علـوم محـیـطی سال هشتم، شماره دوم، زمستان ۱۳۸۹
ENVIRONMENTAL SCIENCES Vol.8, No.2 , Winter 2011
137-146

Surface Water Quality Assessment Using Cluster Analysis: a Case Study of the Gharasou River Basin, Iran

Ebrahim Fataei,[1*] Seied Masoud Monavari, [1]Amir Hesam Hasani, [2]Seied Ahmad Mirbagheri [3] and Abdoreza Karbasi

1- Department of Environmental Management, Graduate School of the Environment and Energy, Islamic Azad University, Science and Research Branch, Tehran, Iran.
2- Department of Civil Engineering, Faculty of Engineering, Kh.N.Toosi University of Technology, Tehran, Iran.
3- Department of Environmental Management and Planning, Graduate Faculty of the Environment, Tehran University, Iran

Abstract

For assessment of water quality in Gharasou River, located in the Northwest of Iran, multivariate statistical analysis was used. During the period of one year, 18 physical, chemical and microbiological parameters were sampled in 11 sampling stations. The measured data were analyzed by a multivariate statistical approach, Cluster Analysis (CA). Based on CA analysis the stations were divided to three groups of highly polluted (HP), moderately polluted (MP), and less polluted (LP) stations. The results of the study revealed that multivariate statistical techniques are an effective statistical method for water quality assessment, identification of pollution sources/factors in water quality for effective water quality management. As Extracted clustered information can be used in reducing the number of sampling sites on the River without missing much information.

Keywords: Water quality, Pollutant sources, multivariate statistical techniques, Surface water.

ارزیابی کیفیت آب‌های سطحی با استفاده از آنالیز کلاستر: (مطالعه موردی رودخانه قره سو، ایران)

ابراهیم فتائی[1*]، سید مسعود منوری[1]، امیرحسام حسنی[1]، سید احمد میرباقری[2]، عبدالرضا کرباسی[3]

۱- گروه مدیریت محیط زیست ، دانشکده محیط زیست و انرژی، واحد علوم و تحقیقات دانشگاه آزاد اسلامی
۲- گروه مهندسی عمران، دانشکده فنی، دانشگاه خواجه نصیرالدین طوسی
۳- گروه مدیریت و برنامه‌ریزی محیط‌زیست، دانشکده محیط زیست، دانشگاه تهران

چکیده

این تحقیق به منظور ارزیابی کیفی رودخانه قره سو که در شمال غربی ایران در استان اردبیل واقع شده است؛ انجام گرفت. نمونه‌برداری در طول یکسال بر روی ۱۸ پارامتر فیزیکی، شیمیایی و بیولوژیکی در ۱۱ ایستگاه انجام گردید. نتایج حاصل از اندازه‌گیری‌ها با استفاده از روش‌های تحلیل چند متغیره تجزیه کلاستر مورد تجزیه و تحلیل قرار گرفتند. بر اساس نتایج حاصل از تجزیه کلاستر ایستگاه‌ها به سه گروه با آلایندگی زیاد(HP)، آلایندگی متوسط (MP) و آلایندگی کم (LP) تقسیم گردیدند. در گروه HP ایستگاههای ۱۰ و ۱۱، در گروه MP ایستگاه ۲ و بقیه ایستگاه‌ها در گروه LP قرار گرفتند. نتایج این مطالعه نشان‌دهنده سودمندی تکنیک‌های آماری چند متغیره در ارزیابی کیفی منابع آب، تشخیص منابع و عوامل آلاینده برای مدیریت موثر کیفیت منابع آبی می‌باشد. بطوری که می توان از اطلاعات خوشه‌بندی شده در راستای کاهش تعداد ایستگاه‌های نمونه‌برداری در رودخانه استفاده نمود بدون این که اطلاعات چندانی از دست رود.

کلمات کلیدی: کیفیت آب ، منابع آلاینده ، روش‌های آماری چندمتغیره، آب سطحی.

* Corresponding author. *E-mail Address*: ebfataei@gmail.com

Introduction

Multivariate statistical techniques are increasingly used in water quality assessment, environmental analysis and the selection of qualitative variables for establishment of water quality monitoring plans (Astel *et al.*, 2006; Zhang *et al.,* 2008). Recently, the use of Principle Component Analysis (PCA) and Factor Analysis (FA) has become common in the indicator variables reduction and better interpretation of the findings (Ouyang, 2005). PCA, FA, CA and discriminate analysis (DA) methods were used in water quality assessment and the apportionment of pollution sources in some other studies. For example, such PCA and CA were used in surface water quality assessment in Tahtali River of Turkey (Boyacioglu, 2007) or in assessment of temporal and spatial fluctuations of the water quality in Gomti River in India (Singh *et al.,* 2005), Daliao River in China (Zhang *et al.,* 2008), and Fuji River in Japan (Shrestha & Kazama, 2007). Water quality in Daliao River, was assessed by using CA, DA and PCA (Zhang *et al.,* 2008). In the recent study, the stations under surveillance were divided into three groups based the sources of pollution. Water quality assessment and controlling water pollution sources in Gomti River (Singh *et al.,* 2004) and in Fuji River, (Shrestha & Kazama, 2007) were investigated using the CA, PCA, FA and DA methods. Boyacioglu (2007) had shown that the multivariate statistical method can be useful in parameters and station numbers reduction and in surface water quality studies as well. Geostatistical kriging was used for designing a water quality monitoring system in Karoun River, Iran (Karamouz *et al.,* 2005). In another study that had been carried out on the Karoun River, one of eight sampling stations was identified as a subordinate station based on PCA analysis (Nouri *et al.,* 2007). The goal of this paper is to introduce the use of multivariate statistical techniques in Gharasou River water quality assessment. In the present study, the efficiency of the cluster analysis technique was applied to evaluate spatial and temporal variations in the water quality data matrix of the Gharasou River (Iran). This study is subjected to CA to extract information about the similarities or dissimilarities between sampling sites, identification of water quality variables responsible for spatial and temporal variations in river water quality and influence of possible sources (natural and anthropogenic) on the water quality parameters, which were generated under the one year monitoring program.

Materials and Methods

Study Area

Gharasou River is one of the main branches of Aras River to the west of the Caspian sea, in Ardabil Province, northwestern Iran. This river originates in the Sabalan and Baghro mountains and exits the Ardabil plain at Samian hydrometric station after joining other streams. This river round has three hydrologic units with a length of 255 km and average slope of 5.7% (Ardabil Regional Water Organization, 2004). The average water discharge of this river, calculated over the long term (30 years) at Samian station, is about 228 cubic meters per year (Ardabil Regional Water organization, 2005). Because of construction of the Yamchi and Sabalan dams upstream and downstream respectively, as well as its role in water consumption for both irrigational and drinking purposes, the river's location has high strategic importance in the study area.

Since the River passes through three urban (Ardabil, Nir, and Sarein) and several rural areas as well as, vast farmlands, and some manufacturing units already established or under construction, it is quite naturally exposed to various sources of the pollution. In order to decrease in the River's water production, on the one hand, and because of the ever-increasing amount of water consumption and urban, industrial, and agricultural sewage discharges, on the other, the water

quality of the river is endangered. Since Ardabil is an agricultural center and is in the process of development, constant monitoring of the water quality in the river is necessary. This research was done to control qualitative variables in Gharasou River for this same reason.

Methods of sampling and analyzing parameters

Sampling stations were selected with regard to natural conditions and access as well as various natural or anthropogenic topographical features such as subordinate river branches, structural geological changes and point and nonpoint pollution sources. The coordination of each sampling station was recorded and was situated on the map by means of GPS (Fig. 1).

Eleven sampling stations were chosen. The sampling process was carried out in a year-long process from September 2007 to September 2008. Samples were collected each month at all eleven sites. The analyses were carried out based on the 21st edition of the Standard Methods for the Examination of Water and Wastewater of the American Public Health Association. Sampling and analysis was conducted on 18 physical and chemical and microbiological parameters. These parameters, including Temp., pH, Turb., DO, BOD_5, COD, EC, TColi., FColi., NO_3^-, PO_4^{3-} , NH_3, Ca^{2+},Mg^{2+}, Na^+, Cl^-, SO_4^{2-} and HCO_3^- were sampled monthly (Table 1).

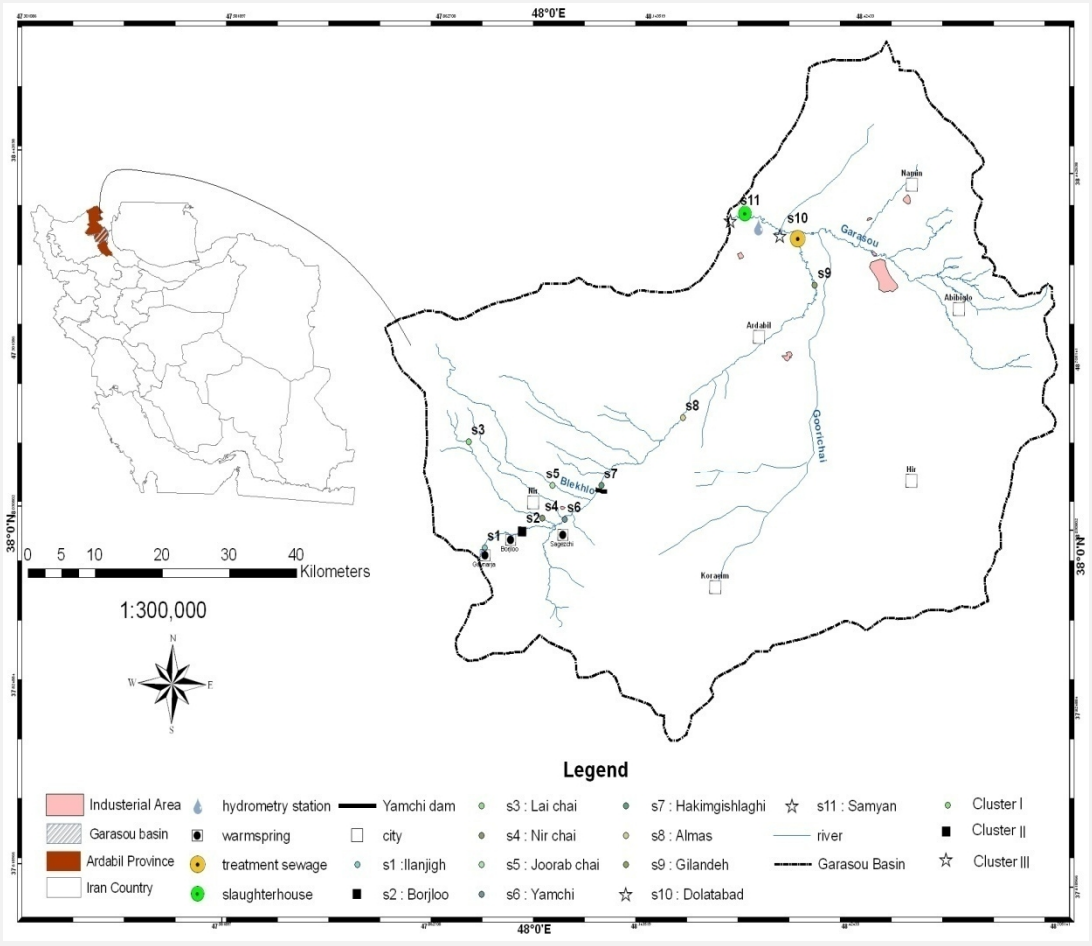

Figure 1- Study area and sampling stations (listed 1-11) in the Gharasou River basin of Ardabil, Iran.

Table 1- Water quality parameters, units and analysis methods.

Parameters	Units	Analytical methods	
EC	μ S cm^{-1}	Portable	Sension156 Hach
DO	mgl^{-1}	Portable	Sension156 Hach
Turb.	NTU	Portable	Sension156 Hach
pH	pH unit	Portable	Sension156 Hach
WT	C°	Portable	Sension156 Hach
NO_3^-	mgl^{-1}	Spectrophotometric	Spectrophotometric
NH_3	mgl^{-1}	Ammonia with vario power pack	Spectrophotometric
COD	mgl^{-1}	COD Total with Vario Tube	Spectrophotometric
BOD	mgl^{-1}	Instrumental method	Dichromate reflex method
TColi.	MPN/100ml	Nine tubes system	Winkler azide method
FColi.	MPN/100ml	Nine tubes system	Multiple tube method
HCO_3	Meq/l	Using HCL and Phenol phthalein	Multiple tube method
Cl^-	Meq/l	Titration with Agcl and Chromate potassium tracer	Titrimetric
SO_4^{2-}	Meq/l	DR2500 Spectrophotometer	Spectrophotometric
Ca^{2+}	Meq/l	Titration with EDTA and Mvraksayd in alkalinity environment	Spectrophotometric
Mg^{2+}	Meq/l	Titration with EDTA and EBT and Ammonia buffering	Flame AAS
Na	Meq/l	Flame photometer (BUCH Scientific)	Flame AAS

Multivariate statistical methods - cluster analysis

Cluster analysis (CA) is a group of multivariate techniques whose primary purpose is to assemble objects based on their characteristics. CA classifies objects, so that each object is 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 clusters) heterogeneity. Hierarchical agglomerative clustering is the most common approach, which provides intuitive similarity relationships between any one sample and the entire data set, and is typically illustrated by a dendrogram (McKenna, 2003). There are two types of cluster analysis: analysis based on distance (Jolliffe, 1986) and analysis based on models (Mohammadi & Prasanna, 2003).

Recently, methods based on distance are more commonly used. These methods themselves are divided into two groups: ordinal and stochastical models. Ordinal models are used more often compared to stochastical models. In this method, in the first stage of grouping, the number of parameters is equal to the number of groups and each group includes only one parameter. In later stages, the more similar groups are put together and, then, these groups themselves join other similar groups. Finally, all the parameters are put in only one group (Vega, 1998). There are different grouping methods like: the Unweighted Paired Group Method Using Arithmetic Averages (UPGMA), Ward's Minimum Variance (WMV), Single linkage (SL), and Complete Linkage (CL). Among these methods, UPGMA is the most commonly used one

(Panchen, 1992). In this method the similarities and differences between parameters and related groups are equal to the similarities or differences between parameters within the group. The distance in different groups is calculated between pairs of parameters. This is while, in Ward's method, grouping is done based on intra-group minimum and inter-group maximum variance (Otto, 1998). In this study, the UPGMA method was used.

The normality of the data distribution was analyzed by one sample Kolmogorov-Smirnov test. All the mathematical and statistical calculations were done by $SPSS_{16}$, and $MINITAB_{15}$.

Results

Different measured parameters and their mean, SD and ranges are given in the Table 2.

To classify water quality in the sampling stations and to determine the sources of pollution, CA was used along with the UPGMA method, using Euclidean distance based on the standardized mean of the 18 measured parameters. With regard to the dendogram cross-section, the stations were divided into three groups based on the greatest Euclidean distance (Laurie *et al*., 2005). Figure 2 represents a cluster analysis dendrogram based on the measured parameters.

Table 2 - Descriptive statistics of water quality variables.

Parameters	**Units**	Mean	Standard Deviation (SD)	Range	
				Minimum	Maximum
EC	μ S cm^{-1}	931	656	125	2069
DO	mgl^{-1}	8.68	0.86	5.22	10.23
Turb.	NTU	15.01	6.17	6.13	26.57
pH	pH unit	7.85	0.16	7.66	8.09
Temp.	C°	10.71	1.44	8.22	12.67
NO_3^-	mgl^{-1}	5.86	4.57	2.69	18.07
NH_3	mgl^{-1}	0.37	0.50	0.09	1.50
PO_4^{3-}	mgl^{-1}	0.91	0.74	0.46	2.82
COD	mgl^{-1}	12.41	2.85	8.30	18.07
BOD	mgl^{-1}	2.15	1.08	0.87	4.55
TColi.	MPN/100ml	475.3	229	127.5	764.1
FColi.	MPN/100ml	192.9	132.2	34.7	372.0
HCO_3	Meq/l	3.29	1.37	0.72	5.89
Cl^-	Meq/l	2.48	2.24	0.35	8.05
SO_4^{2-}	Meq/l	3.36	3.13	0.17	8.10
Ca^{2+}	Meq/l	2.72	1.40	0.66	5.05
Mg^{2+}	Meq/l	1.56	0.82	0.35	2.75
Na	Meq/l	4.91	4.32	0.30	13.83

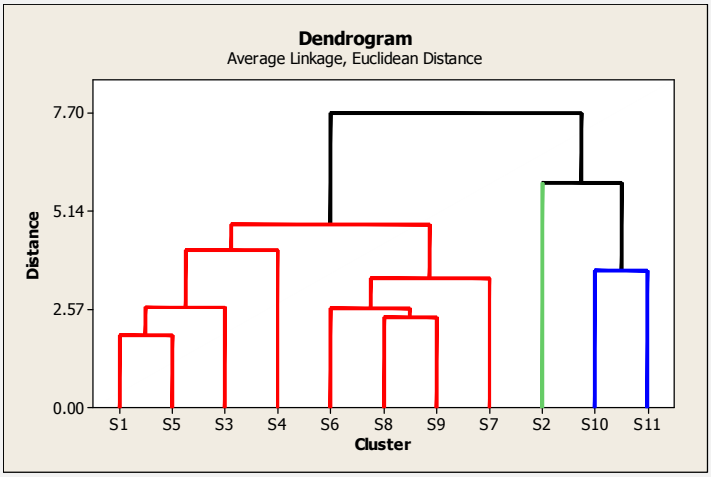

Figure 2 - Cluster analysis dendrogram of the sampling station for surface water quality assessment in Gharasou River basin (S_1, S_2, …, S_{11} are stand for station 1, station 2 , …., station 11 respectively).

The first group included stations S10, and S11 and the changes in water quality in them were mainly due to the agricultural pollutants, sewage from the Ardabil sewage treatment plant, and Ardabil slaughter house. The second group, which only included S_2, is affected by the drainage from warm mineral springs. The third group includes stations S_1, S_3, S_4, S_5, S_6, S_7, S_8, and S_9 where the water quality in these stations is mainly affected by residential sources of pollutants. Therefore the differences between the groups indicate the differences in the sources of pollution.

The dendrogram in Figure 2 shows that stations 10 and 11 have the highest pollution level (HP). These stations are distinguished from other stations concerning the level of pollution and have the most distance from other stations. After that is the second group with moderate pollution (MP), which is related, to station 2. Other stations are among the less polluted (LP) stations. The results of one-way ANOVA confirms the existence of meaningful differences among resulting clusters concerning most studied parameters with a significance level of 0.05 and 0.01. Within-group assessments showed that these parameters were not meaningfully different within groups. This is while there were meaningful differences among clusters concerning the most studied parameters.

Investigation of differences between the resulting clusters indicated that there are no significant differences between each station cluster with regard to the evaluated parameters. But we found that the resulting clusters are different ($P<0.05$, $P<0.01$) based on their evaluated characteristics (Table 3). The Bartlet test also, shows a correlation coefficient of 99% and confirms the multivariate statistical techniques used in this study.

Discussion

The results show the effect of different polluting factors in the environment on the quality of water. According to the findings of this research, these methods can be used, with high confidence level, in surface water resource quality assessment. The findings are in accordance with the findings of Boyacioglu on the Tahtali River in Turkey and Zhang *et al.* in the Daliao River in China with regard to sampling stations clustering.

Table 3 - Mean, SD and variance for each evaluation parameters resulted from cluster analysis.

Cluster	Statistical parameters	EC	DO	Turb.	pH	Temp.	No_3^-	NH_3	PO_4^{3-}	COD	BOD	TColi.	FColi.	HCO3	SO_4^{2-}	Cl^-	Ca^{2+}	Mg^{2+}	Na
1	$\bar{x}$	592	9.90	13.08	7.85	10.25	4.21	0.15	0.54	-1.26	1.64	470.2	204.5	2.78	1.66	1.48	2.01	1.19	2.72
	$\bar{x}_h - \bar{x}_{..}$	-339	1.22	-1.93	0.00	-0.46	-1.65	-0.21	-0.37	-2.65	-0.51	-5.1	11.6	-0.51	-1.7	-1.01	-0.71	0.62	-2.19
2	$\bar{x}$	2068.7	6.59	15.80	8.06	12.18	3.02	0.13	1.01	13.77	2.37	724.38	361.04	5.89	8.10	8.05	5.05	2.75	13.83
	$\bar{x}_h - \bar{x}_{..}$	1137.7	-209	0.79	0.21	1.47	-2.84	-0.24	0.1	1.36	0.22	249.08	168.14	2.61	4.74	5.57	2.33	1.19	8.92
3	$\bar{x}$	1717.3	5.87	22.33	7.72	11.82	13.85	1.37	2.32	16.77	4.08	371.1	62.2	4.03	7.78	3.73	4.44	2.47	9.18
	$\bar{x}_h - \bar{x}_{..}$	786.3	-2.81	7.32	-0.13	1.11	7.99	1	1.41	4.36	1.93	-104.2	-130.7	0.75	4.42	1.25	1.72	0.91	4.27
Mean	Total	931	8.68	15.01	7.85	10.71	5.86	0.37	0.91	12.41	2.15	475.3	192.9	3.28	3.36	2.48	2.72	1.56	4.91
F		2024900.72**	0.942**	82.14^{ns}	0.00^{ns}	4.14*	4078^{ns}	0.72*	1.94**	27.90**	3.95**	67460.87^{ns}	16493.59^{ns}	3335.08**	43.57**	22.65**	8.60**	2.57**	53.58**

Using the CA method, the 11 sampling stations were divided into three clusters with similar qualitative features. The results obtained from groupings, like the findings of Shrestha *et al.* on the Fuji River in Japan, Singh *et al.* on the Gomti River in India, and Zhang *et al.* on the Daliao River in China, showed that the number of sampling stations and associated monitoring costs can be reduced without missing much information. These findings show that selecting a single station from each cluster for rapid water quality assessment can provide acceptable information in surface water monitoring network.

Cluster analysis constitutes a valuable tool that allows the identification of tendencies of different hydrochemical processes that are difficult to characterize using univariate statistical methods. In this case study, hierarchical CA helped to group the eleven sampling sites into three clusters of similar characteristics pertaining to water quality characteristics and pollution sources. Therefore, we can conclude the following. The resulting pollution from warm mineral waters wastes, residential wastes and agricultural drainage are the main factors responsible for deteriorating the quality of water in the Gharasou River. Thus, the multivariate statistical technique is very useful in the analysis and interpretation of large and complex data sets, in water quality assessment, and in the identification of the factors that can affect water quality. Naturally, this will aid in the better understanding of the sources and factors responsible for pollution and designing monitoring networks for the effective management of water resources.

Acknowledgment: We would like to thank the two anonymous reviewers who provided useful comments on this article. The study was carried out as part of a PhD thesis in the Science and Research Branch of Islamic Azad University.

References

Abdul-Wahab, S.A., C.S. Bakheit and S.M. Al-Alawi (2005). Principal component and multiple regression analysis in modeling of ground-level ozone and factors affecting its concentrations. Environmental Modelling & Software, 20 (10): 1263-1271.

Ardabil Regional Water organization (2004). Ardabil dam irrigation and drainage network design. Consulting engineers Band Ab. PP: 135

Ardabil regional water organization (2005). Environmental impact assessment of Sabalan dam and network. Nepta consultative engineering corporation. PP: 239

Astel, A., M. Biziuk, A. Przyjazny and J. Namiesnik (2006). Chemometrics in monitoring spatial and temporal variations in drinking water quality. Water Research, 8: 1706-1716.

Boyacioglu, H.U. and H. Boyacioglu (2007). Water pollution sources assessment by multivariate statistical methods in the Tahtali Basin, Turkey. Evniron. Geol., 54: 275-282.

Brumelis, G., L. Lampina, O. Nikodemus and G. Tabors (2000). Use of an artificial model of monitoring data to aid interpretation of principal component analysis. Environmental Modelling & Software, 15 (8): 755-763.

Glasbery, C.A. (1998). Normal distribution assumptions in discrimination. *Proc. of* IGARSS'88 symposium, Ebinburgh, Scotland, 1789-1791.

Helena, B., R. Pardo, M. Vega, E. Barrado, J. M. Fernandez and L. Fernandez (2000). Temporal

evolution of groundwater composition in an alluvial aquifer (Pisuerga River, Spain) by Principal component analysis. Water Research, 34: 807-816.

Johnson, R.A. and D.W. Wichern (1992). Applied Multivariate Statistical Analysis. Prentice-Hall, Englewood Cliffs, NJ.

Jolliffe, I.T. (1986). Principal component analysis. Springer-Verlag, pp. 271.

Karamouz, M., R. Kerachian and M. Karimi (2005). Water quality monitoring network for River system: Application of genetic algorithm. Pro. Of ASCE world water and Environmental Resources Congress. Alaska, USA.

Laurie, K., Bryan, F. and manly, J. Manly (2005). Multivariate Statistical Methods: A Primer . Chapman & Hall/Crc .

Lillesand, T. M., R. W. Kiefer and J. W. Chimpman (2004). Remote sensing and image interpretation(5th ed.). USA: John Wiley and Sons Inc.

Love, D., D. Hallbauer, A. Amos and R. Hranova (2004). Factor analysis as a tool in groundwater quality management: two southern African case studies. Physics and chemistry of the Earth, 29: 1135-1143.

McKenna, J.E. (2003). An enhanced cluster analysis program with bootstrap significance testing for ecological community analysis. Environmental Modelling & Software, 18 (3): 205-220.

Mohammadi, S. A. and B. M. Prasanna (2003). Analysis of genetic diversity in crop plant salient statistical tools and considerations. Crop Sci, 43: 1235-1248.

Nouri, R., R. Kerachian, A. Khodadadi and A. Shakibaeinia (2007). Assessment of importance of water qulity monitoring stations using principal components analysis and factor analysis: a case study of the Karoon River. Water and Wastewater,63: 60-69.

Otto, M., (1998). Multivariate methods. In: Kellner, R., Mermet, J. M., Otto, M., Widmer, H.M.(Eds), Analytical Chemistry. Wiley-VCH, Weinheim, Germany .

Panchen, A.L. (1992). Classification, evolution and the nature of biology. Cambridge Univ. press, Cambridge, England. P:127-129

Sarbu, C. and H.F. Pop (2005). Principal component analysis versus fuzzy principal component analysis. Acase study: the quality of Danube water (1985-1996). Talanta, 65: 1215-1220.

shrestha, S. and F. Kazama (2007). Assessment of surface water quality using multivariate statistical techniques: A Case study of the Fuji River Basin, Japan. Environmental Modeling and Software, 22: 464-475.

Singh, K.P., A. Malik, D. Mohan and S. Sinha (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.

Singh, K.P., A. Malik, D.Mohan and S. Sinha (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.

Standard Methods for the Examination of Water and Wastewater (2005). 21th edn, American Public Health Association/American Water Works Association/Water Environment Federation, Washington DC, USA.

Ouyang, Y. (2005). Application of principal component and factor analysis to evaluate surface water quality monitoring network. Water Research, 39: 2621-2635.

Vega, M., R. Pardo, E. Barrado and L. Deban (1998). Assessment of seasonal and polluting effects on the quality of River water by exploratory data analysis. Water Research, 32: 3581-3592.

Ward, J.H. (1963). Hieratical grouping to optimize an objective function. J. Am. Statist. Assoc, 58: 236-244.

Wunderlin, D.A., M. Diaz, M.M.V. AME, S. F. pesce, A. C. Hued and M. Bistoni (2001). Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality, A case study: Suquia River basin (Cordoba-Argentina). Water Research, 35: 2881-2894.

Zhang, Q., X. Shi, B. Huang, D.S.Yu, I. Oborn, A. Blomback, H. J. Wang, T. F. Pagella and F. Sinclair (2007) Surface water quality of factory-based and vegetable-based Peri-urban areas in the Uangtze River delta region, China, Catena 69: 57-64.

Zhang Y., F. Guo and W. Meng (2008). Water quality assessment and source identification of Daliao River basin using multivariate statistical methods. Environmental Monitoring and Assessment. DOI 10.1007/S10661-008-0300-z.