Surface water quality assessment using factor analysis

Hülya Boyacioglu

Dokuz Eylul University, Faculty of Engineering. Department of Environmental Engineering, Tinaztepe Campus Buca, 35160, Izmir, Turkey

Abstract

In this study, the factor analysis technique is applied to surface water quality data sets obtained from the Buyuk Menderes River Basin, Turkey, during two different hydrological periods. Results show that the indices which changed the quality of water in two seasons and locations differed. During low-flow conditions, water quality was strongly affected by agricultural uses. On the other hand, the main pollution source changed from agricultural uses to urban land uses in high-flow periods. Therefore major water pollution threats in the basin were urban and agricultural land uses which are defined as non-point sources. This technique is believed to assist decision makers in identifying priorities to improve water quality that has deteriorated due to various land uses.

Keywords: factor analysis, factor scores, Buyuk Menderes River, water quality

Introduction

Water quality monitoring has one of the highest priorities in environmental protection policy (Simeonov et al., 2002). The main objective is to control and minimise the incidence of pollutant-oriented problems, and to provide water of appropriate quality to serve various purposes such as drinking water supply, irrigation water, etc.

The quality of water is identified in terms of its physical, chemical and biological parameters (Sargaonkar and Deshpande, 2003). The particular problem in the case of water quality monitoring is the complexity associated with analysing the large number of measured variables (Saffran, 2001). The data sets contain rich information about the behaviour of the water resources. The classification, modelling and interpretation of monitoring data are the most important steps in the assessment of water quality.

Multivariate statistical methods including factor analysis have been used successfully in hydrochemistry for many years. Surface water, groundwater quality assessment and environmental research employing multi-component techniques are well described in the literature (Praus, 2005). Multivariate statistical approaches allow deriving hidden information from the data set about the possible influences of the environment on water quality (Spanos et al., 2003).

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). There are three stages in factor analysis (Gupta et al., 2005):

- For all the variables a correlation matrix is generated
- Factors are extracted from the correlation matrix based on the correlation coefficients of the variables
- To maximise the relationship between some of the factors and variables, the factors are rotated.

A first step is the determination of the parameter correlation matrix. It is used to account for the degree of mutually shared variability between individual pairs of water quality variables. Then, eigenvalues and factor loadings for the correlation matrix are determined. Eigenvalues correspond to an eigenfactor which identifies the groups of variables that are highly correlated among them. Lower eigenvalues may contribute little to the explanatory ability of the data. Only the first few factors are needed to account for much of the parameter variability. Once the correlation matrix and eigenvalues are obtained, factor loadings are used to measure the correlation between the variables and factors. Factor rotation is used to facilitate interpretation by providing a simpler factor structure (Zeng and Rasmussen, 2005).

This study evaluated the possibility that a smaller group of water quality parameters/ locations might provide sufficient information for water quality assessment. Factor analysis was applied to a surface water quality data set collected from Buyuk Menderes Basin, Turkey using 'the *Statistical Package for the Social Sciences Software-SPSS 10.0 for Windows*'. Water quality monitoring was conducted at 21 stations in the study area during low- and high-flow periods. The selected parameters for the estimation of surface water quality characteristics were: electrical conductivity (EC), total dissolved solids (TDS), sodium (Na^+), potassium (K^+), calcium (Ca^{2+}), magnesium (Mg^{2+}), sulphate (SO_4^{2-}), nitrate-nitrogen (NO_3-N), Kjeldahl Nitrogen, biochemical oxygen demand (BOD_5) and chemical oxygen demand (COD). COD measurements were performed using the potassium dichromate method.

Study area

The Buyuk Menderes River Basin is located in Western Anatolia and covers Uşak, Aydın and Denizli Provinces with a total land area of about 25 000 km^2 (Fig. 1). The basin is endowed with one of the most fertile soils in the country and the economy of the region is heavily dependent on agricultural production. In addition, rapid industrialisation and population growth over the past few decades have created additional stress on the environmental conditions in the region (Boyacioglu et al., 2004). The

☎ +90 232 412 71 31; fax: +90 232 453 11 43;
e-mail: hulya.boyacioglu@deu.edu.tr

Received 16 January 2006; accepted in revised form 24 April 2006.

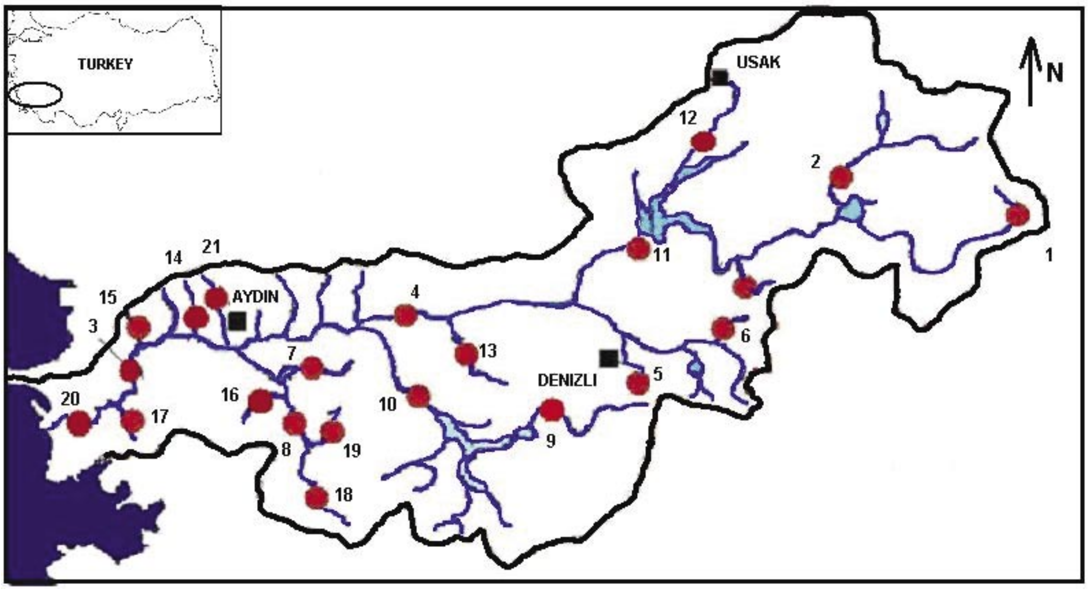

Figure 1
Sketch map of water quality monitoring stations in Buyuk Menderes Basin in Turkey

population of the basin is about 2 500 000 as of the year 2000, living in more than 320 municipalities and settlements, 65% of which have proper sewage systems with only about 12% of them treating their wastewater prior to discharge (State Institute of Statistics, 2005). In this regard, the study area has been subject to increasing rates of pollution originating mainly from anthropological activities.

The pollution sources of Buyuk Menderes River can be organised into three groups:

- Point discharges
- Non-point source contributions
- Other sources

Point discharges originate from either domestic or industrial polluters. While some of these discharges are made to the river after proper treatment, in many cases no treatment is applied prior to the discharge. The basic sources of non-point source pollution in the basin include the diffused transport of contaminants to river channels originating from agricultural practices. In addition, there are also other sources of pollution that degrade the quality of surface waters in Buyuk Menderes River including transport of eroded land, leachates from mining activities and solid waste disposal sites (Boyacioglu et al., 2004).

Assessment of water quality

Low-flow period

As was mentioned above, one of the most fertile soils in the country is found in the Buyuk Menderes Basin. In this region, the economy is heavily dependent on agricultural production and also industrial activities, which are concentrated in the Aydin and Denizli Provinces. The climate of the region is typically Mediterranean: hot and dry in summer and temperate and rainy in winter. So, hydrological conditions of the river during the summer and winter periods are quite different. Thus, assessment of the water quality separately for summer (low flow) and winter (high flow) periods will assist in understanding the main pollutants, their sources and also determining priorities to improve water quality in two different hydrological periods.

Firstly, factor analysis was applied to data sets obtained during the low-flow period (between June-August). Descriptive statistics of the data set are presented in Table 1.

The correlation matrix of variables was generated and factors extracted by the Centroid method, rotated by Varimax rotation (Ahmed et al., 2005). Calculated eigenvalues, per cent

TABLE 1
Descriptive statistics of water quality data under low-flow conditions

Variable	Unit	Number of data	Mean	Median	Std. deviation	Variance	Coeff. of variance	Minimum	Maximum
EC	µS/cm	17	1027.06	660.00	831.95	692147.10	0.81	160.00	2750.00
TDS	mg/ℓ	17	663.53	420.00	531.99	283011.80	0.80	100.00	1760.00
Na^+	mg/ℓ	17	85.24	24.80	129.66	16810.57	1.52	3.60	434.00
K^+	mg/ℓ	17	8.94	4.40	11.96	143.16	1.34	1.30	52.40
Ca^{++}	mg/ℓ	17	68.49	56.10	39.61	1568.79	0.58	20.00	190.40
Mg^{++}	mg/ℓ	17	46.00	43.60	38.82	1506.75	0.84	1.20	147.10
SO_4^{2-}	mg/ℓ	17	163.02	81.20	212.34	45089.62	1.30	24.00	710.60
NO_3-N	mg/ℓ	17	2.29	2.26	1.75	3.06	0.76	0.00	5.25
Kjeldahl-N	mg/ℓ	17	0.83	0.60	0.54	0.29	0.65	0.30	1.80
BOD_5	mg/ℓ	17	6.42	5.50	3.97	15.74	0.62	2.10	17.60
COD	mg/ℓ	17	31.12	32.00	16.68	278.11	0.54	6.00	68.00

TABLE 2
Factor loading matrix and total variance explained (low-flow conditions)

Variable	Factor		
	1	2	3
EC	0.7230	0.6750	0.0770
TDS	0.7280	0.6720	0.0646
Na^+	0.2740	0.9350	-0.0231
K^+	0.0356	0.9670	0.0416
Ca^{++}	0.9380	0.0054	0.1200
Mg^{++}	0.9120	0.2830	0.0868
SO_4^{2-}	0.9010	0.2850	0.0297
NO_3-N	0.4140	0.4310	-0.6000
Kjeldahl-N	0.0605	0.2440	0.7740
BOD_5	0.5980	-0.1220	0.6720
COD	0.1270	-0.0184	0.8120
Eigenvalue	4.20	3.14	2.11
% total variance	38.21	28.54	19.13
Cumulative %	38.21	66.75	85.88

total variance, factor loadings and cumulative variance are given in Table 2.

The factor analysis generated three significant factors which explained 85.9% of the variance in data sets. The following factors were indicated considering the hydrochemical aspects of the water:

- Factor 1: $Ca^{2+,}$ Mg^{2+}, SO_4^{2-}
- Factor 2: Na^+ and K^+
- Factor 3: COD, BOD_5, Kjeldahl –N, NO_3-N

Ca^{2+}, Mg^{2+}, and SO_4^{2-} marked Factor 1 (F1) explained 38.2% of the variance. Na^+ and K^+ were correlated with Factor 2 (F2) and COD, BOD_5, Kjeldahl –N, NO_3-N with factor 3 (F3). The F1 had a high positive loading in $Ca^{2+,}$ Mg^{2+} and $SO_4^{2-,}$ which were 0.93, 0.91 and 0.90.

Urbanisation influences the water cycle through changes in flow and water quality. Urban land use (Na^+, K^+, Cl^-) may be differentiated from other land uses such as agriculture (Ca^{2+}, Mg^{2+}), through the use of biogeochemical fingerprints (Lindeman, 2004). Salts that are commonly found in subsurface drainage water include sulphates, chlorides, carbonates, and bicarbonates of calcium, and magnesium. Tail water also may contain these salts, but generally in much lower concentrations than in drainage water (Jacobsen and Basinal, 2004). Based on the results of the factor analysis and typical sources of water pollutants, it is concluded that F1 can be denoted as the 'agricultural use' factor with presence of Ca^{2+}, Mg^{2+}. As was mentioned above these parameters are mainly found in agricultural drainage water. F2 is strongly correlated with Na^+ and K^+, assigned as the 'urban land-use' factor. Factor loadings were 0.94 and 0.98**.** COD, BOD_5, Kjeldahl –N are included in F3 and are indicators of organic pollution in water, so F3 represents the 'organic pollution' factor.

In summary, three factors representing three different processes are:

- Urban land-use factor
- Agricultural use factor
- Organic pollution factor.

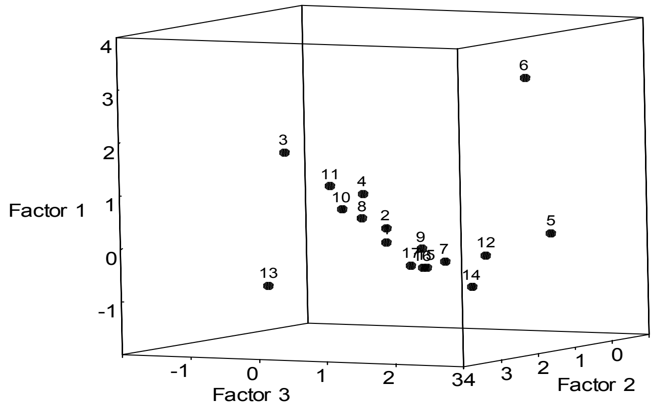

Figure 2
The loading plot of factor scores in low-flow period

Negative factor loading of NO_3-N explained the disproportion between this parameter and F3. COD, BOD_5 and Kjeldahl-N which were correlated with F3, decreased with increasing NO_3-N concentration which was caused by the nitrification process in water.

Therefore, the water quality of the Buyuk Menderes River during the low-level period was mainly controlled by agricultural pollutant sources. The loading plot of factor scores is shown in Fig. 2. Considering the location of the monitoring stations, given in Fig. 1, and the distribution of factor scores, it is concluded that:

- **Factor 1**: Low factor scores of F1 (agricultural use factor) were observed in the west of the basin. The middle and eastern parts where high values were monitored were faced with pollution risks originating from agricultural uses.
- **Factor 2**: High factor scores (urban land-use factor) were obtained in the north-west and also in the regions where population density is relatively high (especially in the centre of the provinces and their surroundings).
- **Factor 3**: F3 (organic pollution factor) scores were distributed in the basin almost uniformly. Depending on the presence of infrastructure and wastewater treatment efficiency, highest and lowest scores were observed even at the stations located next to each other. So, the settlements having no treatment plants increased the organic pollution risk.

High-flow period

The high-flow period may have positive effects with dilution of surface water by rain and stormwater. On the other hand, runoff water increases pollutant concentrations, thereby decreases quality. To assess the water quality of the Buyuk Menderes River under high-flow conditions, factor analysis was applied to data sets obtained from 21 monitoring stations between November-January. Descriptive statistics of the data are presented in Table 3.

Results of factor analysis including factor-loading matrix, eigenvalues and total and cumulative variance values are given in Table 4.

Three factors that are indicated below explained 81.33% of total variance.

- Factor 1: K^+, Na^+, TDS, EC,
- Factor 2: Ca^{2+}, Mg^{2+}, SO_4^{2-},
- Factor 3: COD, BOD_5, Kjeldahl-N.

It is suggested that, F1 represents the urban land-use character-

TABLE 3
Descriptive statistics of water quality data (high-flow conditions)

Variable	Unit	Number of data	Mean	Median	Std. deviation	Variance	Coeff. of variance	Minimum	Maximum
EC	µS/cm	32	1129.38	780.00	812.00	659348.00	0.72	160.00	3400.00
TDS	mg/ℓ	32	738.13	495.00	532.68	283744.80	0.72	100.00	2210.00
Na^+	mg/ℓ	32	101.88	27.05	142.27	20240.78	1.40	2.50	568.00
K^+	mg/ℓ	32	10.84	5.20	13.33	177.72	1.23	2.20	52.40
Ca^{++}	mg/ℓ	32	82.87	73.15	45.74	2092.54	0.55	20.00	230.50
Mg^{++}	mg/ℓ	32	45.36	46.80	28.35	803.99	0.63	0.60	115.50
SO_4^{2-}	mg/ℓ	32	157.43	67.60	165.98	27550.30	1.05	8.60	541.00
NO_3-N	mg/ℓ	32	1.82	1.38	1.46	2.13	0.80	0.00	5.25
Kjeldahl-N	mg/ℓ	32	0.93	0.80	0.54	0.29	0.58	0.10	2.40
BOD_5	mg/ℓ	32	5.62	5.05	3.26	10.64	0.58	2.00	16.50
COD	mg/ℓ	32	46.31	36.00	38.28	1465.51	0.83	4.00	180.00

TABLE 4
Factor loading matrix and total variance explained (high-flow conditions)

Variable	Factor		
	1	2	3
EC	0.8400	0.5010	0.1240
TDS	0.8520	0.4990	0.1260
Na^+	0.9590	0.1320	0.1720
K^+	0.9620	-0.0296	0.0720
Ca^{++}	0.1770	0.8560	0.1800
Mg^{++}	0.2680	0.8500	-0.0770
SO_4^{2-}	0.5360	0.6430	-0.1090
NO_3-N	-0.0357	0.4430	0.1260
Kjeldahl-N	0.5980	0.0519	0.6690
BOD_5	0.2590	0.0528	0.8670
COD	-0.1070	0.1480	0.9470
Eigenvalue	4.11	2.61	2.23
% Total variance	37.33	23.74	20.26
Cumulative %	37.33	61.07	81.33

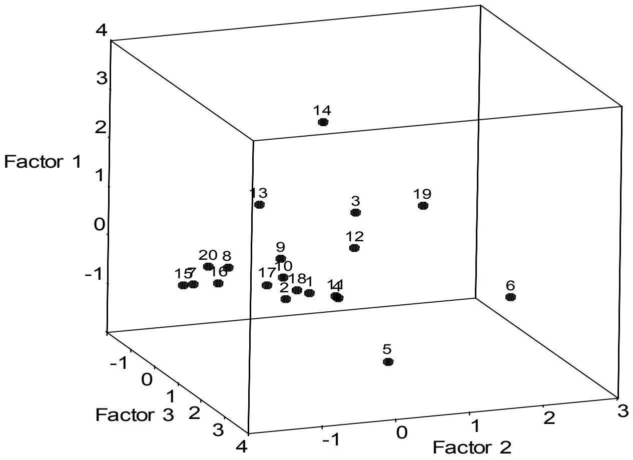

Figure 3
The loading plot of factor scores during the high-flow period

istics shown by presence of K^+ and Na^+. This factor explained 37.33 % of variance. F2 is strongly correlated with Ca^{2+} and Mg^{2+} which are mainly originated from agricultural uses. F3 was marked by BOD_5, COD and Kjeldahl-N. Thus, urban land use was the major pollution source in this hydrological period.

For each section, factor scores are shown in Fig. 3. Considering distribution of factor scores and locations of the monitoring stations, it is concluded that:

- **Factor 1**: High factor scores of F1 (urban land-use factor) were observed at the northwest part, downstream of the basin.
- **Factor 2**: Relatively high values of agricultural use factor (F2) obtained in the middle of the basin, where agriculture is the most important economic activity. Low scores were monitored at the west part.
- **Factor 3**: Low and high scores of organic pollution factor (F3) were distributed in the basin, because F3 depends on point pollution sources and is affected by infrastructure (sewage network and treatment plants) of the settlements.

Conclusions

The factors indicative of water quality in different hydrological periods and locations differed in Buyuk Menderes Basin. Under high-flow conditions pollutants mainly originated from urban land use and 37.3% of total variance was explained by the urban land-use factor. On the other hand water quality was controlled by agricultural pollutant sources during the low-flow period. Although the agricultural use factor explained 38.2% of the variance, for the land-use factor, it was only 28.5 under dry weather conditions. So, the major pollutant source changed from urban land uses to agricultural uses during the low-flow period. The main reason for this was the negative effect of runoff to surface water quality, because the storage ability, the buffering capacity of roads and buildings to rain or stormwater in urban areas, had been drastically weakened.

Thus, major pollution threats in low- and high-flow periods were urban and agricultural land uses which are defined as non-point pollution sources. Therefore priority should be given to minimisation of these sources to improve water quality in the basin.

This study shows that factor analysis is a useful method that could assist decision makers in determining the extent of pollution via practical pollution indicators. It could also provide a crude guideline for selecting the priorities of possible preventative measures in the proper management of the surface water resources of the basin (Boyacioglu et al., 2004).

References

AHMED S, HUSSAIN M and ABDERRAHMAN W (2005) Using multivariate factor analysis to assess surface/logged water quality and source of contamination at a large irrigation project at Al-Fadhli, Eastern Province. Saudi Arabia. *Bull. Eng. Geol. Env.* **64** 315-232.

BOYACIOĞLU H, BOYACIOĞLU H and GUNDUZ O (2004) Application of factor analysis in the assessment of surface water quality in Buyuk Menderes River Basin. *Proc. EWRA Symp. on Water Resources Management Risks and Challenges for the 21st Century* **2**.

GUPTA AK, GUPTA SK and PATIL RS (2005) Statistical analyses of coastal water quality for a port and harbour region in India. *Environ. Monit. Asses.* **102** 179-200.

JACOBSEN T and BASINAL L (2004) *A Landowner's Manual. A Guide for Developing Integrated On-Farm Drainage Management Systems.* California State Water Resources Control Board. Available at: www.cati.csufresno.edu/cit

LINDEMAN MA (2004) Exploring the effects of urban and agricultural land use on surface water quality. *2004 Denver annual meeting.* Paper No, 72-9. *Geological Society of America Abstracts with Programs.* **36** 184.

STATE INSTITUTE OF STATISTICS (2005) Republic of Turkey, Prime Ministry, State Institute of Statistics. 2003 Municipal Sewerage Questionnaire. Ankara.

STATE INSTITUTE OF STATISTICS (2004*) Turkey`s Statistical Yearbook 2004.* Ankara.

PRAUS P (2005) Water quality assessment using SVD-based principal component analysis of hydrological data. *Water SA* **31** 417-422.

SAFFRAN K (2001) *Canadian water quality guidelines for the protection of aquatic life, CCME water quality Index 1, 0. User`s manual.* Excerpt from Publication No.1299, ISBN 1-896997-34-1.

SARGAONKAR A and DESHPANDE V (2003) Development of an overall index of pollution for surface water based on a general classification scheme in Indian context. *Environ. Monit. Assess.* **89** 43-67.

SIMEONOV V, EINAX JW, STANIMIROVA I and KRAFT J (2002) Environmetric modeling and interpretation of river water monitoring data. *Anal. Bional. Chem.* **374** 898-905.

SPANOS T, SIMEONOV V, STRATIS J and XRISTINA X (2003) Assessment of water quality for human consumption. *Microchim. Acta.* **141** 35-40.

SPSS-10 (2000) *Statistical Package for the Social Sciences.* SPSS Inc, Chicago, USA.

YU S, SHANG J, ZHAO J and GUO H (2003) Factor analysis and dynamics of water quality of the Songhua River Northeast China. *Water, Air Soil Pollut.* **144** 159-169.

ZENG X and RASMUSSEN TC (2005) Multivariate statistical characterization of water quality in Lake Lanier, Georgia, USA. *J. Environ. Qual.* **34** 1980-1991.