Tr. J. of Engineering and Environmental Science
23 (1999) , 19 – 26.

Interpretation of Water Quality Data by Principal Components Analysis

Nazire MAZLUM, Adem ÖZER*, Süleyman MAZLUM
Dokuz Eylül University
Environmental Engineering Department
Tınaztepe Campus, Buca-İzmir-TURKEY

Received 19.11.1996

Abstract

A variety of methods are being used to display the information which is concealed in the quality variables observed in a water quality monitoring network. A large portion of these approaches are statistical methods. When the number of variables is greater than two, employment of multivariate analysis techniques gives simpler and more easily interpretable results for the evaluation of observed quality data. In this study, it was attempted to determine factors that caused variations in water quality at the Ağaçköy Monitoring station on the Porsuk Tributary in the Sakarya river-basin by principal components analysis.

Key Words: Multivariate analysis, principal components analysis, quality variables, river-basin, water quality.

Esas Bileşenler Analizi Yardımıyla Su Kalitesinin Değerlendirilmesi

Özet

Su kalitesi gözlem istasyonlarında ölçülen kalite değerlerinin içerdiği bilgiyi ortaya çıkarmak için çeşitli yöntemler kullanılmaktadır. Bu yöntemlerin çoğunluğu istatistiksel metodlara dayanmaktadır. Değişken sayısının ikiden fazla olması durumunda, gözlenmiş mevcut verilerin değerlendirilmesi için çok değişkenli analizlerin (multivariate analysis) kullanılması sayesinde daha basit ve kolay yorumlanabilir sonuçlar elde edilebilmektedir. Bu çalışmada, Sakarya havzasında Porsuk kolu üzerinde kurulmuş bulunan Ağaçköy su kalitesi gözlem istasyonunda gözlemlenen kalite değişimlerine neden olan faktörler esas bileşenler analizi kullanılarak belirlenmeye çalışılmıştır. Başka bir ifadeyle gözlem istasyonu çevresinde su kalitesindeki değişimlere neden olan etkenlerin tespit edilmesi amaçlanmıştır.

Anahtar Sözcükler: Anahtar kelimeler: Akarsu havzası, çok değişkenli analiz, esas bileşenler analizi, kalite değişkenleri, su kalitesi.

Introduction

Multivariate analysis techniques are very useful in the analysis of data corresponding to a large number of variables. Analysis via these techniques produces easily interpretable results. Multivariate data consists of observations on several variables for a number of samples (also called *sample vectors*, or *individuals*). Data of this type arise in all branches of science, ranging from physiology to biology, and methods of analyzing multivariate data constitute an increasingly important area of statistics.

*To whom all correspondence should be addressed

A wide variety of multivariate analysis techniques is available. The choice of the most appropriate technique depends on the nature of the data, problem, and objectives. The underlying theme of many multivariate analysis techniques is simplification. In other words, it is desired to summarize a large body of data by means of relatively few parameters.

One fundamental distinction between the techniques is that some analysis are primarily concerned with relationships between variables, while others are primarily concerned with relationships between samples. Techniques of the former type are called variable-directed, while the latter are called individual-directed (sample-directed) multivariate analysis.

In the analysis of dependence between variables, if the variables do not arise on an equal footing, multivariate regression analysis is recommended. It should be noted that the term 'equal footing' does not imply that some variables are more important than others, though they may be. Rather it indicates that there are dependent and explanatory variables. In multiple regression, the variation in one dependent variable is explained by means of the variation in several explanatory variables. In multivariate regression, more than two dependent variables are in question.

If the variables arise on an equal footing, as for example when different dimensions of different members of a particular species are measured and the primary interest is in the variables, then analysis of interdependence of variables is the subject of interest. When there are only two variables, correlation analysis provides the desired information to some extent. With more than two variables, principal components analysis may be appropriate. This technique aims to transform the observed variables to a new set of variables which are uncorrelated and arranged in decreasing order of importance. The principal aim is to simplify the problem and to find new variables (principal components) which make the data easier to understand.

In this multivariate analysis study, principal component analysis was employed to investigate the factors which caused variations in the observed quality data at the Ağaçköy water quality monitoring station in the Sakarya river-basin. This study also demonstrates the usefulness of the technique in the analysis of water quality data. A literature review on principal components analysis, a technique that was formerly used in the field of hydrology, has shown its appropriateness for water quality data, as confirmed by some recent case studies in the literature (Mahloch, 1974; Schetagne, 1985; Karpuzcu & Şenes, 1987).

Principal components analysis requires the interpreter to be experienced in the field of inquiry. For instance, a mathematician or statistician, applying psychological data to principal components analysis, can not interpret the components reliably. This point, interpretation, constitutes the most problematic aspect of principal components analysis.

In the realization of this study, software programs commercially known as Minitab (1991) and Systat (1990) were used. Minitab (1991) statistical software was used in regression analysis for the substitution of missing values and systat (1990) was principally utilized for the derivation and subsequent rotation of principal components.

2. Principal Components Analysis

2.1. Principal Components from Correlation or Covariance Matrix

Correlation or covariance matrices may be used in principal component analysis. The sums of squares and sums of products of the normalized scores constitute the correlation matrix, R (Hope, 1968). In other words, if one is deriving the principal components from correlation matrix R, this means that the variables have been standardized to have unit variance. The mathematical derivation is the same in the choice of any of these matrices. If the correlation matrix is used, the components turn out to be eigenvectors of R. However, it is important to note that the eigenvalues and the eigenvectors of R will generally not be the same as those of S (covariance matrix). The use of the R matrix for analyzing involves a decision that variables have been considered equally important (Chatfield & Collins, 1980).

Principal components are generally changed by scaling. In other words, principal components depend upon the scales used to measure the variables. If, for example, one variable has a greater variance than all the others, then this variable will dominate the first principal component of the covariance matrix whatever the correlation structure, whereas if the variables are all scaled to have unit variance, then the first principal component will be quite different in kind (Chatfield & Collins, 1980). Karpuzcu and Şene (1987) stated that most applications of this analysis have involved a correlation matrix rather than a covariance matrix. They stated that if the parameters (variables) are in widely different units (mg/l., pH, °C, m^3/min., etc.), then standard variates and correlation matrix should be used.

Considering the importance of the variables in principal components analysis, Chatfield and Collins (1980) stated that if the variables are not to be considered of equal importance, then the analysis of the correlation matrix is not recommended.

2.2. The Identification of Important Components

After computing the variances (eigenvalues, or latent roots) and principal components (eigenvectors) of a correlation (or covariance) matrix, the usual procedure is to look at the first few components which, hopefully, account for a large proportion of the total variance.

Chatfield and Collins (1980) stated that when analyzing a correlation matrix where the sum of the eigenvalues is equal to the number of variables, many social scientists use the rule that eigenvalues less than 1 may be disregarded. This arbitrary policy is a useful rule of thumb but has no theoretical justification. It may be better to look at the pattern of eigenvalues and see if there is a natural breakpoint in the eigenvalues. Chatfield and Collins (1980) claimed that one serious drawback is that there is no objective way of deciding how many components to retain.

2.3. Rotation of Principal Components

In principal components analysis, the variables are rotated to obtain new variables (principal components or principal axes) and later the number of principal components are reduced by eliminating some relatively unimportant components. Sometimes the first few principal components selected are rotated to achieve a new set of components which can be more easily interpreted. A variety of rotation techniques (varimax, equamax, quartimax) may be used for this purpose. Varimax rotation is the most widely used rotation in principal component analysis. This rotation, which includes an orthogonal rotation, is too complicated a technique to explain in this study. The original paper by the author illustrates the theory of the technique (Kaisher, 1958).

When using rotation, the usual procedure is to carry out a principal components analysis, first by subjectively choosing the number of important principal components, and then calculating linear combinations of the selected eigenvectors in the subspace of m dimensions so as to get a new set of components (which will no longer be principal) satisfying some desired property. The idea is that each variable should be heavily loaded on as few components as possible. One such technique for accomplishing this transformation is a varimax rotation. This technique tends to eliminate medium-range correlations between the components and the original variables, thus simplifying the decision as to which of the original variables to include in the components extracted (Chatfield & Collins, 1980).

3. Revealing the Causes of the Variations in Sakarya River-Basin water Quality by Principal Components Analysis

3.1. Selected Station

Water quality data observed at monthly intervals between the years 1979 and 1984 at Ağaçköy monitoring station, situated on Porsuk Tributary in the Sakarya river-basin, were obtained from the records of State Hydraulic Works (DSI, 1985; DSI, 1987) and used in this study. The Sakarya river-basin is located in north-western Anatolia, and has a surface area of 58,000 km^2. The main course of the river originates from the northern part of the city of Afyon city and extends 824 km before discharging into the Black Sea. The average flow of the river is about 200 m^3/s, with a highly variable flow regime. A significant tributary of Sakarya river is Porsuk, which passes through the plains of Kütahya and the city of Eskişehir. The length of Porsuk is 442 km up to its confluence with the Sakarya river near the town of Polatlı (Mazlum, 1992). The Ağaçköy monitoring station is situated on the Porsuk tributary near the city of Kütahya. This situation is subject to pollution from industrial, agricultural and residential sources. Surface runoff and drainage waters also cause nonpoint source pollution along the river.

3.2. Estimation of Missing Values

Water quality data used in this study was that obtained at the Ağaçköy monitoring station between 1979 and 1984 in the Sakarya river-basin. There were some missing values in the observed data. The method proposed by Buck (1960) was used in this study for the estimation of missing values. This method is based on the use of the multilinear regression equation, in which the dependent variable is the variable with the missing value.

Ağaçköy monitoring data were measured at monthly intervals, constituting 72 samples (rows) for six years. Out of 72 rows, 4 did not have any observed values and therefore were completely eliminated from the analysis. The remaining 68 rows were evaluated, and there were determined to be (i) 55 complete samples, (ii) 6 samples missing variables for SS (Suspended Solids) (iii) 3 samples missing variables for DO (Dissolved Oxygen) (iv) 1 sample missing a variable for BOD_5 (v) and 3 samples missing variables for SS (Suspended Solids) and NO_3-N.

3.3. Principal Component Analysis of Water Quality Data

Among the observed variables at Ağaçköy monitoring station, Q, T, pH, EC, SS, MAl, Cl, NH_3-N, NO_3-N, DO, Pv, and BOD_5 were selected for the analysis due to their continuity of measurement in time scale. These symbols signify flowrate, temperature, negative logarithms of hydrogen-ion concentration in water, electrical conductivity, suspended solids, methyl-orange alkalinity, chloride, ammonia-nitrogen, nitrate-nitrogen, dissolved oxygen, permanganate value, and biochemical oxygen demand, respectively.

In the application of principal components analysis to water quality data from the Ağaçköy monitoring station a correlation matrix was used. The reason was that variables were different in scale (as suggested by Karpuzcu and Şeneş, 1987) and equal in importance (as suggested by Chatfield & Collins, 1980).

The results of principal components analysis of the data are presented in Table 1a, and the 12 subsequently derived components were rotated according to varimax rotation in order to make interpretation easier (Table 1b). Table 1b yields 12 factors which may be interpreted as having vital importance to the water quality status of the river-basin.

As Chatfield and Collins (1980) stated, components with an eigenvalue of less than 1 should be eliminated so that fewer components are dealt with. The first four components were extracted (Table 2a) and the other components have been eliminated. When the percentages of the total variances of the 4 extracted components are accumulated, it can be seen that these first four principal components account for 72% of the total variance of the original data. This means that majority of the variance of the original data has been accounted for by these extracted components. These components were later rotated (Table 2b).

Table 1a. Principal components from correlation matrix

	Latent Roots (Eigenvalues or Variances) Explained by Principal Components											
	1	2	3	4	5	6	7	8	9	10	11	12
	3.481	2.456	1.549	1.162	0.989	0.677	0.511	0.412	0.271	0.258	0.135	0.099
	Percent of Total Variance Explained											
	1	2	3	4	5	6	7	8	9	10	11	12
	29.007	20.467	12.905	9.683	8.246	5.645	4.260	3.432	2.261	2.146	1.122	0.827
	Component Loadings											
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
Q	0.813	0.337	-0.262	-0.017	0.007	0.203	0.036	0.049	0.077	0.241	-0.168	-0.138
T	-0.734	0.539	-0.238	0.075	-0.035	0.018	0.084	0.182	0.023	-0.037	0.170	-0.191
pH	-0.088	0.061	-0.246	0.472	0.829	0.032	-0.041	-0.035	0.113	0.006	0.013	0.038
EC	-0.821	0.312	-0.067	0.079	-0.024	-0.203	-0.028	0.206	-0.160	0.304	-0.065	0.103
SS	0.644	0.575	-0.264	-0.003	-0.172	0.149	0.251	0.066	0.110	0.052	0.158	0.162
MA1	-0.463	-0.380	0.357	0.083	0.015	0.682	0.085	0.179	-0.025	0.016	-0.015	0.026
C1	-0.183	0.372	0.646	-0.349	0.301	-0.028	0.323	-0.274	-0.029	0.124	0.032	-0.032
NH_3-N	0.097	0.565	0.641	-0.037	-0.027	0.007	-0.450	0.095	0.213	0.032	0.035	0.015
NO_3-N	0.029	-0.378	-0.163	-0.775	0.333	-0.096	0.026	0.307	0.114	-0.015	0.003	0.014
DO	0.659	-0.623	0.113	0.067	0.095	-0.017	-0.161	0.023	-0.174	0.220	0.208	-0.048
Pv	0.524	0.675	0.061	-0.120	0.244	0.096	-0.091	0.131	-0.349	-0.184	-0.030	0.009
BOD_5	0.467	-0.176	0.539	0.421	-0.007	-0.295	0.298	0.316	0.041	-0.060	-0.041	-0.026

Table 1b. Varimax rotated components

	Variance Explained by Rotated Components											
	1	2	3	4	5	6	7	8	9	10	11	12
	2.470	2.044	1.043	1.014	1.022	1.015	1.033	0.968	0.742	0.369	0.145	0.135
	Percent of Total Variance Explained											
	1	2	3	4	5	6	7	8	9	10	11	12
	20.585	17.037	8.690	8.452	8.517	8.457	8.610	8.070	6.181	3.073	1.206	1.121
	Rotated Loadings											
	1	2	3	4	5	6	7	8	9	10	11	12
Q	-0.293	0.868	-0.093	0.008	0.046	0.174	0.060	0.002	0.168	-0.029	-0.003	0.295
T	0.942	-0.070	0.010	0.088	0.090	-0.047	0.015	-0.173	0.008	0.068	0.235	0.021
pH	0.059	-0.021	-0.020	0.038	0.994	0.006	-0.065	-0.000	0.044	0.012	0.004	0.004
EC	0.725	-0.316	0.053	0.064	0.046	-0.024	0.032	-0.104	-0.072	0.589	0.018	0.001
SS	0.019	0.924	-0.018	0.100	-0.078	0.209	0.032	0.050	0.163	-0.085	-0.000	0.217
MA1	0.032	-0.285	0.054	0.013	-0.007	-0.950	-0.017	-0.011	-0.111	0.008	0.002	0.004
Cl	0.129	-0.062	0.958	-0.046	-0.021	-0.054	0.215	0.044	0.089	0.014	-0.008	0.005
NH_3-N	0.068	0.076	0.229	0.133	-0.080	0.017	0.938	0.086	0.159	0.009	-0.004	0.002
NO_3-N	-0.140	-0.079	0.043	-0.971	-0.041	0.011	-0.119	-0.114	-0.009	-0.017	0.002	0.003
DO	-0.906	0.028	-0.175	-0.097	0.022	-0.009	-0.075	0.192	-0.006	0.077	0.299	.008
Pv	-0.016	0.482	0.151	0.014	0.083	0.182	0.262	0.020	0.796	-0.033	-0.000	0.006
BOD_5	-0.323	0.039	0.049	0.130	-0.000	0.012	0.088	0.930	0.015	-0.031	0.004	0.001

Table 2a. Principal components with an eigenvalue less than1

	Latent Roots (Eigenvalues or Variances) Explained by Principal Components				
	1	2	3	4	
	3.481	2.456	1.549	1.162	
	Percent of Total Variance Explained				
	1	2	3	4	
	29.007	20.467	12.905	9.683	
	Component Loadings				**Communalities**
	PC1	PC2	PC3	PC4	
Q	0.813	0.337	-0.262	-0.017	0.843
T	-0.734	0.539	-0.238	0.075	0.891
pH	-0.088	0.061	-0.246	0.472	**0.294**
EC	-0.821	0.312	-0.067	0.079	0.782
SS	0.644	0.575	-0.264	-0.003	0.815
Mal	-0.463	-0.380	0.357	0.083	**0.493**
Cl	-0.183	0.372	0.646	-0.349	0.710
NH_3-N	0.097	0.565	0.641	-0.037	0.740
NO_3-N	0.029	-0.378	-0.163	-0.775	0.770
DO	0.659	-0.623	0.113	0.067	0.839
Pv	0.524	0.675	0.061	-0.120	0.748
BOD_5	0.467	-0.176	0.539	0.421	0.716

Table 2b. Varimax rotated (the first four) components

	Variances Explained by Rotated Components				
	1	2	3	4	
	2.876	2.778	1.744	1.250	
	Percent of Total Variance Explained				
	1	2	3	4	
	23.969	23.148	14.530	10.414	
	Rotated Loadings				**Communalities**
	1	2	3	4	
Q	0.314	0.861	-0.057	0.025	0.843
T	-0.918	-0.078	0.046	0.200	0.890
pH	-0.127	0.021	-0.281	0.447	**0.295**
EC	-0.793	-0.344	0.080	0.166	0.781
SS	0.058	0.893	0.054	0.102	0.814
Mal	0.010	-0.694	0.102	0.030	**0.493**
Cl	-0.158	-0.100	0.804	-0.173	0.711
NH_3-N	-0.022	0.177	0.824	0.169	0.739
NO_3-N	0.077	-0.070	-0.154	-0.859	0.772
DO	0.884	0.008	-0.223	-0.096	0.840
Pv	0.014	0.764	0.403	0.050	0.748
BOD_5	0.688	-0.044	0.279	0.404	0.578

However, communalities in Table 2a and Table 2b show that variances of variable pH and MA1 have not been described well by these four components. Percentages of the total variance of pH and MA1 are 0.295 and 0.493 respectively (Table 1b). For this reason, the fifth principal component was taken into consideration.

Since the communalities show that variance of the variable MA1 has again not been described well, the next component, component 6, has also been taken into consideration (Table 3a) and subsequently these six components have been rotated (Table 3b). In this case, communalities show that all the variables have been described to an acceptable level.

Table 3a. The first six principal components

	Latent Roots (Eigenvalues or Variances) Explained by Principal Components						
	1	2	3	4	5	6	
	3.481	2.456	1.549	1.162	0.989	0.677	
	Percent of Total Variance Explained						
	1	2	3	4	5	6	
	29.007	20.467	12.905	9.683	8.246	5.645	
	Component Loadings						**Communalities**
	PC1	PC2	PC3	PC4	PC5	PC6	
Q	0.813	0.337	-0.262	-0.017	0.007	0.203	0.884
T	-0.734	0.539	-0.238	0.075	-0.035	0.018	0.893
pH	-0.088	0.061	-0.246	0.472	0.829	0.032	0.983
EC	-0.821	0.312	-0.067	0.079	-0.024	-0.203	0.823
SS	0.644	0.575	-0.264	-0.003	-0.172	0.149	0.866
MA1	-0.463	-0.380	0.357	0.083	0.015	0.682	0.958
C1	-0.183	0.372	0.646	-0.349	0.301	-0.028	0.802
NH_3-N	0.097	0.565	0.641	-0.037	-0.027	0.007	0.741
DO	0.659	-0.623	0.113	0.067	0.095	-0.017	0.848
PV	0.524	0.675	0.061	-0.120	0.244	0.096	0.817
BOD_5	0.467	-0.176	0.539	0.421	-0.007	-0.295	0.803

Table 3b. Varimax rotated (the first six) components

	Variances Explained by Rotated Components						
	1	2	3	4	5	6	
	2.859	2.554	1.669	1.241	1.048	0.942	
	Percent of Total Variance Explained						
	1	2	3	4	5	6	
	23.838	21.285	13.908	10.343	8.737	7.850	
	Rotated Loadings						**Communalities**
	1	2	3	4	5	6	
T	−0.926	-0.092	0.042	0.122	0.102	-0.014	0.893
DO	0.895	0.019	-0.199	-0.080	0.035	0.000	0.848
EC	−0.786	-0.415	0.094	0.114	0.073	0.085	0.824
BOD_5	0.670	-0.110	0.252	0.480	0.046	0.215	0.813
SS	0.014	0.892	-0.017	0.147	-0.111	0.190	0.866
Q	0.290	0.882	-0.079	0.014	0.024	0.123	0.844
Pv	0.014	0.757	0.440	-0.003	0.155	0.165	0.817
Cl	-0.126	-0.085	0.867	-0.155	0.020	-0.063	0.803
NH_3-N	-0.048	0.200	0.773	0.287	-0.137	0.006	0.741
NO_3-N	0.163	-0.121	-0.007	−0.919	-0.035	0.063	0.891
pH	-0.066	0.009	-0.067	0.036	0.986	-0.003	0.982
MA1	-0.012	-0.389	0.054	0.044	0.004	−0.896	0.959

Interpretation of the Results

Component loadings (correlation coefficients) and the variances (eigenvalues) regarding the components were computed for all the variables at the first step (Table 1a). The proportion of the total variance explained by each principle component is additive, with each new component contributing less than the preceding one to the explained variance. In other words, the components are derived in decreasing order of importance in Table 1a. Subsequently, these components were rotated to eliminate medium-range loadings (correlations) to make the interpretation of the components easier (Table 1b). Twelve rotated components which were too numerous to explain and components which explain a relatively small proportion of the total variance of the principal components were eliminated for simplification (Mazlum, 1994).

Then the first four components selected (Table 2a), whose variances are greater than 1, were rotated (Table 2a). In the last step, the number of components considered was 6 (Table 3a). Subsequently, these 6 components were rotated (Table 3b). Table 3b shows that 0.959 of the total variance of MA1 was described. Since all the communalities are larger than 0.7 in this final case, it may be assumed that all the variables were described to an acceptable level. It can also be seen from Table 3b that these six principal components accounted for 86% of the total variance of the original data. Consequently, the conclusive result was that the first six components can be considered significant in the analysis.

In general, component loadings (correlation coefficients) larger than 0.6 may be taken into consideration in the interpretation (Mahloch, 1974). In other words, the most significant variables in the components represented by high loadings have been taken into consideration in evaluating the components. In addition to the significance of high loading values, there exists a difference between the components; the components with larger variances are more desirable since they give more information about the data. When the variances (eigenvalues) of the components are examined in the related tables, it can be seen that principal components are in decreasing order of importance with respect to their variances.

An interpretation of the rotated six principal components in Table 3b is made by examining the component loadings noting the relationship to the original variables. The first component gives information about the variation in dissolved oxygen state, electrical conductivity, temperature and BOD_5. In this component, BOD_5 loading indicates that a domestic discharge exists just ahead of the station. However, when the raw data observed at the Ağaçköy monitoring station is examined, it can be concluded that the observed BOD_5 values are quite low (about 2 mg/lt.), proving that there is a relatively small discharge as compared to river discharge. In the second component, suspended solids (SS) and methyl orange alkalinity (MA1) are important; seasonal effect of flow is demonstrated by the positive relation between suspended solids and flow in this component. In the third component, it can be understood that a domestic waste is discharged into the river due to the significance of Chloride (Cl) and ammonia-nitrogen (NH_3-N) in that component. Due to the significance of nitrate in the fourth component, it can be concluded that nitrification takes place in the vicinity of the station. In the fifth component, pH is the important variable and it can be claimed that an industrial waste discharge is being carried out. The last component shows that alkalinity shows increase, possibly due to the river-basin structure.

Evaluation of the Results and Conclusions

From the 12 components in Tables 1a and 1b, the first six components are sufficient to explain the monitoring area. As can be seen in these tables, these components explain more than 70% of the total variance of the original data set. Moreover, the first six selected principal components explained more than 70% of the variance of each quality variable (see Communalities in Table 3a and Table 3b). As can be seen in the tables, the variance of the principal components and the variance of the related rotated components are nearly the same for all components. Therefore, both tables a and b in this article can be examined to evaluate the percent of variances of the components explained.

Principal components analysis of water quality data from the Ağaçköy monitoring station showed that small domestic waste discharge, industrial waste discharge, nitrification and seasonal effects are the main causes of variations in water quality in that region. Availability of information on the activities in that region would have made the interpretation of the components easier since some activities might have lasted for short periods and not for the entire duration of the observation period.

In the interpretation of principal components, difficulties arose in explaining the physical nature of significant variables (variables with high loadings) in some components (e.g. T, DO, EC, BOD5 in the first component in Table 1b). From this point if view, it can be said that the principal components were not obtained well enough for interpretation of the components with respect to variables. Limited number of data (insufficient data), improper frequency of observation of data and errors in analyzing the quality variables in the laboratory or in situ may have contributed to this failure.

For this reason, it is worthwhile to stress that more systematically observed data should be used in this analysis. More frequently observed (in a certain time period) data or data observed as frequently as necessary may help obtain more meaningful principal components. Furthermore, it may be necessary to observe and subsequently include certain new quality variables in the analysis in order to make the interpretation of principal components easier. With the help of these new variables, it may be easier to understand what it is the factor explain (e.g. nitrification, domestic discharge). On the other hand, the observation of some variables may not be important for this analysis and thus may be eliminated from the monitoring program.

Similar to this analysis, factor analysis was formerly used to discover the factors which affect the field of inquiry (monitoring field). Past experiences with factor analysis were mostly on medicine (specifically psychology), biology and hydrology, and recent applications have intensified in the field of water quality. Factor analysis is being replaced by principal components analysis, which yields more reliable results. Factor analysis includes an error structure while principal components analysis is a pure mathematical technique without any assumption. The results of principal components analysis demonstrate that it provides reliable information with respect to reality in fields of scientific research.

References

Buck, S.F. "A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with an Electronic Computer". The Journal of the Royal Statistics Society, Vol.22, No.2, 302-306, 1960.

Chatfield, C. & Collin, A.J. "Introduction to Multivariate Analysis". Published in the USA by Chapman and Hall in Association with Methuen, Inc. 733 Third Avenue, New York NY. 10017, ISBN 0-412-16030-7. 1980.

"DSI Su Kalitesi Gözlem Yıllığı". 1979-1982. İçme Suyu ve Kanalizasyon Dairesi Başkanlığı. 1985.

"DSI Su Kalitesi Gözlem Yıllığı". 1983-1984. İçme Suyu ve Kanalizasyon Dairesi Başkanlığı. 1987.

Hope, K. "Methods of Multivariate Analysis". University of London Press Ltd. St. Paul's House, Warwick Lane, London EC4, SBN 340 07 331 4 Unibook. 1968.

Kaiser, H.F. "The Varimax Criterion for Analytic Rotation in Factor Analysis". Psychometrica. Vol.23, No.3. 1958.

Karpuzcu, M. & Şenes, S. "Design of Monitoring Systems for Water Quality by Principal Component Analysis and a Case Study". Proceedings of the International Symposium on Environmental Management. Environment '87. Vol.1, 673-690. Istanbul 1987.

Mazlum, N. "Multivariate Analysis of Water Quality Data". Master of Science Thesis, Dokuz Eylül University, Graduate School of Natural and Applied Sciences. Bornova, İzmir 1994.

Mahloch, J.L. "Multivariate Techniques for Water Quality Analysis", J. Environ. Eng. Div. Amer. Soc. Civil Eng., 100, No. EE5, 1119-1132, 1974.

Mazlum, S. "Variable Selection in Water Quality Monitoring Networks". Master of Science Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Environmental Engineering. Dokuz Eylül University, Graduate School of Natural and Applied Sciences. Bornova-İzmir, 1992.

"Minitab Statistical Software". 3081 Enterprise Drive, State College, PA 16801-3008, Release 8.2. Copyright 1991.

Schetagne, R. "The Use of Multivariate Methods in the Interpretation of Water Quality Monitoring Data of a Large Northern Reservoir". Proceedings of the Workshop Held at the Canada Center for Inland Waters, October 7-10, 1985. Edited by A.H. EL-SHARARAWI-R.E. KWI ATKOWSKI. Statistical Aspects of Water Quality Monitoring, 31-43, 1985.

"Systat Statistical Software", Version 5.0. Systat Inc. Copyright 1990.