Article

Multivariate statistical analysis of surface water chemistry: A case study of Gharasoo River, Iran

MH Sayadi [1], **A Rezaei**[1], **MR Rezaei**[1], **K Nourozi**[2]
[1]Environmental Department, University of Birjand, Birjand, Iran
[2]Department of Environmental Protection Kermanshah, Iran
E-mail: mh_sayadi@yahoo.com

Received 8 May 2014; Accepted 15 June 2014; Published online 1 September 2014

Abstract
Regional water quality is a hot spot in the environmental sciences for inconsistency of pollutants. In this paper, the surface water quality of the Gharasoo River in western Iran is assessed incorporating multivariate statistical techniques. Parameters like EC, TDS, pH, HCO_3^-, Cl^-, SO_4^{2-}, Ca^{2+}, Mg^{2+} and Na^+ were analyzed. Principal component and factor analysis is showed the parameters generated 3 significant factors, which explained 73.06% of the variance in data sets. Factor 1 may be derived from agricultural activities and subsequent release of EC, TDS, SO_4^{2-} and Na^+ to the water. Factor 2 could be influenced by domestic pollution and explained the deliverance of HCO_3^-, Cl^- and Mg^{2+} into the water. Factor 3 contains hydro-geochemical variable Ca^{2+} and pH, originating from mineralization of the geological components of bed sediments and soils of watershed area. Likewise, the clustering analysis generated 3 groups of the stations as the groups had similar characteristic features. Pearson correlation analysis showed significant correlations between HCO_3^- and Mg^{2+} (0.775), Ca^{2+} (0.552) as well as TDS and Na^+ (0.726). With reference to multivariate statistical analyses it can be concluded that the agricultural, domestic and hydro-geochemical sources are releasing the pollutants into the Gharasoo River water.

Keywords anthropogenic activities; geological components; Gharasoo River; PCA; water quality.

Proceedings of the International Academy of Ecology and Environmental Sciences
ISSN 2220-8860
URL: http://www.iaees.org/publications/journals/piaees/online-version.asp
RSS: http://www.iaees.org/publications/journals/piaees/rss.xml
E-mail: piaees@iaees.org
Editor-in-Chief: WenJun Zhang
Publisher: International Academy of Ecology and Environmental Sciences

1 Introduction

The surface water quality is truly a sensitive issue today because of its effects on human health and aquatic ecosystems. Rivers are highly vulnerable to pollution attributing to their role in carrying off the municipal and industrial wastewater and runoff from agriculture in their vast drainage basins. Anthropogenic influences, as well as natural processes, deteriorate surface water and impair their use for drinking, industrial, agricultural and recreational purposes (Sayadi et al., 2010; Sayadi and Sayyed, 2011; Sayyed and Sayadi, 2011; Sayadi and Rezaei, 2014).

The environmental impacts of agricultural runoff, municipal wastewater and industrial effluent discharge on receiving water are numerous and inputs of contaminants can be hazardous to the environment. Surface water is most vulnerable to pollution due to natural processes, such as precipitation inputs, erosions, weathering of crystal materials, and anthropogenic influences (Sundaray et al., 2006).

From many methods available, Cluster Analysis (CA) and Principal Component Analysis (PCA) are widely used, being capable of detecting similarities among samples and/or variables (Acque et al., 1995; Wang et al., 2006; Wenning and Erickson, 1994). In this work, PCA and hierarchical cluster analysis have been used to investigate the Gharasoo river water quality and discriminate relative magnitude of anthropogenic and natural influences on the river water quality. The objective of the study is to extract information about; (1) the similarities or dissimilarities between monitoring periods and monitoring sites, (2) the influence of possible sources (natural and anthropogenic) on water quality parameters, and (3) source identification for estimating possible sources determining water quality parameters of the Gharasoo River.

2 Material and Methods

2.1 Study Area

The Gharasoo River system forms with joining of two main branches of Merek and Raz Avar Rivers. Its length is approximately 20.7 kilometers and runs through the city of Kermanshah. Kermanshah city is located in the diplomatic area in western Iran. Gharasoo River is one of the most important tributaries of Saymareh River. It collects water from Kermanshah and Kurdistan provinces and delivers it to Saymareh River. The study area lies between latitudes 46° 36´ - 47° 37´ N and longitudes 34° 00´- 34° 91´ E with the height of 1322 meters above sea level (Fig. 1). The mean annual temperature and rainfall are 14°C and 456.8 mm, respectively. Rainfall occurs during the autumn and winter (December to March).

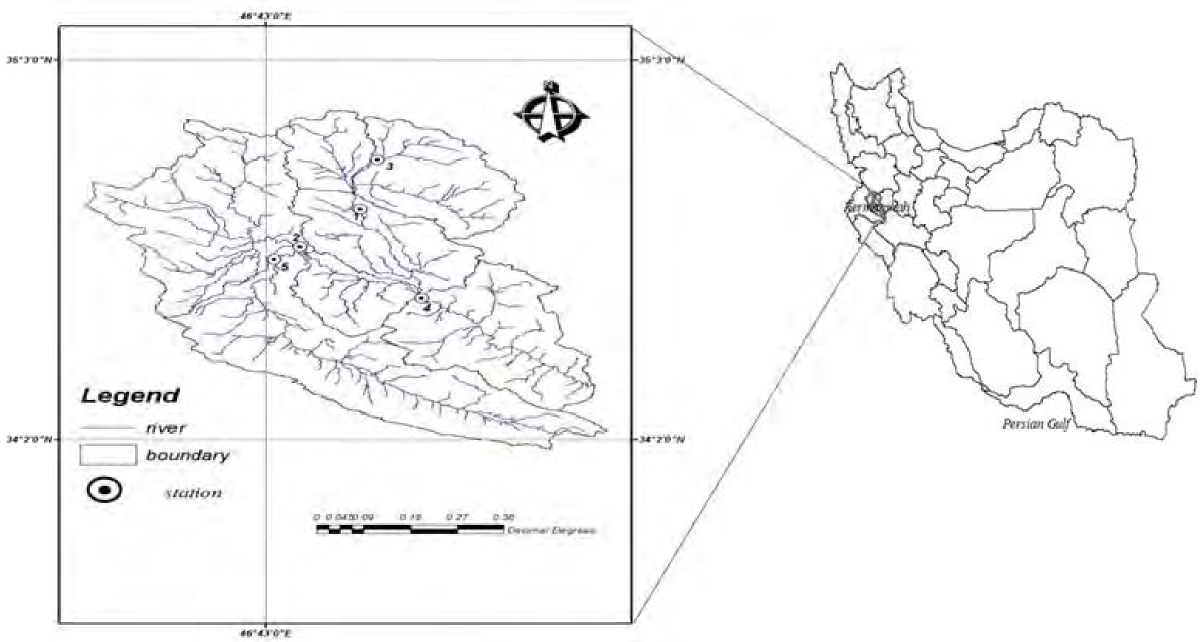

Fig. 1 Location of sample sites on the Gharasoo River.

Data sets of 9 parameters of the water quality were monitored monthly over a period of 2009-2010. Monitoring stations are shown in (Fig. 1). The selected parameter for the determination of water quality characteristics were EC, PH, TDS, bicarbonate (HCO_3^-), chloride (Cl^-), sulfate (SO_4^{2-}), calcium (Ca^{2+}), magnesium (Mg^{2+}) and sodium (Na^+). The parameters were analyzed according to standard methods (APHA-AWWA-WPCF, 1985; APHA, 1999). The results were evaluated via multivariate statistical analysis techniques. All statistical computations were made using SPSS statistical software.

2.2 Principal Component Analysis (PCA)

PCA is designed to transform the original variables into new and uncorrelated variables called the principal components, which are linear combinations of the original variables (Zhang, 2011; Vieira, 2012). It provides information on the most significant parameters due to spatial and temporal variations that describes the whole data set by excluding the less significant parameters with minimum loss of the original information (Helena et al., 2000; Kannel et al., 2007). The principal component can be expressed as

$$Z_{ij}=a_{i1}\, x_{1j} + a_{i2}\, x_{2j} + \ldots + a_{im}\, x_{mj} \qquad (1)$$

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

Factor analysis follows principal component analysis. The main purpose of factor analysis is to reduce the contribution of less significant variables and to simplify even more the data structure coming from the principal component analysis. This purpose can be achieved by rotating the axis defined by principal component analysis according to well established rules, and constructing new variables, also called vary factors. A small number of factors will usually account for approximately the same amount of information as does the much larger set of original observations (Shrestha and Kazama, 2007). The Factor analysis can be expressed as:

$$Z_{ji} = a_{f1}\, f_{1i} + a_{f2}\, f_{2i} + a_{f3}\, f_{3i} + \ldots + a_{fm} f_{mi} + e_{fi} \qquad (2)$$

where **z** is the measured value of a variable, a is the factor loading, f is the factor score, e is the residual term accounting for errors or other sources of variable number, and m is the total number of factors.

2.3 Cluster Analysis (CA)

CA is a multivariate technique, whose primary purpose is to classify the objects of the system into categories or clusters based on their similarities (Zhang, 2012), and the objective is to find an optimal grouping for which the observation or objects within each cluster are similar, but the cluster is dissimilar to each other. Hierarchical clustering is the most common approach in which clusters are formed sequentially. The most similar objects are first grouped, and these initial groups are merged according to their similarities. Eventually as the similarity decreases all subgroups are merged into a single cluster. CA was applied to surface water quality data using a single linkage method. In the single linkage method, the distances or similarities between two clusters A and B are defined as the minimum distance between a point A and a point in B:

$$D\,(A, B) = \min \{d\,(x_i+x_j), \text{ for } x_i \text{ in A and } x_j \text{ in B}\} \qquad (3)$$

where d (x_i +x_j) is the Euclidean distance in (3). At each step the distance is found in every pair of clusters and the two clusters with smallest distance are merged. When over two clusters are merged the procedure is repeated for the next step: the distances between all pairs of clusters are calculated again, and the pair with the minimum distance is merged into a single cluster. The result of a hierarchical clustering procedure can be displayed graphically using a tree diagram, also known as a dendrogram, which shows all the steps in the hierarchical procedure (Alkarkhi et al., 2008; Johnson and Wichern, 2002).

3 Results and Discussion

Table 1 summarizes briefly the mean, maximum and minimum values besides standard deviation and variance of the 9 measured parameters in the river water samples from the five stations. It is interesting to note that the high standard deviations of the parameters indicating changeability in chemical composition between the samples, shows the temporal variations which appears by lithogenic and anthropogenic sources.

Table 1 Simple statistical analysis of water quality parameters at different locations on the Gharasoo River.

Station		EC	pH	TDS	HCO_3^-	Cl	SO_4^{2-}	Ca	Mg	Na
Station1	Mean	372	7.8	2.4	3.3	0.52	0.49	2.7	1.3	0.36
	Std.	53.8	0.46	35.2	0.55	0.22	0.27	0.42	0.38	0.17
	Variance	290	0.21	1.24	0.31	0.05	0.07	0.18	0.15	0.03
	Minimum	172	6.53	108	2.31	0.16	0.1	1.91	0.56	0.09
	Maximum	437	8.57	280	5.06	1.10	1.29	3.41	2.24	0.91
Station2	Mean	437	7.7	280	3.73	0.56	0.49	2.9	1.47	0.43
	Std.	96	0.46	61	0.8	0.25	0.27	0.43	0.65	0.31
	Variance	91.70	0.22	37.45	0.6	0.06	0.07	0.19	0.42	0.09
	Minimum	329	6.70	211	2.56	0.16	0.14	2.01	0.78	0.16
	Maximum	661	8.52	423	5.43	0.96	0.92	3.55	2.80	1.16
Station3	Mean	404	7.79	285	3.43	0.66	0.59	2.94	1.42	0.38
	Std.	56.61	0.36	36.18	0.68	0.14	0.28	0.42	0.42	0.08
	Variance	321	0.13	130	0.46	0.02	0.081	0.18	0.18	0.01
	Minimum	320	7.04	205	2.30	0.38	0.20	2.37	0.81	0.25
	Maximum	540	8.40	346	5.11	0.91	1.36	3.37	2.40	0.59
Station4	Mean	434	7.86	275	3.58	0.52	0.72	2.77	1.56	0.55
	Std.	111	0.37	71.51	0.93	0.22	0.52	0.50	0.58	0.53
	Variance	124	0.13	51.14	0.87	0.05	0.28	0.25	0.34	0.28
	Minimum	312	7.19	199	2.10	0.22	0.16	1.41	0.56	0.06
	Maximum	663	8.66	424	6.46	1.00	2.71	3.49	3.00	2.20
Station5	Mean	494	7.37	336	4.00	0.43	0.91	3.00	1.94	0.45
	Std.	491	0.13	199	0.29	0.04	0.14	0.08	0.41	0.03
	Variance	241	0.02	397	0.08	0.01	0.02	0.01	0.17	0.00
	Minimum	340	7.32	320	3.44	0.36	0.64	2.80	1.00	0.39
	Maximum	520	7.80	390	4.38	0.50	1.11	3.13	2.50	0.49

3.1 Application of PCA to Gharasoo River data set

A particular problem in the surface water quality monitoring is the complexity associated with analyzing a large number of measured variables (Saffran et al., 2001). Therefore, in this study, surface water quality data were grouped using FA. The correlation matrix of variables was generated and factors were extracted by the centroid method, rotated by Varimax. From the results of the FA, the first three eigenvalues were found to be bigger than 1 (Fig. 2). According to the Fig. 2 and a subsequent interpretation of the factor loadings, the first three components were extracted and the other components have been eliminated.

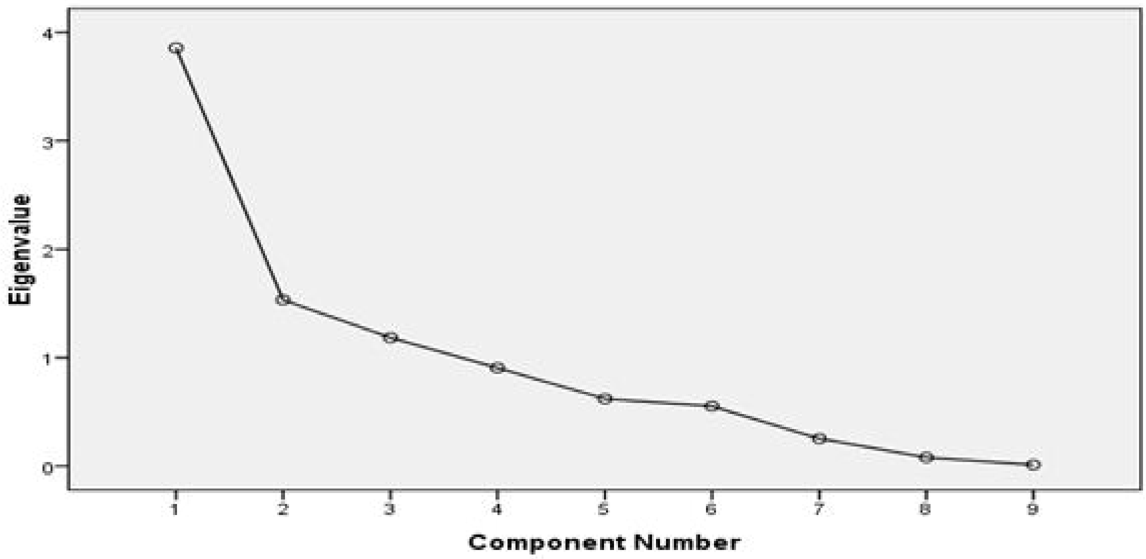

Fig. 2 Screen plot of the eigenvalue and component number.

Table 2 presents the total variance explained by the first three factors for both related and unrelated factor loadings. The parameter loading three factors in the two from FA associated with each factor stations are well defined and contribute slightly to other factors, which help not only in the interpretation of the results but also in the identification of anthropogenic sources of pollution from the surface water quality data. FA generated three significant factors, which explained 73.06% of the variance in data sets, where a correlation greater than 0.75 is considered “strong”; 0.75-0.50, “moderate”; and 0.50-0.30, as “weak” significant factor loading (Liu et al., 2003).

Table 2 Extracted values of various FA parameters.

Component	Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.85	42.86	42.86	2.81	31.23	31.23
2	1.53	17.03	59.89	2.47	27.49	58.72
3	1.18	13.16	73.06	1.29	14.33	73.06

Table 3 Loadings of 9 experimental variables on 3 significant Principal components, rotated factor loadings matrix

Variables	Factor 1	Factor 2	Factor 3
EC	**0.831**	0.360	0.087
pH	-0.180	0.180	**-0.687**
TDS	**0.858**	0.306	0.102
HCO_3^-	0.102	**0.844**	0.422
Cl	0.220	**0.720**	-0.221
SO_4^{2-}	**0.740**	-0.184	0.198
Ca	-0.040	0.389	**0.716**
Mg	0.240	**0.820**	0.046
Na	**0.829**	0.361	-0.136

As shown in tables 2 and 3, the first factor (Factor 1), accounted for 31.23 of the total variance, was high positive loading in EC, TDS, $SO_4^{2-}$and Na^+ which were 0.831, 0.858, 0.740 and 0.829 respectively. This factor represents the contribution of nonpoint pollution from agricultural areas. In these areas, farmers use sulfate fertilizers, and the stream receives sulphate via surface runoff and irrigation waters. As a result shown increase in SO_4^{2-} concentrations may be due to agricultural activities (Krouse, 1997). The contribution of Na^+ to this factor can be considered a result of action-exchange processes in soil–water interface (Guo and Wang, 2004). Factor 2 explains 27.49% of the total variance and is the positively correlated with HCO^-, Cl^- and Mg^{2+}. This factor represents the contribution of point pollution and the physico-chemistry of the stream. While point pollution is from domestic wastewater, nonpoint pollution is from agricultural and livestock farms. Mg^{2+} is a basic metal which increases alkalinity of the environment (Razmkhah et al., 2010). This factor may also be due to anthropogenic activities such as domestic waste water or influents. Nevertheless, the release of domestic effluents into the river water caused the dramatic Cl^- increase. The loading for factor 3 was 18.92% with Ca^{2+} and pH. Thus, this factor contains hydro-geochemical variable Ca^{2+}, originating, at a first glance, from mineralization of the geological components of soils as well as moderate decrease of pH concentration. The contribution of Ca^{2+} to this factor can be considered a result of action-exchange processes in soil–water interface (Guo and Wang, 2004) as the results demonstrated an increase in EC, TDS, SO_4^{2-} and Cl^- concentrations due to agricultural and domestic waste water activities. Sources of dissolved SO_4^{2-} in natural river waters may include dissolution of sedimentary sulfates, oxidation of both sulfide minerals and organic materials, and anthropogenic inputs.

3.2 Pearson correlation

Statistical analysis using Pearson correlation showed that the parameters in the water samples collected from Gharasoo river were weak and moderately correlated to each other at $p < 0.01$ and $p < 0.05$ levels. A significant positive correlation was found to exist between EC and TDS (0.918), HCO_3^- (0.432), Cl^- (0.381), SO_4^{-2} (0.411), Mg^{2+} (0.421), and a positive correlation was found between EC and Na^+ (0.721) at $p < 0.01$ (Table 4). Similarly, there were significant correlations between TDS and HCO_3^- (0.387), Cl^- (0.369), SO_4^{2-}(0.434), Mg^{2+} (0.385) and a positive correlation between TDS and Na^+ (0.726) at $p < 0.01$ in the collected water samples of the study region. The level of TDS reflects the pollutant burden of the water. High levels of dissolved and suspended solids in water systems increase the biological and chemical oxygen demand (Jonnalagadda and Mhere, 2001). Similarly, some correlations were also observed (Table 4) between HCO_3^- and Cl^- (0.373), Na^+ (0.387) and a positive strong correlation was found between HCO_3^- and Mg^{2+} (0.775) and Ca^{2+} (0.552) at $p < 0.01$. Likewise, Li and Zhang (2009) indicated a strong positive correlation between HCO_3^-, Ca^{2+} and Mg^{2+}in Geochemistry of the upper Han River basin, China. There were as well significant positive correlations between Cl^- - Mg^{2+} (0.506), Cl^--Na^+ (0.453), SO_4^{-2} - Na^+ (0.529), and Mg^{2+} - Na^+ (0.459) at $p < 0.01$ level (Table 4). Chloride concentration is higher in wastewater than raw water because sodium chloride, the commonest component of the human diet passes unchanged through the digestive system (WHO, 2008).

It is interesting to note that in this study there is no significant correlation between pH and other parameters. Similarly, Chigor et al. (2012) exhibited nil correlation between pH and other contaminated parameters in surface water sources used for drinking and irrigation in Zaria, Nigeria.

3.3 Hierarchical cluster analysis (HCA)

Spatial similarity and monitoring stations grouping is shown in Fig. 3. In this study, the classification of monitoring stations was performed incorporating HCA, and a dendrogram was composed.

Table 4 Pearson correlation between different water quality parameters of the study site.

Discriminate variables	EC	pH	TDS	HCO_3^-	Cl	SO_4^{-2}	Ca	Mg	Na
EC	1.00								
pH	-0.126	1.00							
TDS	0.918**	-0.177	1.00						
HCO_3^-	0.432**	-0.140	0.387**	1.00					
Cl	0.381**	0.070	0.369**	0.373**	1.00				
SO_4^{-2}	0.411**	-0.086	0.434**	-0.039	0.014	1.00			
Ca	0.169	-0.096	0.148	0.552**	0.175	0.204*	1.00		
Mg	0.421**	-0.017	0.385**	0.775**	0.506**	0.166	0.180*	1.00	
Na	0.721**	-0.017	0.726**	0.387**	0.453**	0.529**	0.000	0.459**	1.00

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

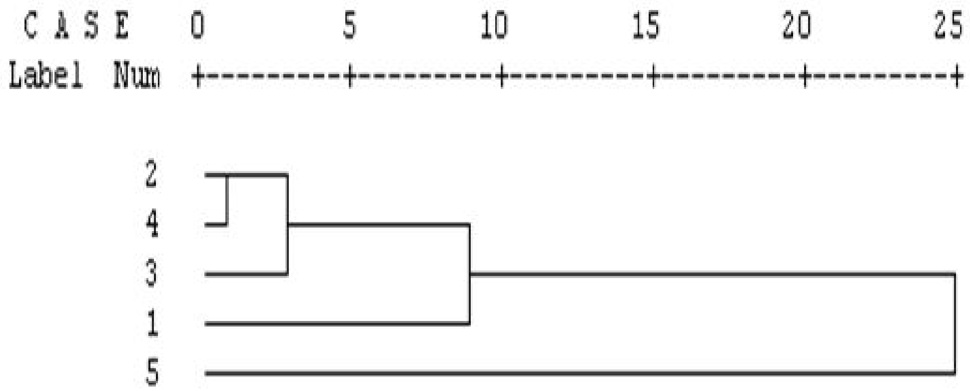

Fig. 3 Dendrogram of the CA according to single linkage method.

The clustering procedure generated 3 groups of stations in a very convincing way, as the sites in these groups have similar characteristic features and natural background source types. Cluster 1 (Stations 2, 3 and 4), Cluster 2 (Station 1) and Cluster 3 (Station 5) correspond to a relatively low to high polluted regions. Hence, the temporal variation in the Gharasoo river water quality was greatly determined by agricultural and municipal activities as well as lithogenic sources which confirm the result of the PCA. In fact, Fig. 3 shows that the patterns of pollution sources of Gharasoo river water.

4 Conclusion

In this study, multivariate statistical methods including factor, principal component and cluster analysis were applied to surface water quality data sets obtained from the Gharasoo River in Iran. The results suggest that anthropogenic activities such as agricultural and domestic pollution sources and lithogenic activities had significant effects on water quality. Three factors explaining the 73.06% of the total variance in the surface water quality data set were determined. Based on the above results, Factor 1 may be derived from agricultural activities and release the EC, TDS, SO_4^{2-} and Na^+ to the environment. Factor 2 could be influenced by domestic pollution and explained the deliverance of HCO^-_3, Cl^- and Mg^{2+} into the surface water of Gharasoo River. Factor 3 contains hydro-geochemical variable Ca^{2+} and pH, originating from mineralization of the geological components of bed sediments and soils of watershed area. Cluster analysis grouped the monitoring stations into 3 clusters of similarity based upon water quality characteristics at different stations. These results reveal that agricultural, domestic and hydro-geochemical sources are responsible for pollutions in terms of water quality in Gharasoo River.

Acknowledgment

The authors would like to appreciate the Department of Environment, Head Office, Kermanshah city for their cooperation and support. Authors are appreciated the authorities of Research Council and Faculty of Natural Resources and Environment, University of Birjand, due to their sincere cooperation. We also like to thank Dr. Mrs. Mahavash F. Kavian for editing the paper.

References

Acque RS, Battegazzore M, Renoldi M. 1995. Integrated chemical and biological evaluation of the quality of the River Lambro (Italy). Water Air and Soil Pollution, 83(3-4): 375-390

Alkarkhi A FM, Ahmad A, Easa A M. 2008. Assessment of surface water quality of selected estuaries of Malaysia: multivariate statistical techniques. The Environmentalist, 29(3): 255-262

APHA-AWWA-WPCF. 1985. Standard methods for the Examination of Water and Wastewater (16^{th} ed), American Public Health Association, American Water Works Association, Water Pollution Control Federation, USA

American Public Health Association (APHA). 1999. Standard methods for the examination of water and wastewater (20th ed.). Washington D.C.: APHA, AWWA, Washington DC, USA

Chigor VN, Umoh VJ, Okuofu CA, et al. 2012. Water quality assessment: surface water sources used for drinking and irrigation in Zaria, Nigeria are a public health hazard. Environmental monitoring and assessment, 1845: 3389-3400

Guo H, WangY. 2004. Hydrogeochemical processes in shallow quaternary aquifers from the northern part of the Datong Basin, China. Applied Geochemistry, 19(1): 19-27

Helena B, Pardo R, Vega M, Barrado E. 2000. Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga River, Spain) by principal component analysis. Water Research, 34(3): 807-816

Johnson RA, Wichern DW. 2002. Applied Multivariate Statistical Analysis (Vol. 5). 1-13, Prentice Hall Upper Saddle River, NJ, USA

Jonnalagadda SB, Mhere G. 2001. Water quality of the odzi river in the Eastern Highlands of Zimbabwe. Water Research, 35(10): 2371-2376

Kannel PR, Lee S, Kanel SR, Khan S P. 2007. Chemometric application in classification and assessment of monitoring locations of an urban river system. Analytica Chimica Acta, 582(2): 390-399

Krouse HR. 1997. Application of the stable isotope composition of S O 4 to tracing anomalous TDS in Nose Creek , southern Alberta , Canada IRegional Bow River. Applied Geochemistry, 12(97): 567-575

Li S, Zhang Q. 2009. Geochemistry of the upper Han River basin, China. 2: Seasonal variations in major ion compositions and contribution of precipitation chemistry to the dissolved load. Journal of hazardous materials, 170(2-3): 605-611

Liu CW, Lin KH, Kuo YM. 2003. Application of factor analysis in the assessment of groundwater quality in a blackfoot disease area in Taiwan. The Science of the Total Environment, 313(1-3): 77-89

Razmkhah H, Abrishamchi A, Torkian A. 2010. Evaluation of spatial and temporal variation in water quality by pattern recognition techniques: A case study on Jajrood River (Tehran, Iran). Journal of environmental management, 91(4): 852-860

Saffran K, Environment A, Cash K, Canada E. 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, Canada

Sayadi MH, Rezaei MR. 2014. Impact of land use on the distribution of toxic metals in surface soils in Birjand city, Iran. Proceedings of the International Academy of Ecology and Environmental Sciences, 4(1): 18-29

Sayadi MH, Sayyed MRG, et al. 2010. Short-term accumulative signatures' of heavy metal in river bed sediments, Tehran Iran. Environmental Monitoring and Assessment, 162: 465-473

Sayadi MH, Sayyed MRG. 2011. Comparative assessment of baseline concentration of the heavy metals in the soils of Chitgar Industrial Area Tehran (Iran) with the comprisable reference data. Environmental Earth Sciences, 63(6): 1179-1188

Sayyed MRG, Sayadi MH. 2011. Variations in the heavy metal accumulations within the surface soils from the Chitgar Industrial Area of Tehran (Iran). Proceedings of the International Academy of Ecology and Environmental Sciences, 1(1): 27-37

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

Sundaray SK, Panda UC, Nayak BB, Bhatta D. 2006. Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of the Mahanadi river-estuarine system (India)--a case study. Environmental geochemistry and health, 28(4): 317-330

Vieira VMNCS. 2012. Permutation tests to estimate significances on Principal Components Analysis. Computational Ecology and Software, 2(2): 103-123

Wang YYS, Lou ZZP, Sun CC C, Wu MML, Han SH. 2006. Multivariate statistical analysis of water quality and phytoplankton characteristics in Daya Bay, China, from 1999 to 2002. Oceanologia, 48(2)

Wenning RJ, Erickson GA. 1994. Interpretation and analysis of complex environmental data using chemometric methods. TrAC Trends in Analytical Chemistry, 13(10): 446-457

World Health Organization. 2008. Guideline to drinking water quality (3rd ed., Vol. 1). World Health Organization, Geneva, Switherlands

Zhang WJ. 2011. Simulation of arthropod abundance from plant composition. Computational Ecology and Software, 1(1): 37-48

Zhang WJ. 2012. Computational Ecology: Graphs, Networks and Agent-based Modeling. World Scientific, Singapore