Application of multivariate statistical techniques for surface water quality assessment: case study of Karaj River, Iran

G. Badalians Gholikandi[1], H. R. Orumieh[2], S. Haddadi[3]
& N. Mojir[4]
*[1]Power and Water University of Technology (PWUT),
Water Research Institute (WRI), Tehran, Iran*
[2]Parsarianab Consulting Engineers, Iran
*[3]Faculty of Environment, University of Tehran/ Water & Wastewater
Research Centre, Water Research Institute (WRI), Iran*
*[4]Industrial Engineering Department, Faculty of Engineering,
University of Tehran, Tehran, Iran*

Abstract

Water quality monitoring has one of the highest priorities in surface water protection policy. Multivariate statistical methods have been used successfully in hydrochemistry for many years. In the present study, evaluation of spatial variations and interpretation of Karaj River water quality data were made by using multivariate analytical techniques including factor analysis and cluster analysis. Data set consisted of 11250 observations of a three-year monitoring program (measurement of 24 variables at 20 stations from April 2006 to March 2009). Factor analysis with principal component analysis extraction of the data set yielded seven varifactors contributing to 82% of total variance and evaluated incidence of each varifactor on the total variance. Three groups of quality parameters, e.g. salinity, organic matter, and nutrients were investigated. Natural sources pollution originates most variation on water quality. Cluster analysis clustered 20 sampling sites into two major groups and demonstrated usefulness of this method for data reduction. This spatial similarity and site clustering makes possible optimal designing of the future sampling strategy that saves both time and costs. The results of cluster analysis came complete with T test and made water quality comparison between two clusters possible. Results of factor

WIT Transactions on Ecology and the Environment, Vol 145, © 2011 WIT Press
www.witpress.com, ISSN 1743-3541 (on-line)
doi:10.2495/WRM110311

analysis were employed to facilitate T test analysis. T test revealed significant difference between the mean of calculated varifactors 1, 2, 6 and 7 between two clusters and any significant difference in the mean of other varifactors 3, 4 and 5 between two groups, in a confidence interval of 95%. The result shows the effect of agricultural fertilizers on downstream stations of the dam, although amount of corresponding parameters do not excess of World Health Organisation's drinking water standards. Results showed also the positive affect of Karaj basin regarding physicochemical water quality.

Keywords: surface water quality, multivariate statistical techniques, factor analysis, cluster analysis, Karaj River.

1 Introduction

Surface water is highly prone to point and non-point pollutions due to its easy accessibility for disposal of wastewaters, difficulty with its protection, its being uncovered, and its high flow velocity. Surface water is influenced by natural processes (precipitation, erosion, weathering of crustal materials) and anthropogenic effects (urban, industrial, and agricultural activities and increasing consumption of water resources) [1–3]. Anthropogenic forces have the immense tendency to accelerate natural processes that affect water quality [4]. Water quality refers to the physical, biological and chemical status of water bodies [5]. Discharges from municipals and industries are considered as a point and constant polluting source while surface runoff is as a seasonal phenomenon and non-point source due to its characteristics that are highly influenced by climate and seasonal changes [6, 7]. Surface waters have heterogeneity in space and time. Temporal and spatial changes both in natural process and anthropogenic influences cause spatiotemporal variations in water quality parameters; therefore, reliable and regular monitoring programs reflecting the variations have to be set. In Iran, governmental companies have carried out water quality monitoring programs, but many of those monitoring programs contain large data sets.

A particular problem in the case of water quality monitoring is the complexity associated with analyzing the large number of measured variables. The data sets contain rich information about the behaviour of the water resources. Classification, modelling and interpretation of monitored data are the most important steps in the assessment of water quality [8]. The application of multivariable statistical methods offers a better understanding of water quality for interpreting the complicated data sets [9]. Multivariate statistical methods have many applications in different environmental studies. They have been presented as appropriate tools in water quality assessment, identification of pollution sources/factors and understanding temporal/spatial variations in water quality for effective river water quality management [2, 6, 7, 10]. They may be particularly useful when there is a large volume of experimental results and sometimes they provide insight into the multidimensional patterns in the data that would be overlooked with univariate analyses [11]. The multivariate methods applied in water quality analyses are cluster analysis (CA), discriminant analysis (DA), factor analysis (FA), principal component analysis (PCA), and exploratory factor analysis (EFA).

Factor analysis was particularly useful for considering several related random environmental variables simultaneously, and so identification of a new, smaller set of uncorrelated variables that accounted for a large proportion of the total variance in the original variables [12]. Varimax factor rotation method rotates the axis such that the two vertices remain 90 degrees (perpendicular) to each other and assumes uncorrelated factors (also referred to as orthogonal rotation). Cluster analysis (first used by Tryon, 1939) classifies objects so that each object can be similar to the others in the cluster with respect to a predetermined selection criterion. The resulting clusters of objects should then exhibit high internal (within-cluster) homogeneity and high external (between-cluster) heterogeneity [13]. The goal is that the objects within a group be similar (or related) to one another and different from (or unrelated to) the objects in other groups. The greater the similarity (or homogeneity) within a group, and the greater the difference between groups, the better or more distinct the clustering.

Principal component analysis and exploratory factor analysis are multivariate statistical techniques used to identify important components or factors that explain most of the variances of a system. They are designed to reduce the number of variables to a small number of indices (i.e., principal components or factors) while attempting to preserve the relationships present in the original data [14]. PCA is a variable reduction technique and is used when variables are highly correlated. It reduces the number of observed variables to a smaller number of principal components which account for most of the variance of the observed variables. The total amount of variance in PCA is equal to the number of observed variables being analyzed. In PCA, observed variables are standardized, e.g., mean=0, standard deviation=1, diagonals of the matrix are equal to 1. The amount of variance explained is equal to the trace of the matrix (sum of the diagonals of the decomposed correlation matrix). The number of extracted components is equal to the number of observed variables in the analysis. The first principal component identified accounts for most of the variance in the data. The second component identified accounts for the second largest amount of variance in the data and is uncorrelated with the first principal component and so on. Components accounting for maximal variance are retained while other components accounting for a trivial amount of variance are not retained. Eigenvalues indicate the amount of variance explained by each component. Eigenvectors are the weights used to calculate components scores [15].

2 Materials and methods

The data were extracted from the monitoring program conducted by Tehran water and wastewater company and includes samplings along Karaj River during a three year period (from April 2006 to March 2009); this program provided a data set of 20 stations (figure 1) in which the samplings were carried out seasonally during April 2007 to March 2008 and almost monthly at other times.

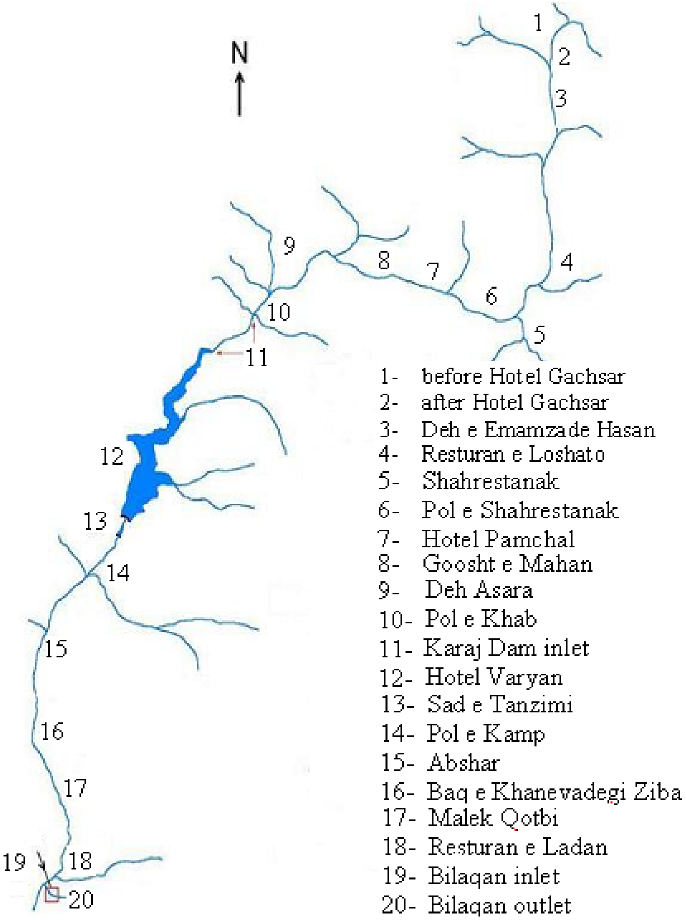

Figure 1: Karaj River.

Due to types of required data for the applied analyses, only 480 out of the many samplings performed were used in this study. Samples were analyzed for 24 variables leading to 11520 observations. The measured variables are: pH, temperature T$^{\circ}$, turbidity, sodium Na, calcium Ca^{2+} , potassium K^{+}, magnesium Mg^{2+}, total hardness TH, phenol- phetalein alkalinity PA, total alkalinity TA, phosphate PO$_4^{3-}$, sulphate SO$_4^{2-}$, silicate SiO$_2^{-}$, ammoniac NH$_3$, nitrite NO$_2^{-}$, nitrate NO$_3^{-}$, chloride Cl^{-}, total dissolved solids TDS, conductivity EC, dissolved oxygen DO, Dissolved Oxygen Percent Saturation (%sat of DO), chemical oxygen demand COD, and biochemical oxygen demand BOD, sodium absorption ratio (SAR). Table 1 presents the basic statistics of the data set.

Two multivariate analytical techniques, cluster analysis and factor analysis have been selected for the data treatment to assess water quality of Karaj River, to determine more influential variables on water quality variation along Karaj

River, to determine pollution sources, cluster sampling sites on the basis of similarity in water quality and assess the effect of Karaj Dam on obtained water quality indices.

Microsoft Office Excel 2007, Minitab 13 and SPSS 13 software packages were employed for data treatment (mathematical and statistical computations).

Table 1: Descriptive statistics for the water quality data used for this study.

parameter	unit	maximum	mean	minimum	Stand. Dev.
temperature	◦C	23	8.89	0	3.83
EC	μS/cm	605	362.15	231	71.55
pH		8.84	8.31	7.84	0.13
Turbidity	NTU	562	22.69	0.6	75.00
TDS	mg/L	359.3	225.55	135.12	44.03
PA	mg/L-$CaCO_3$	20	2.42	0	3.49
TA	mg/L-$CaCO_3$	164	118.01	64	14.84
TH	mg/L-$CaCO_3$	256	159.55	100	30.08
Ca^{2+}	mg/L	84.8	52.41	28.8	10.36
Mg^{2+}	mg/L	17.28	6.84	2.88	2.08
Na^+	mg/L	37	14.62	7	3.87
K^+	mg/L	3	1.66	0.6	0.50
Cl^-	mg/L	50	10.58	3	4.61
SO_4^{2-}	mg/L	125	56.30	24	19.97
SiO_2^-	mg/L	18	9.15	3	2.38
NH_3	mg/L	0.52	0.06	0	0.06
NO_3^-	mg/L	7.2	3.12	0.7	0.97
NO_2^-	mg/L	0.3	0.04	0	0.05
COD	mg/L	7	2.94	0.8	0.92
BOD	mg/L	5.7	1.68	0.5	0.69
PO_4^{3-}	mg/L	0.24	0.03	0	0.03
DO	mg/L	11.5	8.76	6	1.07
%sat of DO	%	102.8694	75.15	58.16609	6.47
SAR	-	1.13752	0.50	0.243673	0.11

2.1 Karaj River

Karaj River located in Tehran province. It is one of the most saturated rivers of the province which originates from Alborz Mountains. This river is about 220 km. in length and has the most copious flow in this region (about 535 MCM per year). The Karaj dam supplies today a major part of Tehran's power and water requirements. The purpose of this dam is to supply water for municipal and agricultural uses. Karaj River has a length and surface watershed about 245 Km and 5000 Km^2 respectively; its width and depth varies from 5 to 8 m and 1 to 3 m respectively. Water of this river is used for industrial, agricultural, and municipal purposes of Tehran province. Its annual mean flow rate at its entrance to Karaj dam is 450 × 106 m^3; minimum and peak annual temporal flow rates in this location are 8.2 and 1450 m^3/s [16].

3 Result and discussion

3.1 Factor analysis

Factor analysis with principal component analysis extraction (PCA) method was investigated. PCA of the data set resulted in seven principal components (PCs) with Eigenvalues >1 accounting for almost 82% of total variance; the higher the eigenvalue, the more significant the corresponding component. Out of the 82% of total variation, PC1, PC2, PC3, PC4, PC5, PC6 and PC7 explain 36%, 11.2%, 9.9%, 7.1%, 6.5%, 6.2% and 5.1%, respectively.

Following results were obtained as summarized in table 2 by using "Varimax rotation method with Kaiser Normalization", removing temperature (loaded under no component), and summing up NO_3-N, NH_3-N and NO_2-N as the nitrogen group (N-group). Axis rotation extracts a new group of variables known as Varifactors (VFs). In contrast to PCs which are linear combinations of

Table 2: Loadings of variables (24) on significant Varifactors for the data set.

Variable	Varifactor						
	1	2	3	4	5	6	7
TH	**.987**	.076	-.042	-.031	.032	-.073	.028
EC	**.950**	.224	.020	-.042	.055	-.116	.051
TDS	**.947**	.288	-.058	-.020	.023	-.077	.022
Ca	**.936**	.111	-.082	.001	.096	-.068	.006
SO_4	**.908**	.222	.067	-.030	.008	-.086	-.138
TA	**.799**	.144	-.261	-.030	.070	-.125	.235
Cl	.653	.525	.151	-.020	.001	-.036	.104
Mg	.627	-.068	.099	-.110	-.178	-.052	.078
SAR	.212	**.956**	-.025	-.003	-.036	-.053	.062
Na	.527	**.835**	-.023	-.018	-.019	-.062	.052
DOsat	-.108	.074	**.764**	.072	.164	-.026	.290
DO	.363	.090	**.752**	-.021	.117	.037	.242
SiO_2	.228	.151	**-.712**	.061	-.027	.151	.207
BOD	-.098	.063	.004	**.918**	-.006	-.045	.043
COD	-.033	-.084	-.011	**.905**	-.043	.116	.067
PA	.018	-.032	.038	.011	**.911**	.000	-.091
PH	.015	-.012	.203	-.068	**.876**	-.110	.061
Turbidity	-.091	-.145	-.009	.023	-.012	**.861**	-.038
PO_4	-.214	.056	-.103	.049	-.092	**.823**	.008
K	-.189	-.091	-.052	.060	-.108	.187	**-.824**
N-group	-.059	.021	.195	.232	-.197	.179	**.717**
Initial Eigenvalue	7.56	2.36	2.07	1.49	1.36	1.31	1.06
% of variance	36.0	11.2	9.86	7.01	6.48	6.23	5.06
Cumulative %	36.0	47.2	57.1	62.2	70.7	77.0	82.0

observed variables, VFs include unobserved, hypothetical, latent water quality variables [17]. VFs and PCs have the same variations. Factor loadings are classified into three groups: 0.3-0.5 as weak, 0.5-0.75 as moderate and >0.75 as strong [18]. Based on this, VF1 (hardness, dissolved solids factor) has high positive loading on TH, EC, TDS, Ca, SO_4 and TA and moderate positive Loading on Cl and Mg. loading of ions (salts groups) on this VF indicates the probable contribution of natural sources e.g. soil erosion. VF2 (Na factor) was loaded highly on Na and SAR. This also represents a natural effect. VF3 shows strongly positive correlation with DO and percentage of DO_{sat}; also strongly negative correlation with SiO_2. This reflects the influence of temperature; higher water temperature leads to lower DO and higher SiO_2 (dissolved soil). Higher temperature leads to lower solubility of oxygen and increases the chemical and biochemical activities which consequently deplete DO while increasing dissolution of soil SiO_2 .VF4 (organic factor) ,strongly correlated with BOD and COD, represents anthropogenic pollution that may have been originated from domestic, commercial and industrial wastewaters-point sources located on the river watershed. VF5 (carbonate alkalinity factor) revealed strongly positive loading on PA and pH which are related to leaching of carbonate compounds out of soil which increases the pH of water. Since in all samples, PA is less than TA/2, PA indicates just the carbonate alkalinity. VF6 has been loaded on turbidity and PO_4. This factor represents non-point source of pollution which can be related to agricultural activities and surface runoff from precipitation. Remaining water quality parameters, K and nitrogen group rooting from non-point source pollutions such as agricultural activities and atmospheric deposition, have been loaded on VF7 (nutrient factor). Loading is however negative for N while positive in the case of K which may be due to use of just one type of the two corresponding fertilizers.

3.2 Cluster analysis

Similarities and differences between water quality sampling sites were identified by using cluster analysis. The clustering method shows groups of similar stations. CA has rendered a dendrogram (figure 2) that classifies stations into two significant clusters: cluster 1 (stations 1, 5, 12-20) and cluster 2 (stations 2-11 excluding 5). Putting stations 1 and 5 aside, cluster 1 and 2 can be defined respectively as downstream- and upstream Karaj Dam clusters as well. This spatial similarity and site clustering makes possible optimal designing of the future sampling strategy that saves both time and costs.

3.3 T test analysis

Results of factor analysis were used to determine water quality differences between two groups. FA correlated variables are loaded on one varifactor, and follow a similar trend, ignoring their signs; therefore comparison between two clusters can be done by comparing the average of parameters loaded on a varifactors (VFs parameters) rather than each parameter. Hence, water quality differences were investigated as follows:

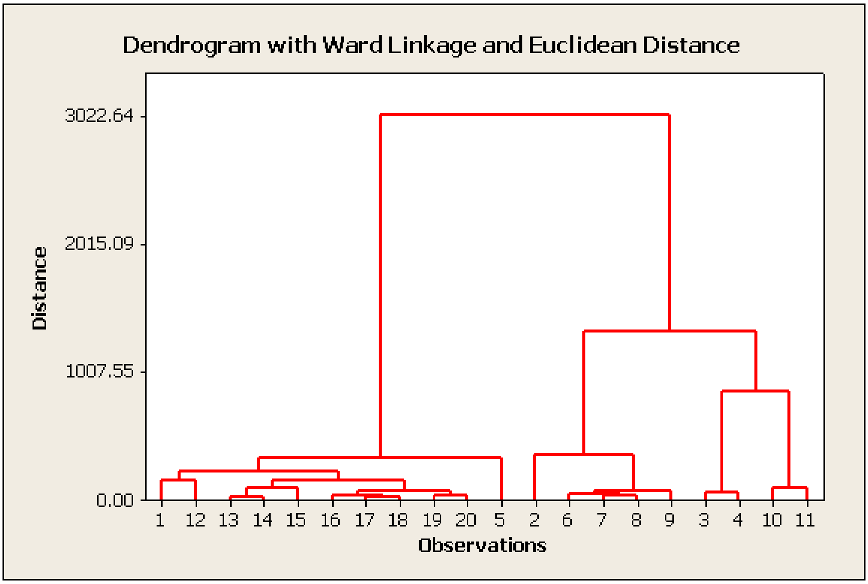

Figure 2: Clustering of monitoring stations on the Karaj River.

1- Transformation of parameters into the standard unit (z) to obtain standardized variables.
2- Calculation of mean standardized values of parameters loaded on each varifactor to reach to an estimation of the value of each varifactor for each of our 480 records.
3- Comparison of calculated averages between two clusters by using independent sample T-test.

T test revealed significant difference between the mean of calculated varifactor values for VF1, VF2, VF6 and VF7 between two clusters (table 3). Our test could not detect any significant difference in the mean of other varifactors (VF3, VF4 and VF5) between two groups, in a confidence interval of 95%.

Similar to cluster 1 and cluster 2, upstream and downstream stations of the dam (excluding stations 1 and 5) have different water quality. Cluster 1 and downstream stations have better quality in VF1, and VF2, VF6, but worse quality in VF7. This result shows the effect of agricultural fertilizers on downstream stations of the dam, although amount of corresponding parameters do not excess of World Health Organisation's drinking water standards.

4 Conclusion

In this case study multivariate techniques such as factor analysis, cluster analysis, and T-test were investigated to determine sources of pollution, group correlated variables, cluster similar sampling sites, and the effect of Karaj Dam on the water quality.

Table 3: Independent Samples T-Test of Varifactors between two clusters.

		Levene's Test for Equality of Variances		t-test for Equality of Means				
	Equal variances	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
VF1	assumed	61.75	.000	-12.22	468	.000	-.8933	.07309
	not assumed			-11.75	354	**.000**	-.8933	.07601
VF2	assumed	1.158	.282	-3.633	468	**.000**	-.3334	.09177
	not assumed			-3.608	434	.000	-.3334	.09242
VF3	assumed	1.101	.295	-1.511	468	**.132**	-.0969	.06414
	not assumed			-1.557	465	.120	-.0969	.06223
VF4	assumed	.436	.509	.691	468	**.490**	.0596	.08626
	not assumed			.693	453	.488	.0596	.08596
VF5	assumed	.228	.633	-.991	468	**.322**	-.1616	.16309
	not assumed			-1.060	386	.290	-.1616	.15242
VF6	assumed	47.27	.000	-3.701	468	.000	-.2945	.07957
	not assumed			-3.449	278	**.001**	-.2945	.08539
VF7	assumed	9.27	.002	2.080	468	.038	.1130	.05429
	not assumed			2.122	468	**.034**	.1130	.05323

The results of factor analysis with principal component analysis showed that natural factors including salts from field sources, organic matter and nutrients explain variation in water quality. Soil leaching rather than anthropogenic pollution plays the main role in water quality variation, though the affect of anthropogenic pollution is considerable. Factor analysis method is useful in summarizing salt water quality, and can help in determining sources of pollution. Cluster analysis techniques classified 20 sampling sites into two groups which indicate its effectiveness for data reduction. Treatment of our data set by cluster analysis technique revealed usefulness of this method for offering reliable classification of stations along Karaj River and data reduction. On the other hand this method is a proper alternative for planning a future spatial sampling strategy in an optimal manner which leads to time and cost benefits. Results of factor analysis technique and T-test verify water quality of two clusters. Stations located after Karaj dam showed better quality. This states that it has a positive effect in water quality, probably because of its role as a precipitation tank. The results could be useful for water quality management regarding monitoring efficiency of Karaj River.

References

[1] Giridharan, L., Venugopal, T., Jayaprakash, M. (2009). *Assessment of Water Quality Using Chemometric Tools: A Case Study of River Cooum, South India.* Arch Environ Contam Toxicol, 56: 654–669.

[2] Alkarkhi, A.F.M., Ahmad, A., Easa, A.M. (2008). Assessment of surface water quality of selected estuaries of Malaysia: multivariate statistical techniques. Environmentalist.

[3] Zhang, Q., Li, Z., Zeng, G., et al. (2008). Assessment of surface water quality using multivariate statistical techniques in red soil hilly region: a case study of Xiangjiang watershed, China. Environ Monit Assess.

[4] Yidanaa, S.M., Ophoria, D., Banoeng-Yakubo, B. (2008). *A multivariate statistical analysis of surface water chemistry data—The Ankobra Basin, Ghana* Journal of Environmental Management, 86(1): 80-87.

[5] Wanga, X., Homerb, M., Dyerb, S D., et al. (2005). *A river water quality model integrated with a web-based geographic information system* Journal of Environmental Management, 75(3): 219-228.

[6] Han, S., Kim, E., Kim, S. (2009). *The Water Quality Management in the Nakdong River Watershed using Multivariate Statistical Techniques.* KSCE Journal of Civil Engineering, 13(2): 97-105.

[7] Shrestha, S., Kazama, F. (2007). *Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan* Environmental Modelling & Software, 22(4): 464-475.

[8] Iscen, C.F., Emiroglu, Ö., Ilhan, S., et al. (2008). *Application of multivariate statistical techniques in the assessment of surface water quality in Uluabat Lake, Turkey.* Environ Monit Assess, 144: 269-276.

[9] Zhang, Y., Guo, F., Meng, W., et al. (2009). *Water quality assessment and source identification of Daliao river basin using multivariate statistical methods.* Environ Monit Assess, 152: 105-121.

[10] Singh, K., Malika, A., Sinha, S. (2005). *Water quality assessment and apportionment of pollution sources of Gomti river (India) using multivariate statistical techniques—a case study* Analytica Chimica Acta, 538: 355-374.

[11] Shtangeeva, I., Alber, D., Bukalis, G., et al. (2009). *Multivariate statistical analysis of nutrients and trace elements in plants and soil from northwestern Russia.* Plant Soil.

[12] Wang, X., Lu, Y., He, G., et al. (2007). *Multivariate Analysis of Interactions between Phytoplankton Biomass and Environmental Variables in Taihu Lake, China.* Environ Monit Assess, 133: 243-253.

[13] Varol, M., Sen, B. (2008). *Assessment of surface water quality using multivariate statistical techniques: a case study of Behrimaz Stream, Turkey.* Environ Monit. Assess.

[14] Ouyang, Y. (2005) *Evaluation of river water quality monitoring stations by principal component analysis* Water Research, 39(12): 2621-2635.

[15] Suhr, D.D., *Principal Component Analysis vs. Exploratory Factor Analysis.* University of Northern Colorado, SUGI 30, Paper 203.

[16] Company, T.W.W., Tehran Water & Wastewater Company, 2006-2009.

[17] Singh, K., Malik, A., Mohan, D., et al. (2004) *Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India)—a case study.* water research, 38: 3980-3992.

[18] Sojka, M., Siepak, M., Zioła, A., et al. (2008). Application of multivariate statistical techniques to evaluation of water quality in the Mała Wełna River (Western Poland). Environ Monit Assess, 147: 159-170.