

Water pollution sources assessment by multivariate statistical methods in the Tahtali Basin, Turkey

Hülya Boyacioglu · Hayal Boyacioglu

Received: 2 September 2006 / Accepted: 14 May 2007 / Published online: 7 June 2007
© Springer-Verlag 2007

Abstract In this study, multivariate statistical methods including factor, principal component and cluster analysis were applied to surface water quality data sets obtained from the Tahtali River Basin, Turkey. Factor and principal components analysis results revealed that surface water quality was mainly controlled by agricultural uses and domestic discharges. Cluster analysis generated two clusters. Based on the locations of the sites consisted by each cluster and variable concentrations at these stations, it was concluded that agricultural discharges strongly affected north and northeast part of the region. These methods are believed to assist water managers to understand complex nature of water quality issues and determine priorities to improve water quality.

Keywords Cluster analysis · Factor analysis · Multivariate statistical methods · Principal component analysis · Tahtali River Basin · Water quality

Introduction

Water quality assessment aims to describe water quality at regional, national, and international scales, determine

whether or not water quality meets previously defined objectives for designated uses and trends in time, etc. Water quality is defined in terms of its physical, chemical, and biological parameters. Traditional approaches to assessing water quality are based on the comparison of experimentally determined parameter values with the existing guidelines but in many cases it does not readily give information on status of the source (Debels et al. 2005).

One of the difficult tasks facing environmental managers is to convert complex data to information for better defining the sources and typology of the pollution, etc. Environmental data is characterized by high variability because of a variety of natural and anthropogenic influences. The best approach to avoid misinterpretation of environmental monitoring data is the application of multivariate statistical methods for environmental data classification and modeling (Reisenhofer et al. 1996).

Surface water, groundwater water quality assessment, and environmental research employing multi-component techniques are well described in the literature (Praus 2005). They allow to derive information from the data set about the possible influences of the environment on water quality (Spanos et al. 2003). Advantages of multivariate statistical techniques for ecological data can be summarised as:

- reflect more accurately the multivariate nature of natural ecological system
- provide a way to handle large data sets with large numbers of variables by summarizing the redundancy
- provide a means of detecting and quantifying truly multivariate patterns that arise out of the correlation structure of the variable set (McGarial et al. 2000).

In this study factor, principal component and cluster analysis were applied to assess water quality.

H. Boyacioglu (✉)
Department of Environmental Engineering,
Faculty of Engineering, Dokuz Eylul University,
Tinaztepe Campus, Buca, 35160 Izmir, Turkey
e-mail: hulya.boyacioglu@deu.edu.tr

H. Boyacioglu
Department of Statistics, Faculty of Science,
Ege University, Bornova, 35100 Izmir, Turkey
e-mail: hayal.boyacioglu@ege.edu.tr

Factor analysis and principal components analysis

Principal component analysis (PCA) and factor analysis (FA) are both variable reduction techniques. PCA includes correlated variables with the purpose of reducing the numbers of variables and explaining the same amount of variance with fewer variables (principal components). Factor analysis attempts to explain the correlations between the observations in terms of the underlying factors, which are not directly observable (Yu et al. 2003).

The difference between PCA and FA is that for the purposes of matrix computations PCA assumes that all variance is common, with all unique factors set equal to zero; while FA assumes that there is some unique variance. The level of unique variance is dictated by the FA model which is chosen. Accordingly, PCA is a model of a closed system, while FA is a model of an open system (Wulder 2007).

General steps of the FA method can be summarized as:

- for all the variables a correlation matrix is generated
- factors are extracted from the correlation matrix based on the correlation coefficients of the variables
- factors are rotated to increase interpretability (Gupta et al. 2005; DeCoster 1998).

Procedural steps of the PCA are:

- number of components equal to number of variables are generated
- the number of components to retain are determined
- components are rotated (rotations is a linear transformation of the solution to make interpretation easier)
- rotated solution is interpreted (Suhr 2005).

Cluster analysis

Cluster analysis (CA) is used to develop meaningful aggregations, or groups, of entities based on a large number of interdependent variables. Specifically the purpose is to classify a sample of entities into a smaller number of usually mutually exclusive groups based on the multivariate similarities among entities (McGarial et al. 2000). CA divides a large number of objects into smaller number of homogenous groups on the basis of their correlation structure. Within-group similarity is maximized and among-group similarity is minimized according to some objective criteria (Zeng and Rasmussen 2005; McGarial et al. 2000). Important characteristics of cluster analysis are:

- organizes observations into discrete classes or groups such that within group similarity is maximized and among-group similarity is minimized according to some objective criteria,

- assesses relationships within a single set of variables; no attempt is made to define the relationship between a set of independent variables and one or more dependent variables, etc. (McGarial et al. 2000).

Typical clustering activity involves following steps:

- definition of observation similarity
- clustering or grouping
- data abstraction (Jain et al. 1999; Boyacioglu and Boyacioglu 2007; Mihailov et al. 2005).

This study comprises application of multivariate statistical techniques to water quality data set obtained from Tahtali River and its tributaries in Turkey. Statistical calculations were performed using the “*Statistical Package for the Social Sciences Software-SPSS 10.0 for Windows*” and “*Minitab 13 Statistical Software*”.

Tahtali River Basin

Tahtali Basin is located in the western part of Turkey (see Fig. 1). The climate of the region is typical Mediterranean: hot and dry in summers and temperate and rainy in winters. Reservoir located on the Tahtali River is the main drinking water source for the Izmir City, where approximately three million inhabitants live in. There are about 30 settlements in the basin and about 60,000 people live in the catchment area as of 2,000. The land area of the Tahtali Basin is 550 km² and primarily consists forest and agricultural land (60 and 18% of the total area). In addition residential areas cover 2% of the basin (Alpaslan et al. 2002; Boyacioglu 2007).

Application of multivariate statistics to Tahtali Basin water quality data

In this study multivariate statistical methods including factor, principal component, and cluster analysis were applied to a surface water quality data set collected from Tahtali River and its tributaries. Data obtained from seven observation stations in the study area along 6 years were processed. Monitoring stations are seen in Fig. 2. The selected parameters for the estimation of surface water quality characteristics were: biochemical oxygen demand (BOD₅), chemical oxygen demand (COD), dissolved oxygen (DO), chloride (Cl⁻), nitrate–nitrogen (NO₃-N), sodium (Na⁺), sulfate (SO₄²⁻), total dissolved solids (TDS), total phosphorus, and oxygen saturation (%).

Descriptive statistics of these parameters are presented in Table 1.

Fig. 1 Location of Tahtali Basin

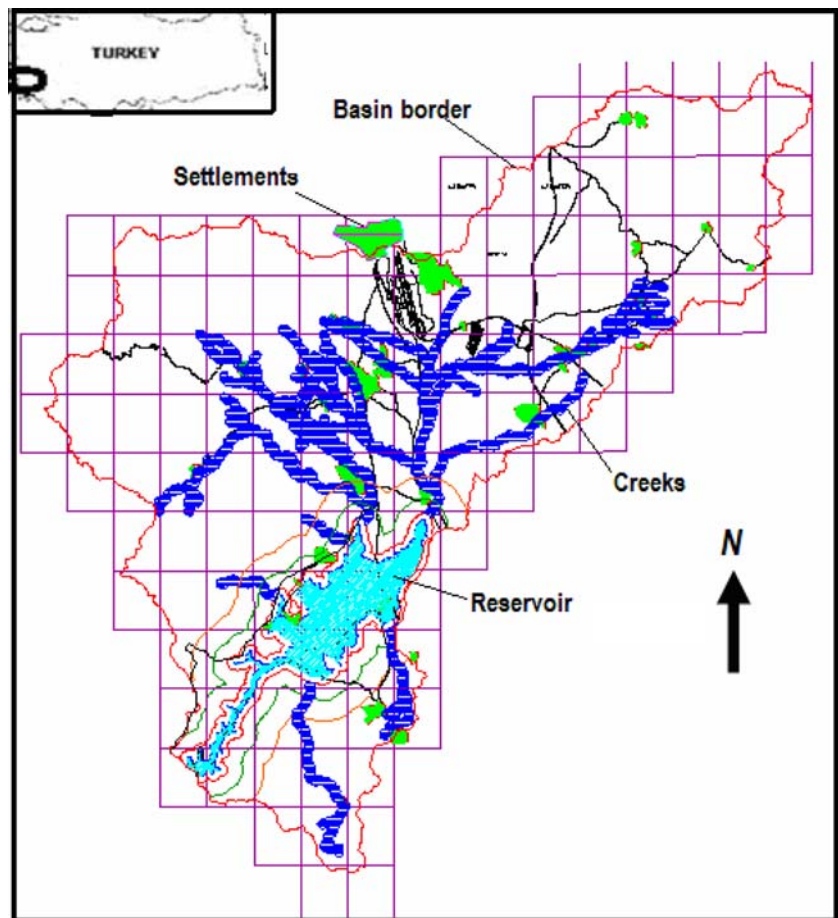
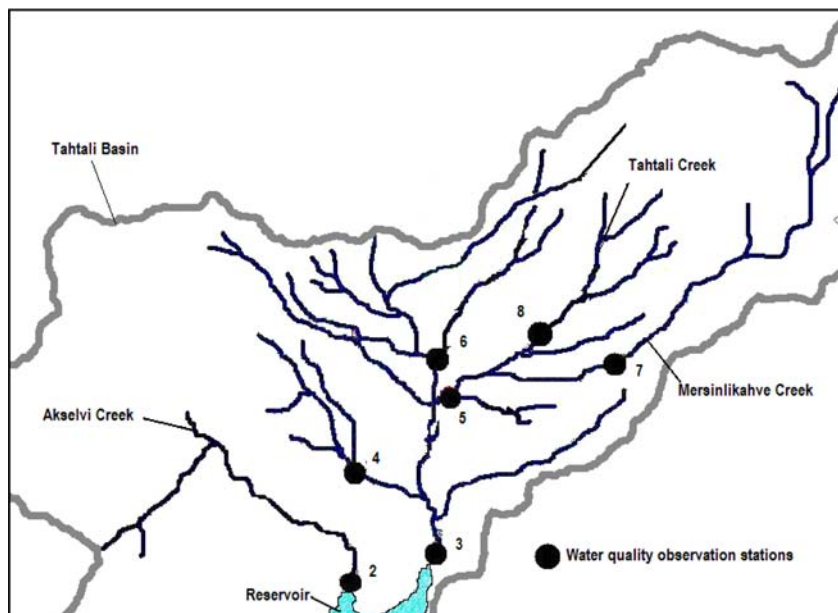


Fig. 2 Tahtali Basin river water quality monitoring stations



Results of factor analysis

In this study, water quality variables were grouped using FA. The correlation matrix of variables was generated and

factors extracted by the Centroid method, rotated by Varimax. Results of factor analysis including factor-loading matrix, eigen values, total, and cumulative variance values are presented in Table 2.

Table 1 Descriptive statistics of water quality variables

	BOD ₅	DO	COD	Cl ⁻	NO ₃ -N	Na ⁺	SO ₄ ²⁻	TDS	Total phosphorus	Oxygen saturation
Unit	mg/l	mg/l	mg/l	mg/l	mg/l	mg/l	mg/l	mg/l	mg/l	%
Mean	2.60	7.27	8.93	41.76	6.29	23.30	45.49	375.45	0.04	82.93
Median	2.00	7.21	8.00	40.00	6.60	22.00	46.00	374.00	0.00	82.00
Mode	2.00	6.70	8.00	40.00	7.00	23.00	46.00	359.00	0.00	77.00
Std. deviation	1.67	0.73	4.01	11.24	3.18	7.81	10.24	67.61	0.08	7.38
Variance	2.80	0.54	16.07	126.33	10.14	61.03	104.91	4570.60	0.01	54.43
Minimum	1.00	4.41	4.00	20.00	1.20	10.00	21.00	219.00	0.00	53.60
Maximum	9.00	8.98	19.00	80.00	22.00	56.00	74.00	552.00	0.50	97.80

Table 2 Factor-loading matrix, eigenvalues and total and cumulative variance values

Variable	Factor 1	Factor 2	Factor 3
BOD ₅	0.189	-0.013	0.800
DO	-0.058	0.975	-0.073
COD	-0.177	-0.117	0.395
Chloride	0.935	-0.093	-0.009
Nitrate–nitrogen	0.746	-0.029	-0.167
Sodium	0.858	0.014	0.186
Sulfate	0.833	0.173	-0.044
TDS	0.744	-0.010	-0.237
Total phosphorus	-0.135	0.034	0.841
Oxygen saturation	0.070	0.975	-0.040
Eigenvalue	3.506	1.957	1.632
% Total variance	35.062	19.571	16.315
Cumulative %	35.062	54.632	70.948

The factor analysis generated three significant factors, which explained 70.95% of the variance in data sets. Parameters were grouped based on the factor loadings and the following factors were indicated:

- Factor 1: Chloride, sodium, sulphate, TDS, nitrate–nitrogen
- Factor 2: DO, oxygen saturation
- Factor 3: BOD₅, total phosphorus, COD

Chloride, sodium, sulfate, TDS, and nitrate–nitrogen marked factor 1 (F1) explained 35.1% of the variance. DO and oxygen saturation (%) were correlated with factor 2 (F2) and BOD₅, total phosphorus, COD with factor 3 (F3). The F1 had a high positive loading in chloride, sodium, sulphate, nitrate–nitrogen, TDS which were 0.94, 0.86, 0.83, 0.75, and 0.74. High positive loadings indicated strong linear correlation between the factor and parameters.

Surface and subsurface drainage effluent contains substances that are potential pollutants. These pollutants may be nutrients, salinity (sodium, calcium, magnesium, and to a lesser extent potassium), major ions (e.g., chloride, sulfate and nitrate), and pesticides, etc. High nitrate

concentrations in subsurface drainage can originate from a number of sources: geologic deposits, natural organic matter decomposition and deep percolation of nitrate resulting from fertilizer applications (FAO 1997). Phosphorus comes from both point and non-point sources. Point sources include municipal waste treatment plants, industrial operations, and large, confined livestock operations. Non-point sources comprise soil erosion and water runoff from cropland, lawns and gardens, private waste treatment systems, urban areas, small livestock confinement operations, etc. (Devlin et al. 2000).

Based on these statements, results of the factor analysis, and also hydro-chemical aspects of the water, it was concluded that, F1 can be denoted as “agricultural pollution” factor with presence of chloride, sodium, sulfate, and nitrate. As has been mentioned above these parameters are mainly found in agricultural drainage water. F2 which is highly correlated with DO and oxygen saturation can be denoted as “oxygen content” factor. COD, BOD₅, and total phosphorus consisted by F3 are the indicator of organic pollution in water. Therefore, F3 represents “organic pollution” factor.

In summary, three factors representing three different processes are,

- agricultural pollution factor
- oxygen content factor
- organic pollution factor caused by urbanization.

As was stated above, forest and agriculture is the primary land uses (78% of the total catchment area) in the Tahtali Basin. Based on this fact and results of the factor analysis it can be concluded that surface water quality of the Tahtali River and its tributaries was mainly controlled by agricultural discharges. On the other hand, uncontrolled domestic discharges caused by rapid urbanization were also threat for the surface water quality (Boyacioglu 2006).

Confirmation of FA results by PCA analysis

Principal component analysis (PCA) was applied to data set to confirm results of FA. A scree plot given in Fig. 3 shows the sorted eigenvalues from large to small as a function of the principal components number. As is seen in this figure PCA generated three significant components (number of components of which the eigenvalues are greater than ‘‘1’’ was three). The components weights are presented in Table 3.

PCA analysis results revealed that the first component was associated with chloride, sulphate, sodium, TDS, and nitrate–nitrogen. The second component comprised DO and oxygen saturation and third component BOD₅, COD, and total phosphorus.

Based on the findings of analysis, it was concluded that PC1 can be called as ‘‘agricultural pollution component’’, PC2 ‘‘oxygen content component’’, and PC3 ‘‘organic pollution component’’. Thus these findings confirmed the results of factor analysis.

Table 3 Principal component weights

Variable	PC1	PC2	PC3
BOD ₅	0.049	0.191	0.616
DO	0.001	-0.675	0.182
COD	-0.121	0.153	0.256
Chloride	0.491	0.117	0.041
Nitrate-nitrogen	0.403	0.028	-0.076
Sodium	0.441	0.082	0.209
Sulfate	0.447	-0.076	0.073
TDS	0.407	0.001	-0.125
Total phosphorus	-0.122	0.149	0.634
Oxygen saturation	0.066	-0.660	0.217

Results of cluster analysis

In this study, sampling site classification was performed by the use of cluster analysis (z-transformation of the input data, squared Euclidean distance as similarity measure and Ward’s method of linkage) and dendrogram was generated.

A dendrogram that clearly differentiates groups of objects has small distances in the far branches of the tree and large differences in the near branches. The following example dendrogram ideally illustrates two clear groups (see Fig. 4).

Stations grouped under each cluster are seen in Fig. 5. Based on the results of cluster analysis, it was concluded that:

Cluster I (Stations 3–4–2–8)

Sites mainly located at the downstream of the Creeks (Station 2–4) were grouped under Cluster I. In addition, Station 8 located upstream part of the Tahtali Creek

Fig. 3 Scree plot of the eigenvalues

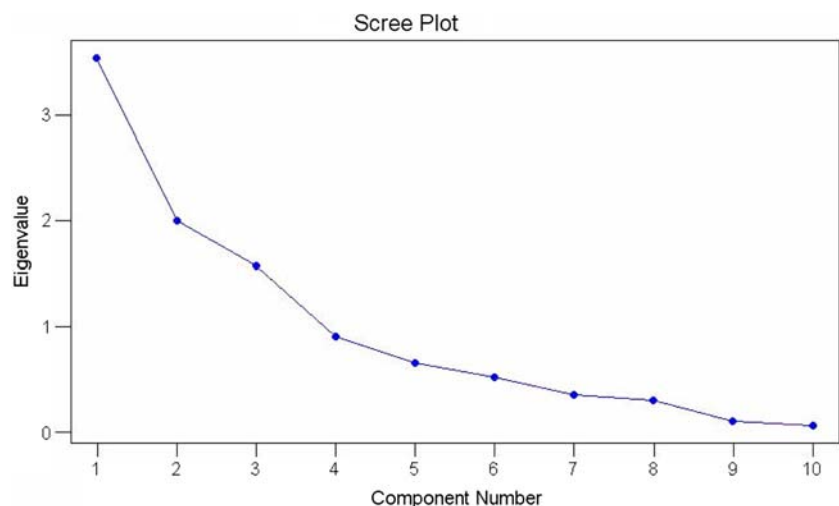


Fig. 4 Ward's minimum variance dendrogram

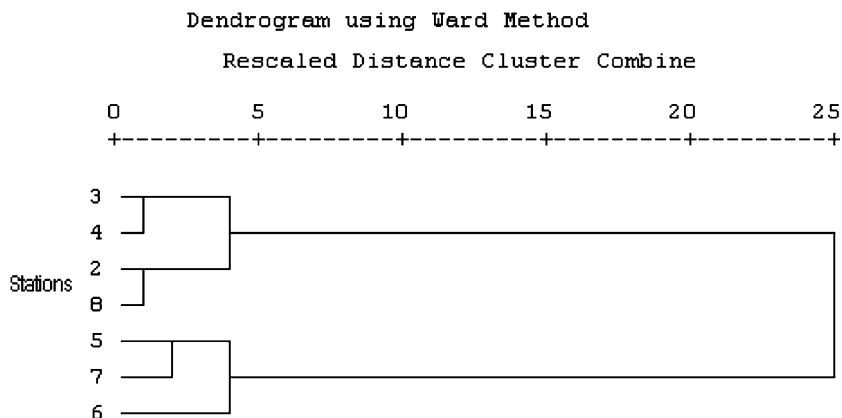
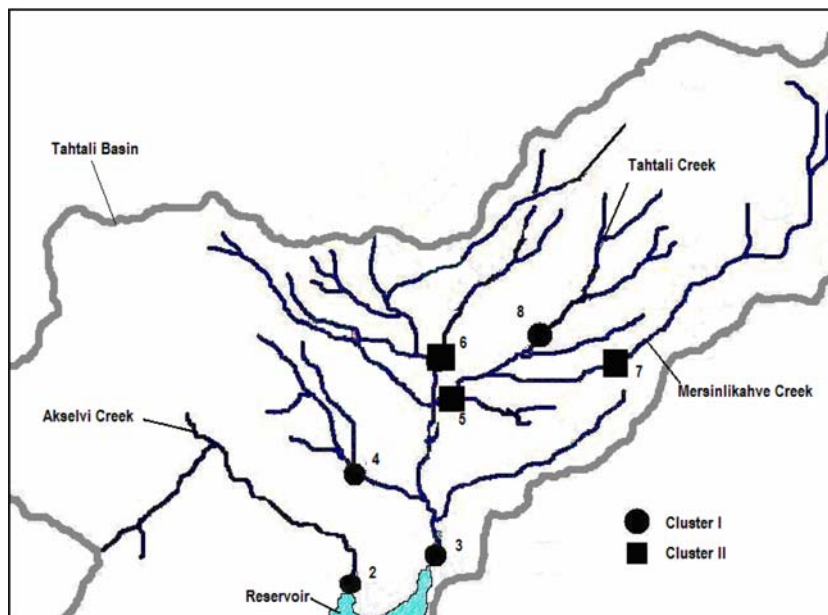


Fig. 5 Grouped stations based on cluster analysis



showed similar water quality characteristics with these stations.

Cluster II (Stations 5–7–6)

This cluster consisted sites located at the north and northeast part of the basin, where population density is considerably high and land use pattern is dominated by agricultural fields (see Fig. 6). The common feature of these sites was relatively high TDS, sodium, chloride, and sulfate concentrations compared to the other monitoring stations (see Fig. 7). As has been discussed above typical sources of these parameters are agricultural drainage waters. Thus, north and northeast part of the basin which were grouped under the same cluster, were strongly affected by agricultural discharges (Boyacioglu 2007). Land use pattern of the region also supports cluster analysis results.

Conclusions

Multivariate statistical techniques including factor, principal component and cluster analysis can successfully be used to derive information from the data set about the possible influences of the environment on water quality and also identify natural groupings in the set of data. These methods are important to avoid misinterpretation of environmental monitoring data due to uncertainties.

In this study, multivariate statistical methods were applied to data set obtained from Tahtali Basin that is the main drinking water source of Izmir, Turkey.

Factor analysis (FA) was performed to explain the correlations between the observations in terms of underlying factors and results were confirmed by PCA. FA generated three significant factors. The factor 1-F1 that was correlated with chloride, sodium, sulphate, TDS, and nitrate–nitrogen

Fig. 6 Land uses in the Northeastern Tahtali Basin

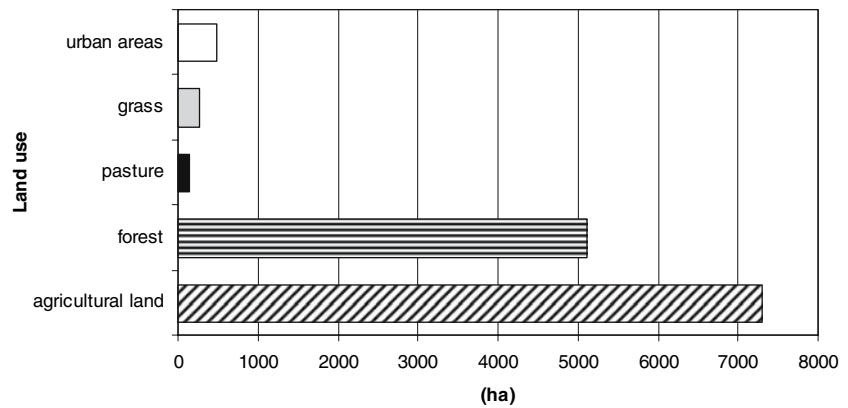
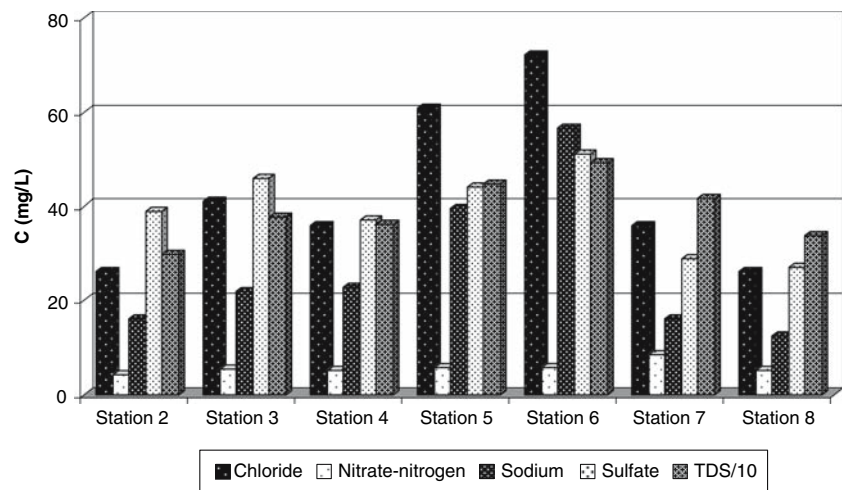


Fig. 7 Chloride, nitrate–nitrogen, sodium, sulfate, and TDS concentrations at river observation stations



explained 35% of the total variance. Factor 2-F2 marked dissolved oxygen and oxygen saturation, and factor 3-F3 comprised BOD₅, COD, and total phosphorus. These three factors explained 71% of the total variance in data set. Based on FA results and considering hydrochemical aspects of the water, land use pattern of the basin which was dominated by forest and agricultural land it was concluded that major water pollution threats in the region were agricultural and domestic discharges.

Cluster analysis (CA) analysis was used to classify stations with similar properties in the study area. Cluster I mainly comprised stations located at the downstream of the creeks. Monitoring sites located at north and northeast of the basin were grouped under Cluster II. At the stations grouped under Cluster II, water quality variables including TDS, sodium, chloride, and sulfate were relatively higher than the others. Therefore, north and northeast basin was the most influenced part by the agricultural pollution. Land uses in this region consisting primarily agricultural fields and forests (with 93% of the land area) explain the origin of the pollution. Consequently priorities should be given to these locations for minimization of effects of agricultural discharges.

This study shows that multivariate statistical methods are useful tool for understanding of complex nature of water quality issues by identifying groupings in the set of data. They assist decision makers in number of ways:

- derive information on possible influences of the environment on water quality in a water basin.
- analyze spatial variations, etc.

Thus, multivariate statistical methods including factor, principal component, and cluster analysis can be used to understand complex nature of water quality issues and determine priorities to improve water quality.

Acknowledgments The authors express their special thanks to members of the Izmir Water and Sewerage Authority (IZSU) for their assistance in providing necessary data for the study.

References

Alpaslan MN, Boyacıoğlu H and Dolgen D (2002) Water quality assessment and control in drinking water reservoirs by deductive approach. The 9th International specialized conference on river

- basin management, Edinburgh, 11–13 September 2002, preprint CD, Scotland, UK
- Boyacioglu H (2006) Development of universal water quality index and environmetric analysis in surface water quality assessment with field implementation and validation. Karlsruhe University. Master of Science thesis supervised by Prof. Dr. Franz Nestmann (2006)
- Boyacioglu H, Boyacioglu H (2007) Surface water quality assessment by environmetric methods. *Environ Monit Assess* (in press)
- Boyacioglu H (2007) Assessment of water quality by total maximum daily load (TMDL) analysis. PhD thesis, Dokuz Eylul University Graduate School of Natural and Applied Sciences, Turkey
- Debels P, Figueroa R, Urrutia R, Barra R, Niell X (2005) Evaluation of water quality in the Chillán River (Central Chile) using physicochemical parameters and a modified water quality index. *Environ Monit Assess* 110:301–322
- DeCoster J (1998) Overview of factor analysis. Retrieved May 24 2006 from <http://www.stat-help.com/notes.html>
- Devlin DL, Whitney DA, McWay KA (2000) Phosphorus and water quality in Kansas, Kansas State University. Retrieved July 4 2006 from <http://www.oznet.ksu.edu/library/h20q12/MF2463.pdf>
- Gupta AK, Gupta SK, Patil RS (2005) Statistical analyses of coastal water quality for a port and harbour region in India. *Environ Monit Assess* 102:179–200
- International Commission on Irrigation and Drainage Food and Agriculture Organization of the United Nations—FAO (1997) Management of agricultural drainage water quality. In: CA Madramootoo, WR Johnston and LS Willardson (eds). ISBN 92-5-104058-3. Retrieved June 19 2006 from <http://www.fao.org/docrep/W7224E/w7224e06.htm#salinity%20and%20major%20ions>
- Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. *ACM Computing Surveys*, vol. 31, no. 3. Retrieved May 26 2006 from <http://kilana.unibe.ch:8080/SCG/uploads/596/p264-jain.pdf>
- McGarral K, Cushman S, Stafford S (2000) *Multivariate statistics for wildlife and ecology research*. Springer, New York
- Mihailov G, Simeonov V, Nikolov N, Mirinchev G (2005) Multivariate statistical assessment of the pollution sources along the stream of Kamchia River. *Bulgaria Water Sci Technol* 51(11):37–43
- Minitab 13 Statistical Software (2000) Minitab, Inc
- Praus P (2005) Water quality assessment using SVD-based principal component analysis of hydrological data. *Water SA* 31:417–422
- Reisenhofer E, Adami G, Favretto E (1996) Heavy metals and nutrients in coastal, surface seawaters (Gulf of Trieste, Northern Adriatic Sea): an environmental study by factor analysis. *Fresenius J Anal Chem* 354:729–734
- Spanos T, Simeonov V, Stratis J, Xristina X (2003) Assessment of water quality for human consumption. *Microchim Acta* 141:35–40
- SPSS- 10.0 (2000) Statistical package for the social sciences. SPSS, Chicago
- Suhr D (2005) Principal component analysis vs. exploratory factor analysis. SUGI 30 Proceedings. Retrieved January 30 from <http://www2.sas.com/proceedings/sugi30/Leadrs30.pdf>
- Wulder M (2007) A practical guide to the use of selected multivariate statistics. Retrieved February 5, 2007 from http://www.pfc.cfs.nrcan.gc.ca/profiles/wulder/mvstats/index_e.html
- Yu S, Shang J, Zhao J, Guo H (2003) Factor analysis and dynamics of water quality of the Songhua River Northeast China. *Water Air and Soil Pollut* 144:159–169
- Zeng X, Rasmussen TC (2005) Multivariate statistical characterization of water quality in Lake Lanier, Georgia, USA. *J Environ Qual* 34:1980–1991