*DOI: http://doi.org/10.21698/simi.2017.0025*

# THE USING OF PRINCIPAL COMPONENT ANALYSIS FOR THE ASSESSMENT OF WATER QUALITY IN KIRMIR BASIN

Ozlem Tunc Dede[1], Filiz Yazıcı Mısır[2], İlker T. Telci[3]

[1]Faculty of Engineering, Environmental Engineering Department, Giresun University, Turkey, *ozlem@tuncdede.com*
[2]Analysis and Reporting Department, Public Hospital Agency of Turkey, Turkey, *filiz.misir@saglik.gov.tr*
[3]Bechtel Oil, Gas and Chemicals Inc. Houston Texas, USA, *ilkertelci@gmail.com*

## Abstract

Water is one of the most important nutrient for human and also for aquatic organism. Poor water quality has adverse effects on human health and aquatic life. Hence, protecting the water resources from pollutants and the monitoring of water quality is important. In recent years, some kind of methods such as water quality index model, regression analysis, factor analysis, principal component analysis, cluster analysis, etc. has been developed for easy assessment and interpretation of large amount of water quality data. Among these methods, multivariate statistical analysis has an advantage of giving an idea about possible sources of pollution. In this study, the assessment of water quality of 10 different sampling station in Kirmir Basin which is one of the most significant drinking water resources of Ankara, the capital city of Turkey has been investigated by using multivariate statistical methods (principal component analysis-PCA and factor analysis-FA). 18 water quality parameters were analysed for each sampling station and used for the statistical analysis. The correlations between parameters and sampling stations were evaluated by using statistical techniques in terms of underlying factors. FA/PCA identified water quality parameters in five groups. The results revealed that Kirmir Basin was mainly affected from agricultural activities, urban land uses and livestock activities. The improving of the water quality in this region can be achieved by controlling these activities.

**Keywords:** *factor analysis, multivariate statistic, water pollution*

## Introduction

Water is an essential nutrient for living organisms. The amount of freshwater resources is very limited in the world. However, water resources are open to pollution due to the growth of population, technological development and increasing industrial activities. Pollutants may disrupt the aquatic ecosystems at a certain concentration. Thus, preserving the quality of water is crucial for the wellbeing of humanity and aquatic organisms. Water quality can be defined by using physico-chemical and biological parameters. Measuring the physico-chemical and biological parameters and comparing the results with the limit values given in standards/
guidelines is a general approach during water quality assessment. However, this approach is very exhaustive and also time consuming due to the evaluation of large number of measured variables (Akkoyunlu & Akiner 2012; Sánchez et al. 2007; Tezcanli Guyer & Genc Ilhan 2011; Tunc Dede et al. 2013).

Different kinds of methods have been developed for easy assessment and interpretation of water quality data such as water quality index methods and statistical multivariate analysis (regression analysis, factor analysis, principal component analysis, cluster analysis, etc.) (Akkoyunlu & Akiner 2012; Boyacioglu & Boyacioglu 2007; Malik & Nadeem 2011; Mazlum et al. 1999).

Among these methods, statistical multivariate analysis for data classification and modeling has an advantage of getting additional information about possible sources of pollution in addition to its capability of handling large amount of raw data (Boyacioglu & Boyacioglu 2007). Principal component analysis (PCA) and factor analysis (FA) aim to explain the relationship between the parameters in terms of the underlying factors, which are not directly observable. These analyses deal with the grouping of the correlated parameters and calculation of the new factors based on data set (Boyacioglu & Boyacioglu 2007). The only difference between PCA and FA is that PCA assumes that all variance is common and all unique factors set equal to zero while FA assumes some unique variance. Parameters can be either positively or negatively correlated. Parameters that are highly correlated means they are influenced by the same factors. PCA provides data reduction by selecting the parameters which represent the whole dataset and summarize the statistical correlation between parameters in water while keeping the original data loss at a minimum level. The main objective of the factor analysis is to determine the common factors affecting the parameters in the data set and to find the relation between the factors and the parameters (Boyacioglu & Boyacioglu 2007).

In this study, the water quality assessment of Kirmir Basin was studied and water quality data was evaluated by using factor analysis (FA) based on principal component analysis (FA/PCA).

**Materials and Methods**

Kirmir is among the most important branches of Sakarya River. It is third longest basins near Ankara, the capital city of Turkey and located between districts of Kızılcahamam and Beypazarı. 67% of the drinking water of Ankara is supplied by the main resources (Çamlıdere, Eğrekkaya and Akyar Dams) located in this basin. However, Kirmir basin is open to pollution due to livestock and agricultural activities, urban land uses, human activities and geothermal facilities (Tunc Dede et al. 2013; DEIAP 2016). For the assessment of surface water quality in Kirmir, water samples were collected on a monthly base from 10 different stations in this area for one-year period from June 2009 to May 2010. Sampling was conducted in accordance with "Standard Methods 1060 Collection and Preservation of Samples" (Eaton & Clescen, 2005). The analysis of water quality parameters was conducted according to Turkish Standards, Standard Methods and U.S. Environmental Protection Agency (EPA) methods. Factor and principal component analyses were employed to evaluate water quality data. Statistical analyses were done using Excel 2016 (Microsoft, 2016) and STATISTICA version 6.0 (SPSS, StatSoft, Inc., USA, 2001) software. The selected parameters for the estimation of surface water quality characteristics were: pH, electrical conductivity (EC), dissolved oxygen (DO), sodium (Na), potassium (K), calcium (Ca), magnesium (Mg), chloride (Cl), sulphate ($SO_4$), biological oxygen demand (BOD), Total coliforms, *Escherichia coli*, manganese (Mn), arsenic (As), boron (B), titanium (Ti), vanadium (V) and barium (Ba).

**Results and Discussion**

In this study, multivariate statistical techniques including factor and principal component analysis were applied to data set obtained from Kirmir Basin and its tributaries in Ankara, Turkey. Selected parameters and their descriptive statistics are given in Table 1.

**Table 1**. Selected parameters for water quality evaluation and their descriptive statistics

| | Unit | Mean | Median | Mode | Std. dev. | Variance | Min. value | Max. value | Limit values |
|---|---|---|---|---|---|---|---|---|---|
| pH | - | 8.16 | 8.20 | 8.20 | 0.47 | 0.22 | 7.17 | 9.17 | 6.5 ≤ ≤ 9.5 |
| EC | μS | 456 | 423 | 549 | 247 | 60985 | 77 | 1210 | 2500 |
| DO | mg $O_2$ | 8.31 | 7.80 | 7.70 | 1.18 | 1.40 | 6.50 | 11.80 | >8[a1] |
| Na | mg $L^{-1}$ | 35.65 | 22.11 | 11.40 | 37.41 | 1400 | 3.20 | 188 | 200 |
| K | mg $L^{-1}$ | 6.99 | 5.94 | 4.60 | 4.49 | 20.15 | 1.40 | 23.97 | 3700[a2] |
| Ca | mg $L^{-1}$ | 30.22 | 25.95 | 18.40 | 19.2 | 366 | 0.17 | 97.47 | 1000[a3] |
| Mg | mg $L^{-1}$ | 19.26 | 16.30 | 5.60 | 13.41 | 180 | 2.10 | 62.86 | 200[a3] |
| Cl | mg $L^{-1}$ | 20.41 | 12.05 | 2.00 | 22.14 | 490 | 0.80 | 90.20 | 200[a1] |
| $SO_4$ | mg $L^{-1}$ | 15.29 | 13.37 | 12.00 | 10.38 | 108 | 2.76 | 59.06 | 250 |
| BOD | mg $O_2$ | 10.53 | 6.10 | 5.00 | 9.87 | 97.41 | 2.80 | 59.10 | <3[a1] |
| Total coliforms | no/100 mL | 1483 | 800 | 3000 | 1937 | 3752790 | 50 | 9600 | 50[a1] |
| *Escherichia coli* | no/100 mL | 194 | 11 | 0 | 608 | 370238 | 0 | 3200 | 20[a1] |
| Mn | μg $L^{-1}$ | 17.15 | 10.03 | n.a. | 18.69 | 349 | 1.17 | 87.35 | 50 |
| As | μg $L^{-1}$ | 25.34 | 18.28 | n.a. | 25.37 | 643 | 0.86 | 127 | 10 |
| B | μg $L^{-1}$ | 390 | 168 | n.a. | 551 | 303379 | 11.87 | 3222 | 1000 |
| Ti | μg $L^{-1}$ | 10.39 | 2.63 | n.a. | 19.90 | 396 | 0.56 | 88.81 | n.a. |
| V | μg $L^{-1}$ | 6.03 | 4.96 | 3.62 | 3.58 | 12.83 | 1.39 | 17.75 | n.a. |
| Ba | μg $L^{-1}$ | 73.55 | 76.13 | n.a. | 37.72 | 1423 | 20.36 | 197 | 100[a1] |

*Notes*: All limits values were given based on TS 266 except: [a1] (TurkishRegulation 2012); [a2] (WHO 2009); [a3] (WHO 2011); no: number; n.a.: not available

In Table 1, standard deviation for some parameters are at high values which demonstrate the variability during the sampling periods due to the anthropogenic effects and variation in precipitation and flow. In order to evaluate mean differences among annual mean values, standard deviation, minimum and maximum values of parameters for all sampling stations, analysis of variance (ANOVA) was employed. Covariance values were found higher than 0.5 which means that the relationship between these 18 parameters are important and application of factor analysis is meaningful. Water quality parameters were grouped by using factor analysis. Factor and principal component analysis was applied in three phases:

1. Preparation of the correlation matrix for all parameters,
2. Extraction of the initial set of factors by using an extraction method (centroid, maximum likelihood, principal component and principal axis extraction, etc.),
3. The rotation of the factors, aiming for increasing the relationship between some of the factors and parameters for simple and easy interpretation. Varimax procedure was

used for the rotation. Varimax factors with eigenvalue 1 were retained. After rotation, the correlation between each parameter and each factor was obtained (Boyacioglu & Boyacioglu 2007; Malik & Nadeem 2011).

The correlation matrix of parameters was generated and five factors were extracted with principal component analysis method rotated by Varimax with Kaiser Normalization. The variance statement ratio of the factors has been optimized so as to be closer to each other by applying six rotations. Results of factor analysis including factor-loading matrix, eigenvalues, total and cumulative variances are presented in Table 2.

**Table 2.** Factor-loading matrix, eigenvalues and total and cumulative variance values (Marked loadings > 0.5)

| Parameter | Factor 1 | Factor 2 | Factor 3 | Factor 4 | Factor 5 |
|---|---|---|---|---|---|
| Na | **0.922** | 0.017 | -0.079 | 0.026 | -0.037 |
| B | **0.914** | -0.120 | 0.023 | 0.121 | 0.085 |
| As | **0.881** | 0.015 | -0.069 | 0.117 | 0.232 |
| EC | **0.841** | 0.151 | -0.157 | 0.287 | 0.287 |
| Cl | **0.812** | -0.043 | -0.092 | 0.152 | 0.046 |
| K | **0.745** | 0.199 | -0.135 | 0.225 | -0.134 |
| Ba | **0.582** | 0.252 | -0.179 | 0.196 | 0.258 |
| V | 0.148 | **0.898** | -0.031 | -0.060 | 0.005 |
| DO | -0.015 | **0.891** | -0.119 | 0.043 | -0.089 |
| *Escherichia coli* | -0.009 | 0.016 | **0.802** | -0.097 | 0.061 |
| Ti | -0.202 | -0.129 | **0.758** | -0.218 | -0.251 |
| pH | 0.201 | 0.222 | **-0.698** | -0.191 | -0.217 |
| Mn | 0.194 | -0.005 | -0.053 | **0.729** | -0.196 |
| Mg | 0.183 | **0.535** | -0.19 | **0.642** | 0.228 |
| $SO_4$ | **0.505** | -0.134 | -0.048 | **0.597** | 0.224 |
| BOD | 0.371 | -0.480 | 0.128 | **0.520** | 0.175 |
| Ca | 0.252 | -0.127 | -0.187 | -0.082 | **0.780** |
| Total coliforms | 0.040 | 0.036 | 0.406 | 0.093 | **0.706** |
| Eigenvalue | 5.356 | 2.363 | 2.074 | 1.908 | 1.632 |
| Total variance% | 29.75 | 13.13 | 11.52 | 10.60 | 9.06 |
| Cumulative variance % | 29.75 | 42.88 | 54.40 | 65.00 | 74.06 |

To confirm the results obtained during factor analysis, principal component analysis was applied to the water quality parameters data set. A scree plot showing the sorted eigenvalues as a function of the principal components number is given in Fig. 1.

As seen from the figure, PCA generated five significant components with eigenvalues greater than 1 (according to the "eigenvalue-one" criterion) (Azhar et al. 2015). The breakpoint in the graph (component:5, eigenvalue:1) will also show us the number of the factors. And the group components are the same as in the factor analysis. Next, five significant factors which explained 74.06% of total variance of original variables were generated by using the factor analysis (Table 2):

Factor 1 (FAC1): Na, B, As, EC, Cl, K, Ba, $SO_4$

Factor 2 (FAC2): V, DO, Mg

Factor 3 (FAC3): *Escherichia coli*, Ti, pH

Factor 4 (FAC4): Mn, Mg, $SO_4$, BOD

Factor 5 (FAC5): Ca, Total coliforms

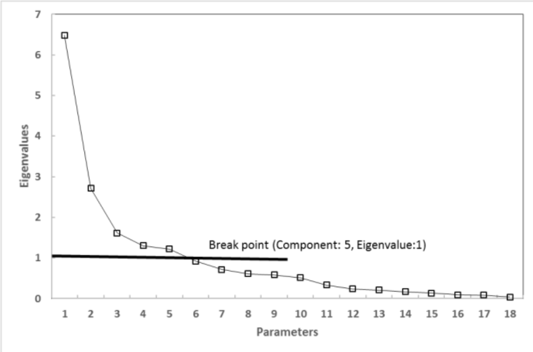


**Figure 1**. Scree plot of the eigenvalues

Based on factor loadings, FAC1 explained 29.75% of the variance and was strongly correlated with Na, B, As, EC, Cl, K, Ba and $SO_4$. The elements Na, K and Cl are essential for living organisms and they are among the most abundant elements on earth. Water may expose to serious amounts of Na, K and Cl due to the urban land use. The higher amounts of these elements can cause serious health problems. For example, the excess amount of chloride may result in hearth and kidney problems and also corrosion in the pipes of water distribution systems (TSE 1998; WHO 2011). Atmospheric activities, deposition from sedimentary rocks, sewage effluents and agricultural runoffs can be the reason of Cl in surface waters. Anthropogenic activities may increase K amount in water.

As is also an abundant element in Earth's crust and can be found in the form of sulfides and metal arsenide or arsenates. Exposure to high concentrations of arsenic can cause widespread health effects in humans. Volcanic rocks can be the reason increasing in arsenic amount in water. B occurs naturally in groundwater due to the rocks and soils containing borates and borosilicate. Surface water can be contaminated with boron as a result of wastewater discharge. Excess amounts can cause serious health problems (Malik & Nadeem 2011, TSE 1998, WHO 2011).

EC is an indirect measurement of total dissolved solids (TDS) which often is associated to waste water discharge and increase in dissolved salt amount cause increase in EC. Unexpected increase in EC is an indicator of the anthropogenic pollution in water (Akkoyunlu & Akiner 2012; Akoteyon et al. 2011).

$SO_4$ can be found naturally in the environment at high concentrations. Soil and industrial effluents can cause the sulfate contamination in water. High amounts of sulfate in drinking water can cause bowel problems in humans and also bad taste in water (Akoteyon et al. 2011).

High amount of Ba contributes to hardness of water. The main sources of hardness in water is due to the sedimentary rocks and also some runoff from soils. Ca and Mg are the principal elements found in many sedimentary rocks and can cause hardness in water, too. High values of hardness cause aesthetic problems in water (WHO 2011; TSE 2000).

In general Cl and $SO_4$ indicate some biological and anthropogenic activities in the environment (Avtar et al. 2011).

FAC2 was marked by V, DO and Mg. This factor explained 13.13% of the total variance. DO indicates inorganic pollution in the water. DO is one of the important parameters for aquatic life. Water containing high amounts of organic and inorganic substances results in oxygen depletion due to oxygen consumption (Malik & Nadeem 2011; Boyacioglu & Boyacioglu 2007). FAC3 explained 11.52% of variance and related with the parameters *Escherichia coli*, Ti and pH. The bacteriological monitoring (*Escherichia coli*, Total coliforms, etc.) is one of key components in water quality evaluation. Livestock activities, animal wastes and domestic waste water discharges may cause bacteriological pollution in water. pH is another important factor as an operational parameter. While it was reported that pH has no direct effect on consumers, pH level in surface water may affect respiration in aquatic life (WHO 2011).

FAC4 was correlated with the parameters of Mn, Mg, $SO_4$ and BOD. 10.60% of the total variance was explained by this factor. BOD is another important parameter for the determination of the quality of water and related with the oxygen demanding capacity of organic materials. It indicates organic pollution. High BOD concentration may cause death of fish and increase eutrophication (Akkoyunlu & Akiner 2012; Cude 2001). Mn is also one of the most abundant and essential elements in the world. High concentrations of manganese cause color problems in water and health problems in human. FAC4 represents the agricultural land use characteristics shown by presence of Mg and $SO_4$.

The last factor FAC5 was related with the parameters Ca and Total coliforms and explained 9.06% of total variance. High Ca concentration yields high Mg amounts. High metal contents in water are due to runoff carrying metals from municipal wastewater, urban areas and agricultural lands.

## Conclusions

This paper introduces assessment of water quality data based on 10 monitoring stations in Kirmir Basin around the capital city of Turkey, Ankara by using statistical multivariate analysis (factor and principal component analysis). FA/PCA identified water quality parameters in five groups. The results revealed that Kirmir basin is mainly deteriorated with non-point pollution sources including mainly agricultural activities, urban land uses and livestock activities. Minimization and controlling these activities will be helpful for improving the water quality and providing better ecology for living organisms in Kirmir Basin. Multivariate statistical analysis was successfully applied for the assessment of water quality in Kirmir basin. In comparison to other techniques, this method has an advantage of providing additional information about potential pollution sources while analyzing large amount of raw data.

**Acknowledgements**

This study was conducted in School of Civil and Environmental Engineering, Georgia Institute of Technology. The author was supported by Islamic Development Bank (IDB) under Merit Scholarship and by General Directorate of State Hydraulic Works (DSI) during her studies in USA. The author would like to thank to those who contributed to collection of water samples and laboratory analysis in DSI Laboratories.

**References**

Akkoyunlu, A & Akiner, ME 2012, 'Pollution evaluation in streams using water quality indices: A case study from Turkey's Sapanca Lake Basin', *Ecological Indicators*, vol. 18, pp. 501–511.

Akoteyon, IS, Omotayo, AO, Soladoye, O & Olaoye, HO 2011, 'Determination of Water Quality Index and Suitability of Urban River for Municipal Water Supply in Lagos-Nigeria', *European Journal of Scientific Research*, vol. 54, no. 2, pp. 263–271.

Avtar, R, Kumar, P, Singh, CK & Mukherjee, S 2011, 'A Comparative Study on Hydrogeochemistry of Ken and Betwa Rivers of Bundelkhand Using Statistical Approach', *Water Quality Exposure and Health*, vol. 2, no. 3-4, pp. 169–179.

Azhar, SC, Aris, AZ, Yusoff, MK, Ramli, MF & Juahir, H 2015, 'Classification of River Water Quality Using Multivariate Analysis', *Procedia Environmental Sciences*, vol. 30, pp. 79–84.

Boyacioglu, H & Boyacioglu, H 2007, 'Surface water quality assessment by environmetric methods', *Environmental Monitoring Assessment*, vol. 131, no. 1-3, pp. 371–376.

Cude, CG 2001, 'Oregon Water Quality Index a Tool for Evaluating Water Quality Management Effectiveness', *Journal of the American Water Resources Association*, vol. 37, no. 1, pp. 125–137.

DEIAP 2016, *Çevre ve Şehircilik İl Müdürlüğü Ankara İli 2015 Yılı Çevre Durum Raporu*, Ministry of Environment and Urbanisation of Republic of Turkey, available on: http://www.csb.gov.tr/db/ced/editordosya/Ankara2015.pdf [23.08.2016].

Eaton, AD & Clescen, LS 2005, 'Standart methods for the examination of water and wastewater (STMD 2005)', *American Public Health Association (APHA)*, Washington.

Malik, RN, & Nadeem, M 2011, 'Spatial and temporal characterization of trace elements and nutrients in the Rawal Lake Reservoir, Pakistan using multivariate analysis techniques', *Environmental Geochemistry and Health*, vol. 33, no. 6, pp. 525–541.

Mazlum, N, Ozer, A & Mazlum, S 1999, 'Interpretation of Water Quality Data by Principal Components Analysis', *TR Journal of Engineering and Environmental Science*, vol. 23, pp. 19–26.

Sánchez, E, Colmenarejo, MF, Vicente, J, Rubio, A, García, MG, Travieso, L, & Borja, R 2007, 'Use of the water quality index and dissolved oxygen deficit as simple indicators of watersheds pollution', *Ecological Indicators*, vol. 7, NO. 2, pp. 315–328.

Tezcanli Guyer, G & Genc Ilhan, E 2011, 'Assessment of pollution profile in Buyukcekmece Watershed, Turkey', *Environmental Monitoring Assessment*, vol. 173, pp. 211–220.

TSE 1998, 'Water quality - determination of dissolved anions by liquid chromatography of ions - Part 1: Determination of bromide, chloride, fluoride, nitrate, nitrite, phosphate and sulfate. Section 1: Method for less polluted waters`. *Turkish Standard Institution*, Turkey.

TSE 2000, 'Water quality - determination of dissolved $Li^{+}$, $Na^{+}$, $NH_4^{+}$, $Mn^{2+}$, $Ca^{2+}$, $Mg^{2+}$, $Sr^{2+}$ and $Ba^{2+}$ using ion chromatography - Method for water and waste water'. *Turkish Standard Institution,* Turkey.

Tunc Dede, O, Telci, IT & Aral, MM 2013, 'The Use of Water Quality Index Models for the Evaluation of Surface Water Quality: A Case Study for Kirmir Basin, Ankara, Turkey', *Water Quality, Exposure and Health*, vol. 5, no. 1, pp. 41–56.

Turkish-Regulation 2012, *İçme suyu elde edilen veya elde edilmesi planlanan yüzeysel suların kalitesine dair yönetmelik*, no. 28338, Ministry of Forestry and Hydraulic Works of Republic of Turkey, available on http://www.resmigazete.gov.tr/eskiler/2012/06/20120629-9.htm. [23.08.2016].

WHO 2009, 'Guidelines for drinking water quality', World Health Organization, available on: http://apps.who.int/iris/bitstream/10665/70050/1/WHO_HSE_WSH _09.05_eng.pdf [23.08.2016].

WHO 2011, '*Hardness in Drinking-water: Background Document for Development of WHO Guidelines for Drinking-water Quality'*, World Health Organization, available on: http://www.who.int/water_sanitation_health/dwq/chemicals/hardness.pdf, [23.08.2016].