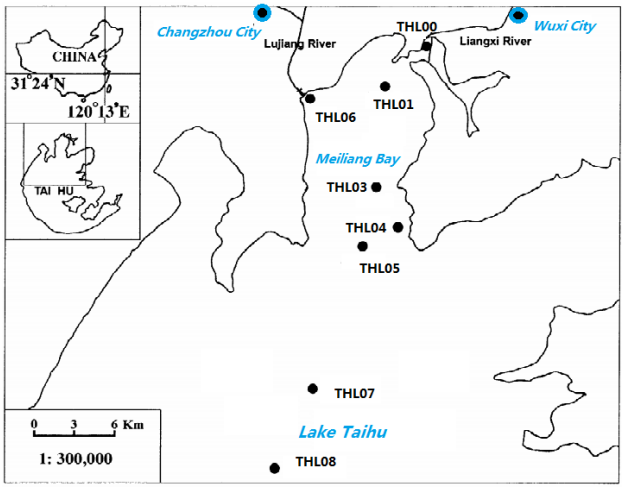

Figure 1. Location of monitoring points on the northern part of Lake Taihu, China.

surface water quality analyses. A previous work (Astel et al., 2007) has indicated that SOM can be used to reach a specific "resolution" of the proposed classification scheme compared with more traditional methods, such as CA. Although SOM is the standard method in environmetric studies (Chon, 2011), its application has not been fully examined in water quality studies in China.

The main purpose of the study is to evaluate the spatio-temporal patterns in water quality in the northern part of Lake Taihu, China and to demonstrate how more advanced SOM approaches could contribute to a better understanding of the data collected during monitoring episodes of a long period of observation. This study is the first to investigate the water quality in Lake Taihu based on the SOM approach. The results could be useful in water quality assessment.

2. Study Area and Data Set

Lake Taihu, with a surface area of approximately 2,400 km^2, is the third largest freshwater lake in the People's Republic of China in terms of area. Lake Taihu is located approximately 150 km west of Shanghai, Eastern China on the border of the Jiangsu and Zhejiang provinces, and the lake center coordinates are at 31°10′0″ N, 120°9′0″ E. The waters of the lake belong to Jiangsu province in its entirety, and part of its southern shore forming the boundary is between the two provinces with an area of 2,250 km^2 and an average depth of 2 m. The Taihu drainage basin is 36,500 km^2. The lake has more than 30 input sources, which range from rivers to small streams and manufactured drainage canals. Water exits the southeastern corner of Lake Taihu via the Taipu River, which drains through Shanghai into the East China Sea (Paerl et al., 2011).

Meiliang Bay is one of the most eutrophied bays in the northern part of Lake Taihu and is the site of recurring and intensifying Microcystis spp. blooms (Qin et al., 2007; Chen et al., 2003a, b). Eutrophication affects the multiple uses of Lake Taihu, including drinking water abstraction and fisheries. Thus, establishing monthly monitoring is important to find remedies against eutrophication by focusing on the effects on the northern part of Lake Taihu. The National Ecosystem Research Network of China (CNERN) Taihu Laboratory for Lake Ecosystem Research (TaiLLER) has established eight sampling stations covering the Meiliang Bay (inner and outer bays, including monitoring stations THL00, THL01, THL03, THL04, THL05, and THL06) and the lake center (main lake, including monitoring stations THL07 and THL08) (Figure 1).

Fourteen parameters from eight long-term positioning stations in Lake Taihu from CNERN TaiLLER in seven years (2000 ~ 2006) were used for analysis. The environmental monitoring data included water depth (WD, m), temperature (T, °C), Secchi depth (SD, m), suspended solids (SS, mg/L), dissolved oxygen (DO, mg/L), ammonium (NH_4^+-N , mg/L), nitrite (NO_2^--N , mg/L), nitrate (NO_3^--N , mg/L), total nitrogen (TN, mg/L), phosphate (PO_4^{3-} , mg/L), total phosphorus (TP, mg/L), sulfate (SO_4^{2-}, mg/L), chlorophyll a (Chl-a, µg/L), and pheophytin (Pheo, µg/L). The selected parameters WD, SD, and T were measured on-site, and the other parameters were measured in the laboratory of CNERN based on the environmental quality standard methods for the surface water of China (Wei et al., 2002) and the standard methods for observation and analysis in China (Huang et al., 1999). Table 1 shows the summary descriptive statistics of these parameters.

3. Methodology

3.1. Self-Organizing Map

The SOM was proposed by Teuvo Kohonen (Kohonen, 1982a,b; Kohonen and Makisara, 1989; Kohonen, 1990) and is used for the visualization and interpretation of large high-dimensional data sets. The SOM is a type of artificial neural network that can be trained using the unsupervised learning algorithm to produce a low-dimensional (typically two-dimensional), discretized representation of the input space of the training samples, called a map. The SOM is an automatic data analysis method (Kohonen, 2008) and is different from other artificial neural networks in the sense that it uses a neighborhood function to preserve the topological properties of the input space. The SOM also has the properties of vector quantization and vector projection algorithms (Vesanto, 2000). Currently, the SOM has been frequently applied as a powerful and effective data mining tool for the detection of data characteristics by pattern recognition, classification, and visualization onto two-dimensional arrays. Statistically, the SOM is utilized in exploratory analysis of large multivariate statistical data (Kohonen, 2008).

An SOM consists of neurons organized on a regular low-dimensional grid, and the number of neurons may range from a few dozen up to several thousands. The neurons are connected to adjacent neurons by a neighborhood relation, which determines the topology or structure of the map (Vesanto et al., 2000), and similar objects (in this case, sampling locations) should be mapped close together on the grid (Astel et al., 2008). The typical structure of an SOM consists of two layers, namely, an input layer, which classifies data based on their similarity, and a Kohonen map or output layer of neurons

Table 1. Basic Statistics (N = 672) (Monthly Averages) of Water Quality Variables in the Northern Part of Lake Taihu from January 2000 to December 2006

Variable Name	Abbreviation	Unit	Mean	Median	Minimum	Maximum	Standard Deviation
Water depth	WD	m	2.3	2.5	1	3.4	0.46
Temperature	T	°C	18	19	2.1	32	8.4
Secchi depth	SD	m	0.42	0.4	0	2.5	0.23
Suspended solids	SS	mg/L	51	43	3.4	230	36
Dissolved oxygen	DO	mg/L	8.2	8.6	0.47	15	2.8
Ammonium	NH_4^+-N	mg/L	1.5	0.31	0.002	20	2.2
Nitrite	NO_2^--N	mg/L	0.073	0.038	0.001	0.74	0.086
Nitrate	NO_3^--N	mg/L	0.93	0.79	0.001	4.3	0.69
Total nitrogen	TN	mg/L	4.3	3.5	0.39	14	2.8
Phosphate	PO_4^{3-}	mg/L	0.014	0.005	0	0.17	0.022
Total phosphorus	TP	mg/L	0.15	0.11	0.023	2.1	0.12
Sulfate	SO_4^{2-}	mg/L	79	74.9	31	210	28
Chlorophyll a	Chla	μg/L	20	11	0	520	33
Phaeophytin	Pheo	μg/L	4.8	3.55	0	42	4.9

arranged as a two-dimensional map. The input layer contains a neuron for each variable (e.g., T, DO) in the data set. The output layer neurons are connected to every neuron in the input layer through adjustable weights or network parameters. The weight vectors in the Kohonen layer provide a representation of the distribution of the input vectors in an ordered manner. The successive procedures required to apply the SOM can be divided into three categories, namely (Kalteh et al., 2008), (i) data gathering and normalization, (ii) training, and (iii) extracting information from the trained SOM. Post-processing of data sets is well-documented in several references (Kohonen and Somervuo, 2002; Vesanto, 2000; Kohonen, 20 01, 2003a,b), and the advantages of the SOM in relation to conventional ordination methods, such as PCA and CA, are discussed elsewhere (Astel et al., 2007).

Map quality is estimated by the correspondence between input data and trained map measured using quantization errors and topographic errors (TE). A low TE (close to 0) indicates that the SOM is good at preserving the topology.

For the total number of map units, a heuristic formula of $m = 5\sqrt{N}$ (where $\sqrt{N}$ is the number of data samples) is generally utilized (Vesanto et al., 2000). When the map size that was determined is large, the amount of detailed patterns that can be identified increases. However, the topographical proximity of clusters decreases. The heuristic rule can generate an optimized map size simultaneously considering the accuracy of pattern classification and topographical adjacency among clusters (Jin et al., 2011).

In the present study, we selected the nonhierarchical *k*-means (Wu et al., 2008) classification algorithm for clustering. Different values of *k* (predefined number of clusters) were used, and the sum of squares for each run was calculated. In addition, the SOM classification provides one more informative output, namely, the unified distance matrix (U-matrix) plane. The U-matrix plane is used to visualize the distance between the nodes in the grid and determine the aforementioned cluster structure of the map. The high values in the U-matrix plane imply a cluster border, and areas of low values indicate clusters themselves (Tsakovski et al., 2010b; Tobiszewski et al., 2010). Finally, we determined the best classifycation method based on the lowest Davies-Bouldin Index (DBI) (Davies and Bouldin, 1979; Jin et al., 2011). A detailed calculation of the DBI can found in the R document (Desgraupes, 2013).

In the present study, all calculations were conducted using MATLAB 6.5 running on Windows XP platform. To implement SOM-based classification, a free SOM toolbox 2.0 was utilized, which can be download together with documenttation (Vesanto et al., 2000) from the website http://www.cis.hut.fi/projects/somtoolbox/.

3.2. Data Set Preprocessing

To ensure that all variable parameters are given the same or similar importance, the monthly mean values must be transformed properly before the application of SOM. In particular, the results of the SOM application are highly sensitive to the data preprocessing method utilized because the SOM is trained so it can be organized based on the Euclidean distances between input data (Alvarez-Guerra et al., 2008). Generally, the five methods for standardization of data preprocessing are variance scaling (its mean to 0), range scaling into [0, 1], logarithmic transformation, logistic or soft-max normalization, discrete histogram equalization, and continuous equalization (Vesanto et al., 2000). In the present study, the skewness of the frequency distribution of each parameter was preliminarily analyzed by plotting histograms, as shown in Figure 2(a), with Pheo as an example. Logarithmic transformation was applied to decrease the positive skewness (see Figure 2(b)) of all parameters, except for DO, WD, and T, which did not have any clear skewness. Logarithmic transformation is used to smooth the data and decrease the influence of extreme values. Otherwise, the biased distribution may remain, causing inappropriate classification by SOM. Variance scaling was then conducted for all parameters so that the transformed data were distributed symmetrically with the same mean value and standard deviation, as shown in Figure 2(c).

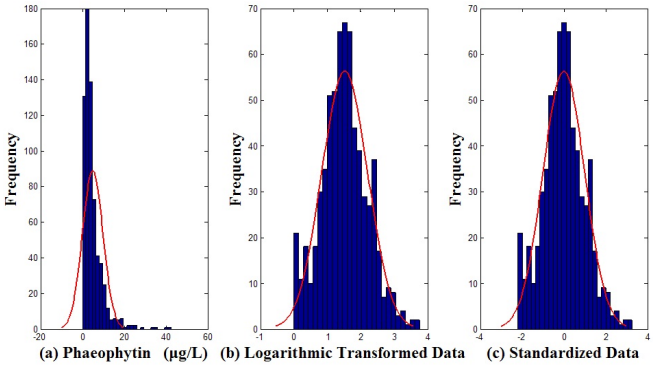

Figure 2. Histograms for (a) raw data, (b) logarithmic-transformed data, and (c) standardized data of phaeophytin.

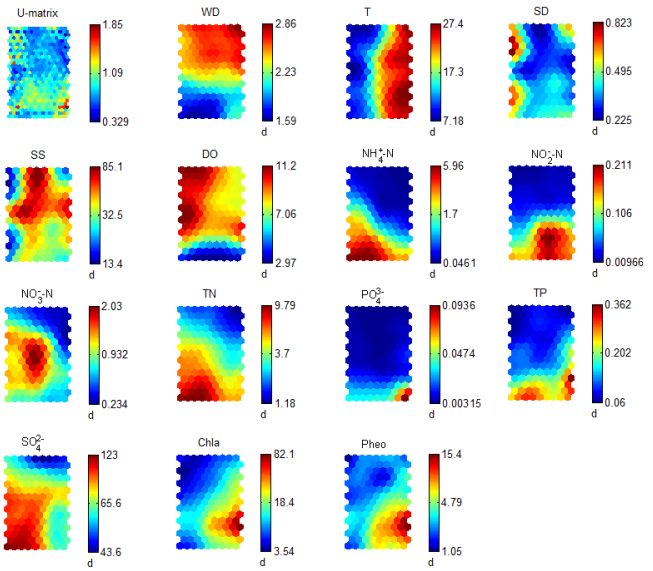

Figure 3. SOM visualization of the distribution of water quality parameters for all sampling stations (Note: The U-matrix is a representation of the SOM and is used to visualize the distances between neurons and to assist in determining and identifying the cluster structure of the map. High values of the U-matrix indicate a cluster border; uniform areas of low values indicate clusters themselves. Each component plane displays the values of one variable in each map unit. The color tone pattern and color bar labeled as "d" provide information regarding species abundance calculated through the SOM learning process, where "d" denotes denormalized data values on the color bar).

4. Results and Discussion

The data set utilized for exploratory analysis consists of 672 samples for all selected stations in Lake Taihu as each one is described by 14 variables derived on a monthly basis. Based on the methodology of SOM described previously, an SOM size of 135 ($\approx 5\sqrt{672}$) nodes (a hexagonal array with 15 nodes for a vertical direction and 9 nodes for a horizontal direction) was used for pattern classification of the standardized data set. The U-matrix and all variable planes for the input data set are shown in Figure 3. On the SOM map, the distribution of each variable and the distances between nodes in the U-matrix plane were determined using a color scale. For example, the objects with high NH_4^+-N , TN, and TP concentrations are located at the lower left part of the SOM plane, whereas the objects with high PO_4^{3-}, NO_2^--N , and TP concentrations are located mainly at the lower right part of the SOM plane.

Detecting the relationships between the variables observed for all stations and periods of monitoring is important. These relationships are shown in Figure 4. The location and distance of variables in the SOM as well as the analysis of color tone patterns provide semiquantitative information regarding the correlation coefficient. The order of variable planes showed six well-defined groups of correlated variables and several variables with specific location. The first group included the water quality parameters SO_4^{2-} and NO_3^--N . This fact is an indication of the similar information value of the two parameters. The second group revealed the connection between the SS and SD, which could be explained by the fact that SD is commonly influenced by SS. The third group included TN, NH_4^+-N , and WD, and a positive correlation exists between TN and NH_4^+-N and a negative correlation exists between NH_4^+-N and WD. The fourth well-defined group was formed by DO, PO_4^{3-}, TP, and NO_2^--N , and a positive correlation exists between PO_4^{3-}, TP, and NO_2^--N and a negative correlation exists between DO and NO_2^--N mainly because NO_2^--N is easily oxidized under the condition that DO is sufficient. The next group included the biological indicators Chl-a and Pheo, and a positive correlation was observed between them. The last variable, T, does not belong to any group and evidently possesses a more specific function in determining water quality. Table 2 quantitatively confirms the strength of the relationship between parameters by utilizing the standardized reference vectors. TN and NH_4^+-N showed the highest correlation coefficient of 0.84. The correlation coefficients between the remaining parameters were also easily determined.

In Figure 5, the clusters formed by the objects of observation (eight sampling stations for 672 episodes of monthly monitoring for seven years) are presented as an SOM. Based on the *k*-means clustering algorithm, the number of significant clusters (10 in this study) was determined by the lowest value of the DBI, as shown in Figure 5(a). Figure 5(b) shows the pattern classification map of the 10 clusters. The numbers of data classified into each node are also shown in Figures 5(c) and 5(d). Comprehensive consideration of the component planes (Figure 3) and the pattern classification results (Figure 5 (d)) determined the kind of data the respective clusters include. The exact content of the clusters for the sampling period is presented in Table 3. We conclude that the clusters are homogeneous based on the application of the Kolmogorov-Smirnov test of the difference between levels of quality indicators of water parameters for Clusters 1 to 10 determined using the SOM algorithm.

Cluster 1 (central part on the SOM of Figure 5(b)) contained primarily most sampling stations, except for THL00.

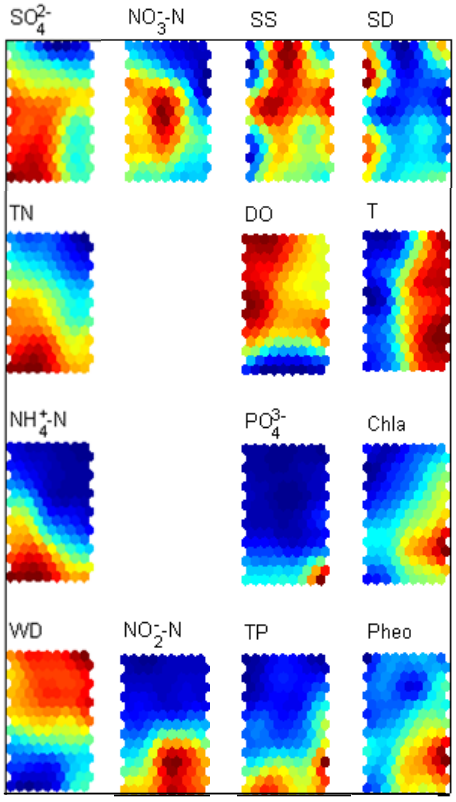

Figure 4. Water quality variable parameters similarity pattern was determined using the SOM approach (the distance between variables on the map, with analysis of color tone patterns, provides semiquantitative information regarding the nature of correlations between them).

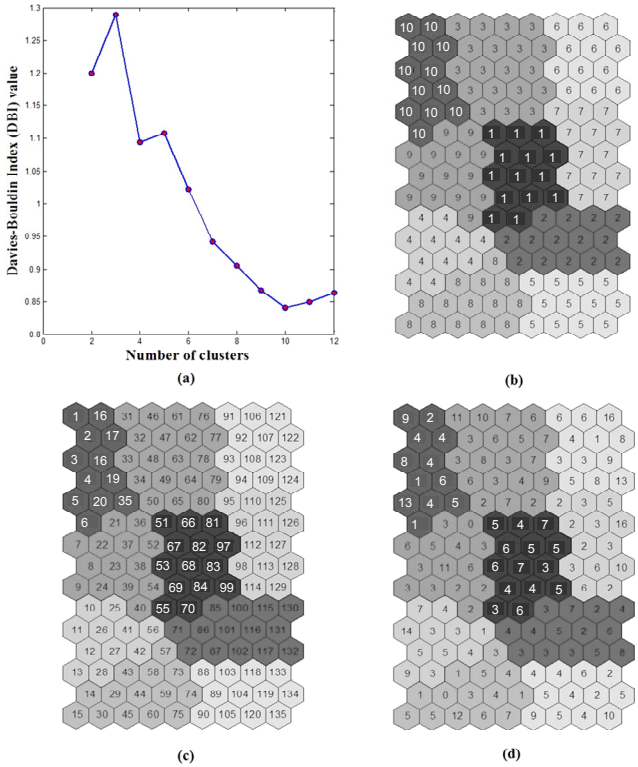

Figure 5. (a) SOM classification of all selected variables and clustering pattern according to Davies-Bouldin index minimum value. (b) both color scale hexagons in each SOM unit and digits represent the clusters; (c) both color scale hexagons in each SOM unit and digits represent node number belonging to particular clusters; and (d) both color scale hexagons in each SOM unit and digits represent the number of samples belonging to particular clusters.

Most water quality parameters were close to the averages of all data (e.g. physical indicators) with low values for NH_4^+-N and PO_4^{3-}. This pattern was observed in the same part of the respective component planes for each parameter, as shown in Figure 3. The worst water quality condition with extremely high chemical parameters (e.g., NO_2^--N , NO_3^--N , TN, PO_4^{3-}, TP, and SO_4^{2-}) and significantly low DO located at the bottom part of each component plane, as shown in Figure 3, is associated with Clusters 4, 5, and 8, as shown in Figure 5(b). By contrast, the better water quality condition with low chemical parameters and biological parameters (e.g., Chl-a) situated at the upper right of each component plane, as shown in Figure 2, was associated with Cluster 6, as shown in Figure 5 (b).

In addition, more quantitative information can be extracted and interpreted from the obtained reference vectors than the visualized pattern classification. The 25th percentile, median, and 75th percentile for the respective clusters were calculated using the standardized reference vectors to numerically characterize the classified data. For example, the quartiles for Cluster 1 were calculated using the standardized reference vectors of the 14 nodes classified into the cluster.

Figure 6 shows the radar graph of the 14 parameters for the 10 clusters with the 25th percentile, median, and 75th percentile plotted. In the case, where WD and T were not considered, the most ideal water quality condition can be defined as a value of 0 for SS, NH_4^+-N , NO_2^--N , NO_3^--N , TN, PO_4^{3-}, TP, SO_4^{2-}, Chl-a, and Pheo and a value of 1 for DO, as shown in Figure 6.

The visible patterns of Cluster 6 (Figure 6(f)), Cluster 1 (Figure 6(a)), Cluster 2 (Figure 6(b)), and Cluster 7 (Figure 6(g)) were similar, as shown in the figures. The pattern with superior physical parameter values but low chemical parameters associated with Cluster 6, which represents the best water quality condition of all clusters. Cluster 1 represents a similar water quality condition with Cluster 6 but with slightly higher SO_4^{2-}, TN, NO_3^--N , and DO concentrations; with Cluster 2 but with slightly higher SO_4^{2-} concentration; and with Cluster 7 but with slightly higher SO_4^{2-} and NO_3^--N concentrations. The visible patterns of Clusters 4, 5, and 8 (Figures 6(d), 6(e), and 6(h)) show the highest SO_4^{2-} and TN concentrations, which represent worse water quality conditions. The visible patterns of Cluster 3, 9, and 10 (Figures 6(d), 6(e), and 6(h)), as shown in Figure 5(b), represent medium water quality conditions.

The 10 classified clusters could be divided into three main environmental patterns. Relatively better water quality conditions were associated with Clusters 1, 2, 6, and 7, as

Table 2. Pearson Correlation Coefficient (N = 672)

	WD	T	SD	SS	DO	NH_4^+-N	NO_2^--N	NO_3^--N	TN	PO_4^{3-}	TP	SO_4^{2-}	Chla	Pheo
WD	1.00	0.10	-0.12	0.21	0.39	-0.60	-0.52	-0.18	-0.66	-0.37	-0.37	-0.43	-0.11	-0.26
T		1.00	-0.12	-0.10	-0.40	-0.24	0.21	-0.16	-0.26	0.12	0.13	-0.26	0.34	0.37
SD			1.00	-0.47	0.00	0.12	0.04	0.13	0.07	0.10	-0.12	0.03	-0.13	-0.08
SS				1.00	0.16	-0.18	-0.17	0.00	-0.07	-0.18	0.11	0.00	0.06	-0.07
DO					1.00	-0.49	-0.46	0.08	-0.41	-0.57	-0.36	-0.19	0.03	-0.12
NH^+_4-N						1.00	0.46	0.13	0.84	0.46	0.40	0.55	-0.04	0.09
NO^-_2-N							1.00	0.18	0.52	0.39	0.35	0.25	0.23	0.37
NO^-_3-N								1.00	0.40	-0.04	-0.06	0.35	-0.02	0.06
TN									1.00	0.39	0.53	0.68	0.15	0.19
PO_4^{3-}										1.00	0.44	0.11	0.05	0.18
TP											1.00	0.21	0.65	0.40
SO_4^{2-}												1.00	-0.05	-0.04
Chla													1.00	0.63
Pheo														1.00

Table 3. Mean Values Calculated from the Raw Data of Each Parameter for the Entire Data Set and the Data Classified into the Respective Clusters

Cluster	WD	T	SD	SS	DO	NH_4^+-N	NO_2^--N	NO_3^--N	TN	PO_4^{3-}	TP	SO_4^{2-}	Chl a	Pheo	TN/TP
	m	°C	m	mg/L	mg/L	mg/L	mg/L	mg/L	mg/L	mg/L	mg/L	mg/L	μg/L	μg/L	-
1	2.46	22.03	0.34	59.28	8.35	0.32	0.06	1.83	3.97	0.00	0.09	91.83	16.77	4.34	44
2	2.06	25.80	0.39	43.63	9.25	1.10	0.19	1.08	4.84	0.01	0.24	72.73	77.45	13.25	20
3	2.66	12.73	0.26	79.57	9.95	0.13	0.02	0.57	2.15	0.00	0.09	59.94	6.84	2.48	24
4	1.88	10.78	0.61	24.04	9.23	3.79	0.08	1.28	6.77	0.02	0.15	102.73	14.23	3.85	45
5	1.95	25.42	0.45	35.62	3.67	3.27	0.19	0.67	5.81	0.06	0.24	70.17	28.78	8.24	24
6	2.73	25.14	0.48	30.52	7.94	0.06	0.02	0.32	1.44	0.01	0.08	52.15	15.03	3.73	18
7	2.59	25.50	0.29	67.29	7.99	0.13	0.04	0.45	2.59	0.01	0.17	76.56	27.91	4.92	15
8	1.64	13.07	0.42	44.43	3.59	5.97	0.15	1.12	9.76	0.03	0.29	120.19	13.53	5.45	34
9	2.43	8.18	0.29	82.00	10.78	1.95	0.04	1.42	5.66	0.01	0.12	98.19	8.61	2.32	47
10	2.48	9.40	0.75	25.83	10.43	0.45	0.02	1.07	2.99	0.01	0.06	72.86	5.87	2.24	50
All data	2.34	17.89	0.42	50.78	8.24	1.49	0.07	0.93	4.26	0.01	0.15	78.90	20.26	4.85	28

shown on the right-hand side of Figure 5(b). The second group included Clusters 4, 5 and 8, which represent relatively worse water quality conditions, as shown on the lower left-hand side of Figure 5(b). The third group contains Clusters 3, 9 and 10, which represent medium water quality conditions.

Tabel 2 shows the mean values calculated from the raw data of each parameter for the entire data set and the data classified into the respective clusters. The high chemical parameters of Clusters 1, 2, 6 and 7 indicate lower mean values than the entire data set. Cluster 6 showed lower mean values for pollutants, such as NH_4^+-N , NO_2^--N , NO_3^--N , TN, PO_4^{3-}, TP, SO_4^{2-} , Chl-a, and Pheo, than those for the entire data set. This finding confirms that the cluster represented high water quality, as mentioned previously. Clusters 1, 2, and 7 showed slightly higher mean values for DO, NH_4^+-N, NO_2^--N, NO_3^--N, TN, PO_4^{3-}, TP, SO_4^{2-}, Chl-a, and Pheo than those of Cluster 6. However, the mean values of Clusters 3, 9, and 10 ranged between slightly lower and higher for all parameters compared with the mean value of the entire data set. In particular, Clusters 4, 5, and 8 represented considerably higher mean values for the pollutants than those for the entire data set, thereby showing significant deterioration of water quality.

Furthermore, the frequency of data classified into each cluster was investigated in the respective stations on a monthly basis to better understand the spatiotemporal variability. Figure 7 shows the spatiotemporal grids for the clusters listed previously, and the number of data occurrences was counted. The horizontal axis in the mesh represents each month, whereas the vertical axis represents the eight stations. The maximum data frequency of a particular month and station was seven because the data measurement period is seven years. The sums of the data frequencies for each station are shown in the column to the right of the grids, whereas the sums of the data frequencies for each month are shown in the row to the bottom of the grids.

With regards to temporal variations, based on Table 3, the environmental monitoring data associated with Cluster 6, which represent the best water quality condition, were mainly measured during the summer and autumn seasons, as shown in Figure 7(f). Clusters 2 and 7 (Figures 7(b) and 7(g)) mainly contain data also measured during the summer and autumn seasons, but the data for Cluster 1 (Figure 7(a)) were measured during the spring season.

The data classified into Clusters 4, 5 and 8, which re-

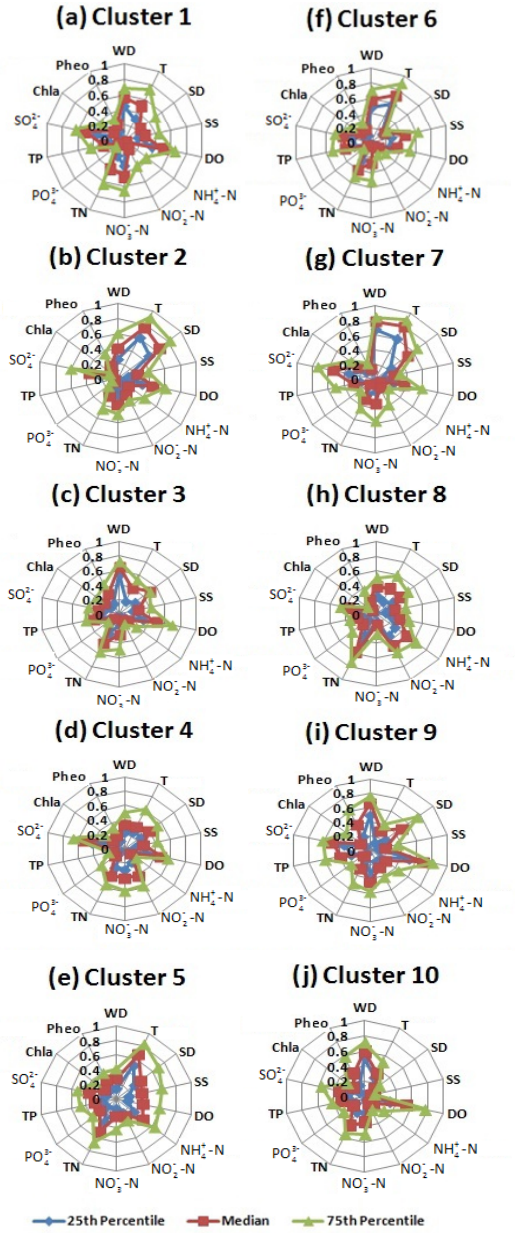

Figure 6. Radar graphs for the respective clusters with the 25th percentile, median, and 75th percentile of the standardized data.

present the worst water quality condition, were mostly measured during the spring and winter seasons (Figures 7(d), 7(e), and 7(h)). The data in Clusters 3, 9, and 10 (Figures 7(c), 7(i), and 7(j)) were observed mostly during the spring season. The temporal distribution of the data showed that the clusters including the data measured during the spring and winter seasons generally showed worse water quality conditions (e.g. Cluster 8) with low DO values.

When the spatiotemporal variation of each cluster was characterized in detail, Cluster 6 included water quality mostly measured from the THL01, THL03, THL04, THL05, THL07, and THL08 stations during the summer and autumn seasons. Clusters 1 and 7 were related to the data measured from the same six stations during the summer and autumn seasons. Cluster 2 data were mostly measured from the THL00, THL01, THL03, and THL06 stations during the summer season.

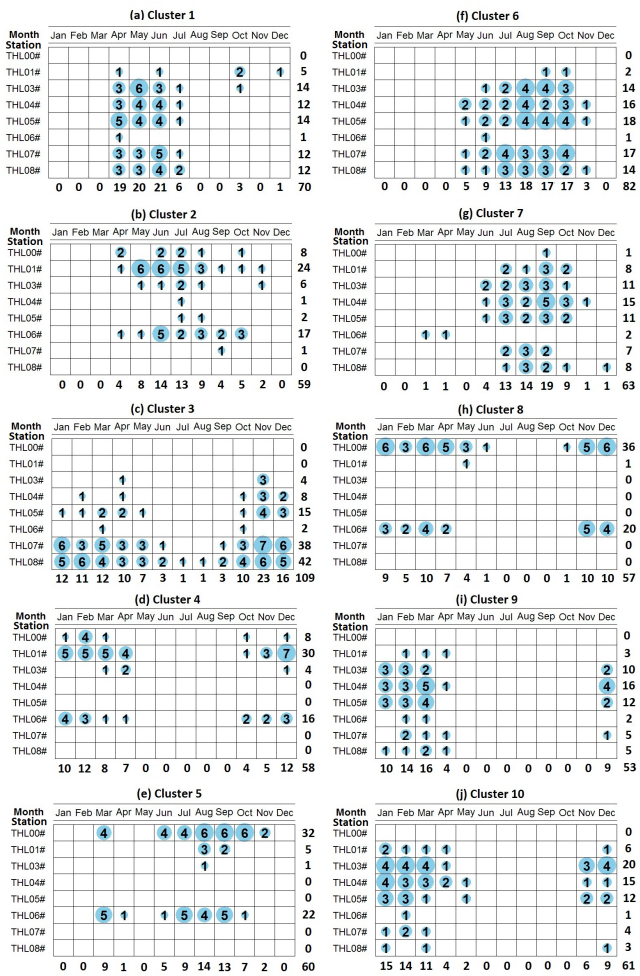

Figure 7. Balloon plot of the spatiotemporal patterns of Clusters 1 to 10.

Clusters 4, 5 and 8 were mostly measured from the THL00, THL01 and THL06 stations. The water quality of Clusters 4 and 8 was measured primarily during the winter and spring seasons, whereas that of Cluster 5 was measured during the autumn season. The data measured during the winter and spring seasons were associated with Clusters 3, 9, and 10. The data of these groups were measured from most stations, expect THL00, which is located near the estuary.

Spatiotemporal grid analysis was performed on the results from pattern classification by SOM application, and the use of SOM was confirmed to classify the parameters into 10 clusters, which was reasonable and feasible for the lake. Spatiotemporal grid analysis also summarized the spatiotemporal distribution of the respective 10 clusters with readily understandable visualization. The spatiotemporal grids analysis proposed in the present study was thus useful for characterizing and understanding the spatial and temporal variability and in-

terdependence of water quality parameters measured in multiple stations.

For lake managers, eutrophication is a significant issue. However, the eutrophication of lakes is mainly caused by nitrogen and phosphorus pollutants. The mass ratio of nitrogen and phosphorus (TN/TP) varies with lake trophic status and reflects the source of nutrients. TN/TP in oligotrophic lakes ranged from 21 to 240, in mesotrophic lakes from 17 to 96, in eutrophic lakes from 4 to 71, and in hypereutrophic lakes from 0.5 to 9 (Downing and McCauley, 1992). Based on these results, the average TN/TP mass ratios ranged from 15 to 50, as shown in Table 3.

Accordingly, one can speculate that the status of the northern part of Lake Taihu is mesotrophic or eutrophic. Therefore, controlling the emissions of nitrogen and phosphorus is recommended. Moreover, to control nitrogen emission, ammonia should be used as one of the main control factors based on the previous correlation analysis. As shown in Figures 6 and 7, the water quality conditions in the THL00, THL01, THL03, and THL06 stations are worse based on seasonal changes. Measures should be implemented to decrease pollutant emissions from the rivers into the lake.

Nowadays, many similar studies (Jin et al., 2011) presented the application of SOM in Korea and classified the environmental data into nine clusters in terms of DBI value. The study of Jin et al. (2011) focused on the water quality of the river, and the hierarchical cluster tree was used for the classification. By contrast, in our study, our main consideration is the water quality of the lake, and we selected and used the nonhierarchical *k*-means algorithm for the classification. Other studies (Tobiszewski et al., 2010; Tsakovski et al., 2010b; Yang et al., 2012) identified spatiotemporal patterns by using not only the SOM but also other methods for analysis, including the Hasse diagram technique, hierarchical CA, and DA.

A previous study confirmed the classification and visualization ability of the SOM algorithm for substantial environmental data, and the specific "resolving power" classification of the SOM was compared with more traditional methods, such as CA or PCA (Astel et al., 2007). Despite having a good pattern recognition ability, the SOM cannot detect the year-to-year trend. In this study, we were able to determine the seasonal water quality conditions of each site in the past seven years. In summary, the SOM is a feasible method for water quality assessment. The combination of SOM and other conventional and nonconventional analytical methods should be further investigated in the future.

5. Conclusions

We conclude that the application of SOM for analysis is suitable for handling environmental data sets describing variations in 14 chemical and biological quality parameters sampled monthly for seven years at eight sampling stations in the northern part of Lake Taihu, China. Visualization of the monitoring results of SOM enables the classification of different water quality patterns for all stations under consideration and for the entire monitoring period. The 25th, 50th, and 75th percentiles of the reference vectors were plotted on the radar graph to display the fundamental characteristics of each cluster. Moreover, the number of data on occurrences in the respective stations on a monthly basis for each cluster was displayed in the spatiotemporal grids to characterize the spatiotemporal variability of the environmental monitoring data. The spatiotemporal distribution of the environmental monitoring data was examined based on the characteristics of the respective clusters. The spatial distribution revealed generally better water quality conditions at the center of the lake than the estuary. The temporal distribution showed a distinct seasonal effect.

In addition, based on the applicability and feasibility of the SOM presented in this study, further research on the application of SOMs for the integrated assessment of a lake basin with simultaneous consideration of ecological, environmental, and geographical factors should be conducted.

Acknowledgments. The authors would wish to express their sincere gratitude to CNERN TaiLLER for providing the monitoring data. The authors thank Dr. Naveed (College of Environmental and Resources Sciences, Zhejiang University) for language help. The authors also deeply appreciate the anonymous reviewers and the editor for their insightful reviews. The provided valuable comments/suggestions have contributed much to improving the manuscript.

References

Alvarez-Guerra, M., González-Piñuela, C., Andrés, A., Galán, B. and Viguri, J.R. (2008). Assessment of self-organizing map artificial neural networks for the classification of sediment quality. *Environ. Int.,* 34(6), 782-790. http://dx.doi.org/10.10 16/j.envint.2008.01.006

Astel, A., Tsakouski, S., Barbieri, P., and Simeonov, V. (2007). Comparison of self-organizing maps classification approach with cluster and principal components analysis for large environmental data sets. *Water Res.,* 41(19), 4566-4578. http://dx.doi.org/10.1016/j.watres.2007.06.030

Astel, A., Tsakovski, S., Simeonov, V., Reisenhofer, E., Piselli, S., and Barbieri, P. (2008). Multivariate classification and modeling in surface water pollution estimation. *Anal. Bioanal. Chem.,* 390 (5), 1283-1292. http://dx.doi.org/10.1007/s00216- 007-1700-6

Chen, Y.W., Fan, C.X., Teubner, K., and Dokulil, M. (2003a). Changes of nutrients and phytoplankton chlorophyll-alpha in a large shallow lake, Taihu, China: An 8-year investigation. *Hydrobiologia,* 506(1-3), 273-279. http://dx.doi.org/10.1023/b:hydr.0000008604.09751.01

Chen, Y.W., Qin, B.Q., Teubner, K., and Dokulil, M.T. (2003b). Long-term dynamics of phytoplankton assemblages: Microcystis-domination in Lake Taihu, a large shallow lake in China. *J. Plankton Res.,* 25(4), 445-453. http://dx.doi.org/10.1093/plankt/25.4.445

Chon, T.S. (2011). Self-organizing maps applied to ecological sciences. *Ecol. Inf.,* 6(1): 50-61. http://dx.doi.org/10.1016/j.ecoinf.2010.11.002

Davies, D., and Bouldin, D. (1979). A cluster separation measure. *Pattern Anal. Mach. Intell., IEEE Trans. on PAMI,* 1(2), 224-227. http://dx.doi.org/citeulike-article-id:4173429

Desgraupes, B. (2013). *Clustering Indices,* Package clusterCrit for R, University Paris Ouest, Lab Modal'X: 1-34.

Downing, J.A., and McCauley, E. (1992). The nitrogen: phosphorus relationship in lakes. *Limnol. Oceanogr.,* 37(5), 936-945. http://dx.

doi.org/10.4319/lo.1992.37.5.0936

Huang, X.F., Chen, W.M., and Cai, Q.M. (1999). *Survey, observation and analysis of lake ecology,* Beijing, Standards Press of China.

Jin, Y.H., Kawamura, A., Park, S.C., Nakagawa, N., Amaguchi, H., and Olsson, J. (2011). Spatiotemporal classification of environmental monitoring data in the Yeongsan River basin, Korea, using self-organizing maps. *J. Environ. Monit.,* 13(10), 2886-2894. http://dx.doi.org/10.1039/c1em10132c

Kalteh, A.M., Hiorth, P., and Bemdtsson, R. (2008). Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application. *Environ. Model. Software,* 23(7), 835-845. http://dx.doi.org/10.1016/j.envsoft.2007.10.00 1

Khalil, B., and Ouarda, T. (2009). Statistical approaches used to assess and redesign surface water-quality-monitoring networks. *J. Environ. Monit.,* 11(11), 1915-1929. http://dx.doi.org/10.1039/b9095 21g

Kohonen, T. (1982a). Analysis of a simple self-organizing process. *Biol. Cybernetics,* 44(2), 135-140. http://dx.doi.org/10.1007/BF0 0317973

Kohonen, T. (1982b). Self-organized formation of topologically correct feature maps. *Biol. Cybernetics,* 43(1): 59-69. http://dx.doi.org/10.1007/BF00337288

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE,* 78(9), 1464-1480. http://dx.doi.org/10.1109/5.58325

Kohonen, T. (2001). Self-organizing maps of massive databases. *Eng. Intell. Syst. Electr. Eng. Commun.,* 9(4), 179-185. http://dx.doi.org/10.1007/978-3-642-56927-2

Kohonen, T. (2003a). Self-organized maps of sensory events. *Philos. Trans. R. Soc. Lond., Seri. A: Math. Phys. Eng. Sci.,* 361(1807), 1177-1186. http://dx.doi.org/10.1098/rsta.2003.11 92

Kohonen, T. (2003b). A computational model of visual attention. *Proceedings of the International Joint Conference on Neural Networks 2003,* Vols 1-4: 3238-3243.

Kohonen, T. (2008). Data management by self-organizing maps. *Computat. Intell. Res. Front.,* 5050, 309-332. http://dx.doi.org/10.1007/ 978-3-540-68860-0_15

Kohonen, T., and Makisara, K. (1989). The self-organizing feature maps. *Physica. Scripta.*, 39(1), 168-172. http://dx.doi.org/10.108 8/0031-8949/39/1/027

Kohonen, T., and Somervuo, P. (2002). How to make large self-organizing maps for nonvectorial data. *Neural Networks,* 15(8-9), 945-952. http://dx.doi.org/10.1016/S0893-6080(02) 00069-2

Lovchinov, V., and Tsakovski, S. (2006). Multivariate statistical approaches as applied to environmental physics studies. *Cent. Eur. J. Phys.,* 4(2), 277-298. http://dx.doi.org/10.2478/s11534-006-00 12-3

Noori, R., Karbassi, A., Khakpour, A., Shahbazbegian, M., Badam, H., and Vesali-Naseh, M. (2012). Chemometric Analysis of Surface Water Quality Data: Case Study of the Gorganrud River Basin, Iran. *Environ. Model. Assess.,* 17(4), 411-420. http://dx.doi.org/10.1007/s10666-011-9302-2

Omo-Irabor, O.O., Olobaniyi, S.B., Oduyemli, K., and Alunna, J. (2008). Surface and groundwater water quality assessment using multivariate analytical methods: A case study of the Western Niger Delta, Nigeria. *Phys. Chem. Earth,* 33(8-13), 666-673. http://dx.doi. org/10.1016/j.pce.2008.06.019

Oyana, T.J. (2009). Visualization of high-dimensional clinically acquired geographic data using the self-organizing maps. *J. Environ. Inf.,* 13(1), 33-44. http://dx.doi.org/10.3808/jei.2009001 38

Paerl, H.W., Xu, H., McCarthy, M.J., Zhu, G.W., Qin, B.Q., Li, Y.P., and Gardner, W.S. (2011). Controlling harmful cyanobacterial blooms in a hyper-eutrophic lake (Lake Taihu, China): The need for a dual nutrient (N & P) management strategy. *Water Res.,* 45(5), 1973-1983. http://dx.doi.org/10.1016/j. watres.2010.09.018

Qin, B.Q., Xu, P.Z., Wu, Q.L., Luo, L.C., and Zhang, Y.L. (2007). Environmental issues of Lake Taihu, China. *Hydrobiologia,* 581, 3-14. http://dx.doi.org/10.1007/s10750-006-0521-5.

Shin, P.K.S., and Fong, K.Y.S. (1999). Multiple discriminant analysis of marine sediment data. *Mar. Pollut. Bull.,* 39(1-12), 285-294. http://dx.doi.org/10.1016/s0025-326x(99)00113-7

Su, S.L., Zhi, J.J., Lou, L.P., Huang, F., Chen, X., and Wu, J.P. (2011). Spatio-temporal patterns and source apportionment of pollution in Qiantang River (China) using neural-based modeling and multivariate statistical techniques. *Phys. Chem. Earth,* 36(9-11), 379-386. http://dx.doi.org/10.1016/j.pce.2010. 03.021

Tobiszewski, M., Tsakovski, S., Simeonov, V., and Namiesnik, J. (20 10). Surface water quality assessment by the use of combination of multivariate statistical classification and expert information. *Chemosphere,* 80(7), 740-746. http://dx.doi.org/10.1016/j.chemospher e.2010.05.024

Tsakovski, S., Astel, A., and Simeonov, V. (2010a). Assessment of the water quality of a river catchment by chemometric expertise. *J. Chemometrics,* 24(11-12), 694-702. http://dx.doi.org/10.1002/cem. 1333

Tsakovski, S., Tobiszewski, M., Simeonov, V., Polkowska, Z., and Namiesnik, J. (2010b). Chemical composition of water from roofs in Gdansk, Poland. *Environ. Pollut.,* 158(1), 84-91. http://dx.doi.org/10.1016/j.envpol.2009.07.037

Vesanto, J. (2000). *Using SOM in data mining,* the Degree of Licetiate of Science and Technology, Helsinki University of Technology.

Vesanto, J., Himberg, J., Alhoniemi, E., and Parhankangas, J. (2000). *SOM Toolbox for MATLAB 5,* Laboratory of Computer and Information Science, Helsinki University of Technology.

Wei, F.S., Qi, W.Q., Bi, T., Kong, Z.G., Huang, Y.R., and Shen, Y.W. (2002). *Methods for water and wastwater monitoring and analysis,* Beijing, China Environmental Science Press.

Wu, X.D., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.H., Steinbach, M., Hand, D.J., and Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowl. Inf. Syst.,* 14(1), 1-37. http://dx.doi.org/10.1007/s10115-007-0114-2

Yang, Y., Wang, C., Guo, H., Sheng, H., and Zhou, F. (2012). An integrated SOM-based multivariate approach for spatio-temporal patterns identification and source apportionment of pollution in complex river network. *Environ. Pollut.,* 168(0), 71-79. http://dx.doi.org/10.1016/j.envpol.2012.03.041

Zhang, Y., Guo, F., Meng, W., and Wang, X.Q. (2009). Water quality assessment and source identification of Daliao river basin using multivariate statistical methods. *Environ. Monit. Assess.,* 152(1-4), 105-121. http://dx.doi.org/10.1007/s10661- 008-0300-z.

Zhou, F., Liu, Y., and Guo, H.C. (2007a). Application of multivariate statistical methods to water quality assessment of the watercourses in northwestern new territories, Hong Kong. *Environ. Monit. Assess.,* 132(1-3), 1-13. http://dx.doi.org/10.1007/s10661-006-9497-x

Zhou, F., Huang, G.H., Guo, H.C., Zhang, W., and Hao, Z.J. (2007b). Spatio-temporal patterns and source apportionment of coastal water pollution in eastern Hong Kong. *Water Res.,* 41(15), 3429-3439. http://dx.doi.org/10.1016/j.watres.2007.04. 02