Asian Journal of Chemistry Vol. 21, No. 5 (2009), 4054-4062

Assessment of Surface Water Quality using Multivariate Statistical Analysis Techniques: A Case Study from Tahtali Dam, Turkey

SUHEYLA YEREL
Bozuyuk Vocational School, Bilecik University, Bozuyuk, Bilecik, Turkey
Fax: (90)(228)3141195; Tel: (90)(228)3141195
E-mail: syerel@anadolu.edu.tr

In this paper, the surface water quality of the Tahtali dam and tributaries in Izmir, Turkey are assessed by using multivariate statistical analysis techniques known as cluster analysis (CA), factor analysis (FA) and principal component analysis (PCA). Multivariate statistical analysis techniques were applied to the physical and inorganic chemical parameters including Cl^-, NO_3^-, NH_4^+–N, oxygen saturation, colour, Na^+, SO_4^{2-}, total phosphorus, dissolved oxygen, BOD_5 and COD obtained from the seven different surface water quality observation stations. These analyses results represent that domestic waste and nutrient pollution caused differences in terms of water quality in the northwest part of the area. Thus, this study show that the usefulness of multivariate statistical analysis techniques for analysis and interpretation in the water quality problem.

Key Words: Water quality, Observation station, Cluster analysis, Factor analysis, Principal component analysis, Pollution, Turkey.

INTRODUCTION

The surface water quality is a matter of serious concern today. Rivers, due to their role in carrying off the municipal and industrial wastewater and runoff from agricultural land in their vast drainage basins are among the most vulnerable water bodies to pollution. The surface water quality in a region is largely determined both by the natural process and the anthropogenic influence of water quality[1-3]. Many different sources and processes are known to contribute to the deterioration in quality and contamination of surface water. Thus, a thorough understanding of the nature and extent of contamination in an area requires detailed hydrochemical data[4,5]. Few studies have so far been undertaken combining the effect of multiple water quality variables in order to evaluate the water quality and the extent and nature of contamination[6].

Multivariate statistical analyses are used to incorporate larger numbers of variables measured in water systems[7]. The application of different multivariate statistical techniques, such as cluster analysis (CA), principal component analysis (PCA) and factor analysis (FA), help in the interpretation of complex data matrices to better understand the water quality and ecological status of the studies systems, allows the identification of possible factors that influence water systems and offers a valuable tool for reliable management of water resources as well as rapid solution to pollution

problems[8-13]. The intention underlying the use of multivariate statistical analysis is to achieve great efficiency of data compression from the original data, add to gain some information useful in the interpretation of the environmental geochemical origin. This method can also help indicate natural associations between samples and/or variables[14]. This multivariate treatment of environmental data is widely successfully used to interpret relationship a variables so that the environmental system could be better managed[15].

Some studies conducted in environmental sciences by using multivariate statistical analyses are as follows, Spears and Zheng[16] used single linkage cluster method for relationships between major elements in some UK coals. These relationships are illustrated on a dendrogram. Silicon, Al, K and Ti are closely associated and reflect the influence quartz and clay minerals. There is greater separation on the dendrogram for calcium. Muri[17] has investigated basic physical and chemical characteristics of water in lakes using cluster analysis. The condition of lakes was assessed. Although the water quality has deteriorated in some lakes, most of the lakes are still in a good condition. Simeonov *et al.*[18] show a description of multivariate statistical assessment of water quality of northern Greece based on the evaluation of a large and complex dataset. Cluster analysis were used for site similarity analysis, whereas for the identification of sources of pollution, PCA followed by absolute principal component scores were applied.

In this study, a large data matrix obtained during a 5 year monitoring program is subjected to cluster analysis (CA), principal component analysis (PCA) and factor analysis (FA) to extract information about the similarities between observation stations, identification of water quality variables for variations in Tahtali dam and tributaries in Izmir (Turkey).

EXPERIMENTAL

The Tahtali dam is located in the southwestern part of Izmir city, southwest Turkey (Fig. 1). The construction of the Tahtali dam was completed in 1996. The watershed of the Tahtali dam covers a surface area of 515 km^2, extending along the Cuma plain. The watershed is the water collection area of the Tahtali dam and tributaries, which runs in the northeast-southwest direction, through the Cuma plain[19]. About 50 % of the area is covered by forests and 30 % of the land is used for agricultural purposes, whereas, the remaining area represents urban area[20].

Dataset: Surface water quality dataset covers for the length of 5 years and contains the values of selected pollution indicators for 7 observations stations from the Tahtali dam and tributaries in Izmir (Turkey). Locations of the observations stations were depicted in Fig. 1 and selected pollution indicators were given in Table-1, respectively. Descriptive statistics of the data set were presented in Table-2.

Multivariate statistical analysis techniques: In this study, surface water quality dataset were performed multivariate statistical analysis techniques including cluster analysis (CA), principal component analysis (PCA) and factor analysis (FA).

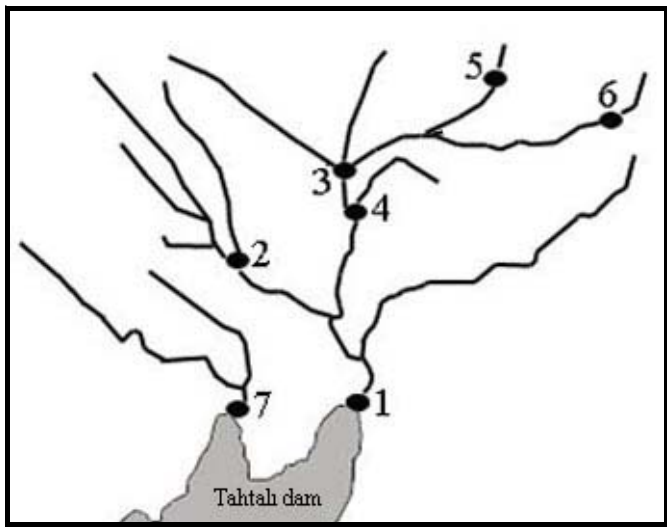

Fig. 1. Location of observation stations

TABLE-1
SURFACE WATER QUALITY DATA DESCRIPTION

Parameter	Symbol	Units
Chloride	Cl^-	mg/L
Nitrates	NO_3^-–N	mg/L
Ammonium	NH_4^+–N	mg/L
Oxygen saturation	OS	%
Colour	Col	Pt-Co
Sodium	Na^+	mg/L
Sulfate	SO_4^{2-}	mg/L
Total phosphorus	P-tot	mg/L
Dissolved oxygen	DO	mg/L
Biochemical oxygen demand	BOD_5	mg/L
Chemical oxygen demand	COD	mg/L

TABLE-2
DESCRIPTIVE STATISTICS OF WATER QUALITY DATA

	Cl^-	NO_3^-–N	NH_4^+–N	OS	Col	Na^+	SO_4^{2-}	P-tot	DO	BOD_5	COD
Mean	44.04	4.07	0.28	79.41	14.71	27.82	41.78	0.10	7.05	2.85	8.90
Median	38.00	3.10	0.00	79.70	10.00	20.75	40.00	0.00	7.01	2.00	8.00
Mode	36.00	3.00	0.00	75.00	10.00	15.00	45.00	0.00	6.00	2.00	8.00
SD	22.26	3.41	0.65	12.17	11.30	18.18	13.17	0.22	1.00	1.94	4.02
Variance	495.72	11.65	0.43	148.20	127.70	330.66	173.50	0.05	0.99	3.75	16.16

Cluster analysis (CA): Cluster analysis is an exploratory data analysis tool for solving classification problems. Its objective is to sort cases into groups or clusters, so that the degree of association is strong between members of the same cluster and weak between members of different clusters. Each cluster thus describes, in terms of the data collected, the class to which its members belong and this description may be abstracted through use from the particular to the general class type[21,22]. It is evident that the cluster analysis is useful in offering reliable classification of surface water in the whole region and would make possible to design a future spatial sampling strategy in an optimal manner. Thus, the number of observation stations in the monitoring network will be reduced, hence cost without loosing any significance of the outcome[3].

In this case of cluster analysis, the similarities-dissimilarities are quantified through Euclidean distance measurements, the distance between two objects, i and j, is given as:

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

where d_{ij}^2 donates the Euclidean distance, z_{ik} and z_{jk} are the values of variable k for object i and j, respectively and m is the number of variables[22]. Euclidean distance and the Ward method were used to obtain dendrograms.

Principal component analysis (PCA) and factor analysis (FA): Principal component analysis is designed to transform the original variables into new, uncorrelated variables (axes), called the principal components, which are linear combinations of the original variables. The new axes lie along the directions of maximum variance. Principal component analysis provides an objective way of finding indices of this type so that the variation in the data can be accounted for as concisely as possible. Principal component (PC) provides information on the most meaningful parameters, which describes a whole data set affording data reduction with minimum loss of original information[13]. The principal component can be expressed as in eqn. 2.

$$z_{ij} = a_{i1}x_{i1} + a_{i2}x_{i2} + a_{i3}x_{i3} + ... + a_{im}x_{im} \quad (2)$$

where z is the component score, a is the component loading, x the measured value of variable, i is the component number, j is the sample number and m the total number of variables, respectively.

The main purpose of factor analysis is to reduce the contribution of less significant variables to simplify even more of the data structure coming from principal component analysis[13]. This can be achieved by rotating the axis defined by principal component analysis, according to well established rules and constructing new variables, also called varifactors. Principal component is a linear combination of observable water quality variables, where as varifactors can include unobservable hypothetical, latent variables was performed to extract significant principal components and to

further reduce the contribution of variables with minor significance. As a result, a small number of factors will usually account for approximately the same amount of information as do the much larger set of original observations. The factor analysis can be expressed as in eqn. 3;

$$z_{jk} = a_{f1}f_{1i} + a_{f2}f_{2i} + a_{f3}f_{3i} + ... + a_{fm}f_{mi} + e_{fi} \quad (3)$$

where z is the measured variable, a is the factor loading, f is the factor score, e the residual term accounting for errors or other source of variation, i the sample number and m the total number of factors, respectively.

RESULTS AND DISCUSSION

In this study, cluster analysis was performed in order to determine the similarities between the observation stations by using surface water quality data. In addition to this, surface water quality data belonging to Tahtali dam and its tributaries were classified by using principal component analysis and factor analysis. These multivariate analyses were performed by using SPSS statistical software.

Cluster analysis was used to determine similarity groups between the observation stations. The dendrogram which was obtained by using cluster analysis is given in Fig. 2. It was observed from the dendrogram that surface water quality parameters were formed into two different clusters. Cluster I is formed of Stations 1, 3, 4, 5 and 6. These stations are located in the north and northeast sections of the Tahtali dam. Stations 1 is at the inlet of Tahtali dam reservoir and Stations 3, 4, 5 and 6 are remote points which feeding the inlet of reservoir. Cluster II is formed of Stations 2 and 7 and are located in the northwest section of the Tahtali dam reservoir. Stations 7 is at the inlet of Tahtali dam and Station 2 are remote points which feeding the inlet of Tahtali dam.

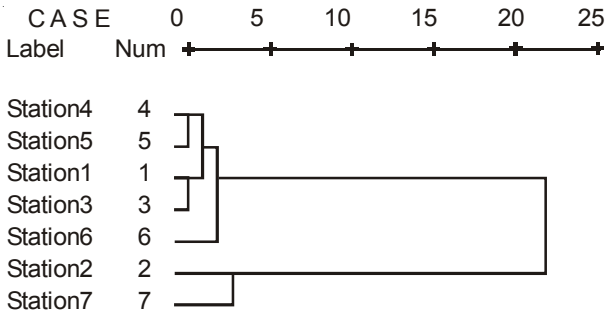

Fig. 2. Dendrogram of the Ward method

Principal component analysis was actually performed on the correlation matrix between the different parameters followed by varimax rotation, with the same being used to examine the relationship between them[23]. Principal component analysis screen plot is presented in Fig. 3. When this screen plot is examined, it is observed that the number of basic component is 2. Principal component analysis results are

given in Table-3. In view of this analysis, it was determined that PCA1 is composed of Stations 1, 3, 4, 5 and 6 whereas PCA2 is composed of Stations 2 and 7.

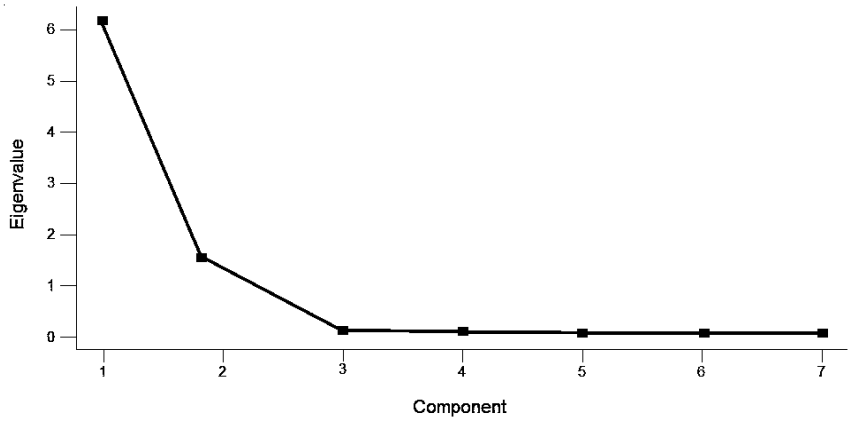

Fig. 3. Screen plot for the PCA by stations

TABLE-3
PRINCIPAL COMPONENT ANALYSIS (PCA) WEIGHTS

Station No.	PCA 1	PCA 2
Station 1	-0.391	–
Station 2	–	0.657
Station 3	-0.388	–
Station 4	-0.387	–
Station 5	-0.379	–
Station6	-0.372	–
Station 7	–	0.492

The factor analysis generated two significant factors, which explained *ca.* 94 % of the variance in observation stations dataset. The correlations matrix of observation stations was generated and factors extracted by the centroid method, rotated by varimax. Results obtained from factor analysis are given in Table-4.

TABLE-4
FACTOR LOADINGS FOR OBSERVATION STATIONS

Station No.	Factor 1	Factor 2
Station 1	0.876	–
Station 3	0.856	–
Station 4	0.828	–
Station 5	0.799	–
Station 6	0.763	–
Station 7	–	0.883
Station 2	–	0.825
% Total variance	88.340	5.660
% Cumulative variance	88.340	94.000

When the results of the analyses are examined, it is observed that principal component analysis and factor analysis support cluster analysis. it is observed that observation stations form 2 different groups. It was determined that, among these groups, Factor1 is composed of Stations 1, 3, 4, 5 and 6 whereas Factor 2 is composed of Stations 2 and 7. An assessment of the cause of the connected accumulate of stations 2 and 7 seen during analysis being different from that of the other stations may be supported by the different environmental effects that the stations are exposed to.

Human activities near the Tahtali dam site have had direct and indirect effects on the contamination rates of surface in the Tahtali watershed area. Direct effects include dissolution and transport of excess quantities of fertilizers with associated materials and hydrologic alterations related to irrigation and drainage. Indirect effects include changes in water-rock reactions in soils and aquifers caused by increased concentrations of dissolved oxidants, protons and major ions[14]. The results of the study revealed different properties. After the statistical grouping of the observation stations, water quality parameters were separately examined by using factor analysis. Factor loadings belonging to water quality parameters are given in Table-5. Liu *et al.*[24] presented the factor loading as strong, moderate and weak balancing to loading values of > 0.75, 0.75-0.50 and 0.50-0.30, respectively.

TABLE-5
MATRIX OF FACTOR LOADING BY WATER QUALITY PARAMETERS

Parameters	F1	F2	F3	F4
Cl^-	0.848	–	–	–
NH_4^+–N	0.790	–	–	–
Na^+	0.893	–	–	–
SO_4^{2-}	0.628	–	–	–
P-tot	0.682	–	–	–
NO_3^-–N	–	–	0.835	–
Col	–	–	-0.748	–
DO	–	0.813	–	–
OS	–	0.789	–	–
BOD_5	–	–	–	0.924
COD	–	–	–	0.483
% Total variance	29.46	14.32	11.97	9.60
% Cumulative variance	29.46	43.78	55.75	65.35

The first factor (F1) is related to the parameters Cl^-, NH_4^+–N, Na^+, SO_4^{2-}, P-tot and explained 29.46 % of the total variance. This factor represents pollution from domestic waste and nutrient. The second factor (F2) is positively loaded with parameters DO and OS. This factor accounts for 14.32 % of the total variance and is strongly and positively loaded with this factor. Factor 3 (F3) explained 11.97 % of the total variance and related to the parameters NO_3^-–N and Col. While the parameter Col is negatively loaded with this factor, NO_3^-–N is strongly and positively loaded with this factor. Factor 4 (F4) is related to the parameters BOD_5 and COD. COD is

widely used for determining waste concentration and is applied primarily to pollutant mixtures such as domestic, agricultural and industrial waste. BOD_5 is the local anthropogenic pollution and also addition of local domestic waste of this site[25]. The discharges of the surface water from many factors and especially from municipal, fertilizers and factories waste contribute to the pollution of the Tahtali dam and tributaries.

The data of the F1 were calculation into mean value to compare the aspects of the variation in surface water quality data collected from seven different sites as shown in Fig. 4. Among the mean value, all parameters were found to be high at Stations 2 and 7 showing high pollution of these sites. Thus, the figure is supported multivariate statistical analysis results.

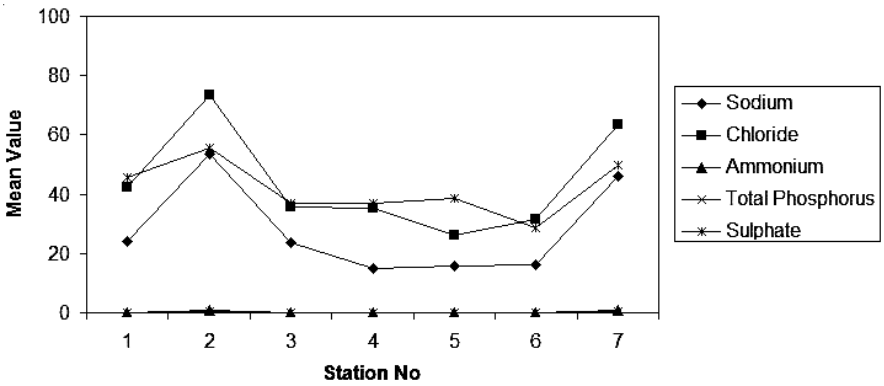

Fig. 4. Sodium, chloride, ammonium, total phosphorus and sulphate mean value at observation stations

Conclusion

Tahtali dam is an important drinking water source of Izmir (Turkey). The multivariate statistical analysis techniques including cluster analysis, principal component analysis and factor analysis generally confirmed surface water classification. In this study, multivariate statistical analysis techniques were applied to surface water quality dataset.

Cluster analysis, principal component analysis and factor analysis were used to classify observation stations. Cluster analysis results show that, Cluster II stations locations are at the northwest of the dam and Cluster I stations locations are at the north and northeast of the dam. Afterwards, these analyses were supported by principal component analysis and factor analysis. Based on the above results, it may be results that of the observation stations explained by the four factors, it is the F1 (Cl^-, NH_4^+–N, Na^+, SO_4^{2-}, P-tot) that best observed variances in the data.

The study enabled us to show similarities among the observation stations that were not clearly visible from an examination of the data in the tables. These results

represent that domestic waste and nutrient pollution caused differences in terms of water quality in the northwest part of the area. Finally, it was determined that the multivariate statistical analysis techniques usefulness for analysis and interpretation of water quality dataset.

ACKNOWLEDGEMENTS

The author would like to thank the Izmir Water and Sewerage Authority (IZSU) in Turkey for their help in providing necessary data. The author would also like to thank Prof. Dr. A. Sermet Anagun for their helpful contributions.

REFERENCES

1. S.R. Carpenter, N.F. Caraco, D.L. Corell, R.W. Howarth, A.N. Sharpley and V.H. Smith, *Ecol. Appl.*, **8**, 559 (1998).
2. H.P. Jarvier, B.A. Whitton and C. Neal, *Sci. Total Environ.*, **210/211**, 79 (1998).
3. K.P. Singh, A. Malik and S. Sinha, *Anal. Chim. Acta*, **538**, 355 (2005).
4. B.A. Helena, M. Vega, E. Barrado, R. Pardo and L. Farnandez, *Water Air Soil Pollut.*, **112**, 365 (1999).
5. S.M. Ahmed, M. Hussain and W. Abderrahman, *Saudi Arabia, Bull. Geol. Environ.*, **64**, 319 (2005).
6. Y. Shuxia, J. Shang, J. Zhao and H. Guo, *Water Air Soil Pollut.*, **144**, 159 (2003).
7. M.T. Hussein, *Hydrogeol. J.*, **12**, 144 (2004).
8. M. Vega, R. Pardo, E. Barrado and L. Deban, *Water Res.*, **32**, 3581 (1998).
9. J.Y. Lee, J.Y. Cheon, K.K. Lee, S.Y. Lee and M.H. Lee, *J. Environ. Qual.*, **30**, 1548 (2001).
10. D.A. Wunderlin, M.P. Diaz, M.V. Ame, S.F. Pesce, A.C. Hued and M.A. Bistoni, *Water Res.*, **35**, 2881 (2001).
11. R. Reghunanth, T.R.S. Murthy and B.R. Raghavan, *Water Res.*, **36**, 2437 (2002).
12. V. Simeonov, P. Simeonov and R. Tsitouridou, *Chem. Eng. Ecol.*, **11**, 449 (2004).
13. S. Shresthe and F. Kazama, *Environ. Modelling Software*, **22**, 464 (2007).
14. E. Ericksson, Principles and Application of Hydrochemistry, Chapman and Hall, London (1985).
15. K. Chen, J.J. Jiao, J. Huang and R. Huang, *Environ. Pollut.*, **147**, 771 (2007).
16. D.A. Spears and Y. Zheng, *Int. J. Coal Geol.*, **38**, 161 (1999).
17. G. Muri, *Acta Chim. Slov.*, **51**, 257 (2004).
18. V. Simeonov, J.A. Stratis, C. Samara, G. Zachariadis, D. Voutsa, A. Anthemidis, M. Sofoniou and T. Kouimtzis, *Water Res.*, **37**, 4119 (2003).
19. H. Elhatip, M.A. Hinis and N. Gulbahar, *Stoch. Environ. Res. Risk. Assess.*, **2**, 391 (2008).
20. I. Atis, Drinkable Water Supply for Izmir, Present, Past and Future, Water Congress, Izmir, Turkey (1999).
21. J.W. Einax, D. Truckenbrodt and O. Kampe, *Microchem. J.*, **58**, 315 (1998).
22. T. Kowalkowski, R. Zbytniewski, J. Szpejna and B. Buszewski, *Water Res.*, **40**, 744 (2006).
23. M.R. Kuppusamy and V.V. Grindhar, *Environ. Int.*, **32**, 174 (2006).
24. C. W. Liu, K.H. Lin and Y.M. Kuo, *Sci. Total Environ.*, **313**, 77 (2003).
25. A.M. Silva, E.L.B. Novelli, M.L. Fascinelli and J.A. Almeida, *Environ. Pollut.*, **105**, 243 (1999).

(*Received*: 30 October 2008; *Accepted*: 25 February 2009) AJC-7286