Journal of Biodiversity and Environmental Sciences (JBES)
ISSN: 2220-6663 (Print) 2222-3045 (Online)
Vol. 3, No. 8, p. 156-164, 2013
http://www.innspub.net

RESEARCH PAPER **OPEN ACCESS**

Qualitative classification and determining pollution sources of Iran's Talkherud river using multivariate statistical methods

Ebrahim Fataei[1*], Hamed Hassan Pour Kourandeh[2]

[1]*Department of Environmental Engineering, Ardabil Branch, Islamic Azad University, Ardabil, Iran*

[2]*Young Researchers Club, Tonekabon Branch, Islamic Azad University, Tonekabon, Iran*

Article published on August 24, 2013

Key words: Cluster Analysis, water quality, pollution sources, Talkherud River.

Abstract

This researchwas carried out for qualitative evaluation of Talkherud which is located in northwest Iran in East Azerbaijan Province. The sampling has been conducted on 13 physical and chemical parameters at eight stations within one year. The results obtained from the measurements were analyzed using multivariate cluster analysis (CA). With respect to the obtained results of cluster analysis, the studied stations were divided into three groups: (a) High Pollution(HP), (b) Middle Pollution(MP), and (c) Low Pollution (LP). The stations which were alike in amount of pollution fell into one group. This showed the difference in pollution sources and the amount of pollution in different regions of the river. The study of the differences between the mentioned groups shows, that in terms of assessed parameters, there is no significant difference between the stations in each cluster. While based on most assessment parameters, there was a significant difference among the clusters at the 5% and 1% probability levels. The total results of this study show the usefulness of multivariate statistical techniques in the interpretation of multiple datasets, qualitative evaluation of water, and the identification of pollution sources and causes.

***Corresponding Author:** Ebrahim Fataei ✉ ebfataei@gmail.com

Introduction

Themultivariate statisticalmethod is oneof the methods in the examination of water and environmental analysis and control of thequalitative variables of rivers and it has been widelyused in recent years(Zhang *et al*, 2000 ;Fataei *et al*, 2010).Multivariate statistical methodsaremethodsthat simultaneously investigate the changes in severalvariables and the statistical inferences derived from them. These methods are much more reliable than univariate methods, because, in them, random errors of a variable canbe compensated by othervariables to some extent. The study of each of the elements cannot provide us with as much informationabout their relationship as the study of all of the variables can. Multivariate statistical studies onthese parametersallow the various elements and their relationships to be examined. Some studies have been conducted on the control of qualitative variables andthe control of qualitative water monitoring stationsusing clusteranalysis. These items can include the determination ofsurface water quality in a region of northern Greece and Turkey, and the evaluation ofspatial and temporalvariations inthe water quality of India, China and Iran (Boyacioglu *et al*, 2007;Fataei *et al*, 2010 ;Zhang *et al*, 2009).In a study onthe Mahanadi River of India, cluster analysis was used to assess spatial and temporal variables and according to various qualitative parameters, thesampling locations were divided into groups with the highest similarity (Sandaray and Panda, 2006). In a study on Xiangjiang River using cluster analysis to evaluate water quality, 34 sampling stations were divided into three categories based on similarity and water quality characteristics (Zhang *et al*, 2009). Given that Talkherud River has been one of the most important rivers of Orumieh Lake Basin, and absorbs a variety of urban, industrial and agricultural pollutants on its way, this research was conducted to assess water quality of Talkherud and determine pollution sources and the important parameters affecting their quality through multivariate cluster analysis.

Material and methods

Study area

Talkherud River is one of the most important rivers ofthe East Azerbaijan Province which collectsrelatively broad waters from this province (cities of Tabriz, and Bostan Abad)and carries them to the Orumieh Lake and runs in the general east-westdirection. Tabriz, Sarab, Bostan Abad, Heris and Osku are considered as major urban areas of this area.Talkherud River springs from the southern slopes of Sabalan and it is a long and important river that drains an extensive area to the Orumieh Lake.

The maximum altitude of Talkherud Basin is 3882m in the northeastand its lowest altitude is 1280m on the Orumieh Lake shores. The precipitation in the basin varies from approximately 600mm in thenorth-easternhighlands to about 250mmneartheVaniyar dam. The average precipitation in the Orumieh Lake Basin is 312mm.The length of the river to the Orumieh Lake Delta is about 276km. This river is located in the geographical coordinates of 47° 54' to 45° 40' east longitude and 38° 30' to 37° 34' north latitude. Talkherud River, with an area of 11490 km², is composed of various branchesincluding: Aqmion Chay, Tajyar, Razliq Chay,Vanq Chay, Mehraban,Nahand Chay, Saeid Chay, Mehranroud, Sardroud& Onsorroud (Ghods Niroo Consulting Engineers, 2006).

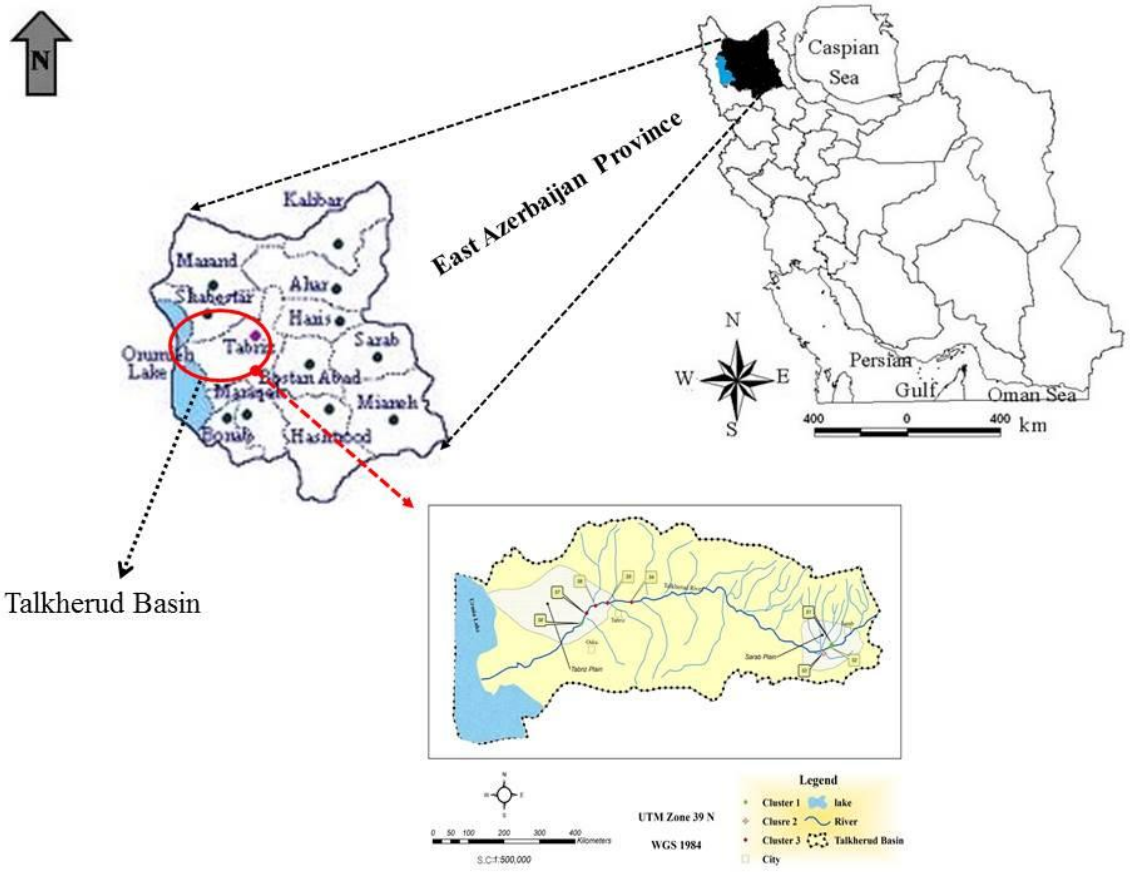

Fig.1 The location of sampling stationsin the studied area

The method of sampling, measurement and analysis of parameters

In selectingsamplingstations factors such aspollution sources, basin land use, establishment of residential centers and the existing industries, geological structures, major and secondary branches of the river and ease of access were considered. After the specification of the sampling stations, ituation and coordinates of these stations were determined on the map through global positioning system(GPS). (table 1).The location of stations is shown in figure 1.

Table 1. Geographical location of sampling stations

station	Geographical condition	
	North Latitude ° ´	East longitude ° ´
S1	37-52	47-29
S2	37-58	47-40
S3	38-02	47-30
S4	38-11	46-28
S5	38-09	46-17
S6	38-04	46-22
S7	38-11	46-09
S8	38-01	46-02

The number of sampling was determined 8 stations. Sampling was conducted over a one year period from September 2008 to September2009 on a monthly basis. The methods and apparatus used to were analyze samples for are summarized in table 2

Table 2. Parameters of water quality, units and the analysis methods used.

Parameters	Abbreviations	Units	Analytical methods
DO	Dissolved oxygen	Mgl^{-1}	Winkler azide method
Turb.	Turbidity	NTU	Turb. -meter
PH	PH	PH unit	PH-meter
$T.PO_3$	Nitrate Total	Mgl^{-1}	Spectrophotometric
Temp.	Temperature	C°	Mercury thermometer
$T.PO_4$	Phosphate Total	Mgl^{-1}	Spectrophotometric
COD	Chemical Oxygen Demand	Mgl^{-1}	Dichromate reflex method
BOD	Biochemical Oxygen Demand	Mgl^{-1}	Winkler azide method
TColi.	Total coliform	MPN/100ml	Multiple tube method
FColi.	Fecal coliform	MPN/100ml	Multiple tube method
T.S.S	Total suspended solids	Mg/l	Gravimetric
T.D.S	Total dissolved solids	Mgl^{-1}	Gravimetric
T.A	Total Alkalinity	Mg/l	Titration

Table 3. Statistical Profile of water quality variables.

Variable	Mean	Standard Deviation	Range of Variations	
			Min.	Max.
Dissolved oxygen	8.28	1.500	6.50	10.25
Turbidity	26.75	9.238	16	38
PH	8.11	0.165	7.8	8.3
Temperature	8.50	5.001	1.07	15.2
Nitrate Total	2.51	0.634	1.72	3.59
Phosphate Total	1.64	0.755	1.06	2.66
Chemical Oxygen Demand	116.38	52.180	24	170
Biochemical Oxygen Demand	17.36	11.940	2.66	33.35
Total coliform	821.25	259.447	585	1130
Fecal coliform	579.69	17.950	560	620
Total dissolved solid	2492.50	2560.89	350	6250
Total suspended solid	166.38	71.669	60	251
Total Alkalinity	127	33.316	56	161

To analyze the data for qualitative classification of the studied stations and determining pollution sources, the researcher used statistical cluster analysis methods.All statistical and mathematical calculations were performed through the software EXCEL2007, MINITAB16 and SPSS17.

Multivariate statistical analysis method (clusteranalysis)

Cluster analysis is one of the multivariate analysis methods that mainly aim to classify stations based on the intended features.The goal of cluster analysis is to search for the natural categories of items or variables. Thereby, the stations withsimilar characteristicsfall into the same categoryand the classification will be accurate when the intragroup stations will be more homogeneous and the intergroup stations are more heterogeneous (Mckenna 2003).There aretwo cluster analysismethods, distance-based methods (Jolliffe, 1986)and model-based methods (Mohammadi *et al*,2003).Currently, distance-based methods are the most widely used methods.

The main goal of cluster analysis is to search for natural categories of variables.For this purpose, we first need to find a qualitative standard by whichto measurethe similarity between objects.There are so many ways to measure the similarity between pairs of objects,including the use of two standards of Euclidean distance and city block distance (Niroomand,2000). To determine Euclidean distance and city block distancebetween the two observations P

(1)

$$\underline{X}_j = [x_{j1}, x_{j2}, \ldots, x_{jp}] \quad , \quad \underline{X}_i = [x_{i1}, x_{i2}, \ldots, x_{ip}]$$

Formulas 2 and 3 are used respectively, as following:

(2) Euclidean distance

$$d_{(i,j)} = \sqrt{\sum_{f=1}^{p} (x_{if} - x_{if})^2}$$

(3) City block distance

$$d_{(i,j)} = \sum_{f=1}^{p} |x_{if} - x_{if}|$$

Inclusteranalysis, inaddition to the issue of determination of distance (similarities), another important issue must be considered that is the selection of clustering methodbased on distance. Cluster analysis methods are divided into two general categories:sequential model and non-sequential model.Sequential model is more used.

In this method, in the first stage of classification, the numberof parameters is equal to the number of groups and each group contains one parameter. In the next stages, the most similar parameters are classified into groups and these groups will form with other groups based on their similarities to one another, and in the end, all the parameters fell into one group (Vega *et al*, 1998). There are different sequential methodsfor classification including Unweighted Paired Group Method using Arithmetic Averages(UPGMA), Ward's Minimum Variance Method(WMV), nearest neighbor and furthest neighbor.

In UPGMAmethod, the similarity or difference between parameters and the respective group is the average similarity or difference of that parameter with other parameters in the group and the distance of people indifferentgroups is measured two by two, while in the Ward method,Wardclassification is conductedclassification based on the minimum intragroup variance and the maximum intergroup variance.Also, in this method, categories are integrated based on the minimum increases in sum of squares in each category(Otto *et al,* 1998).In this presentstudy,Ward method was usedfor data clustering.

Overall algorithm of cluster analysis is as follows:
In the firstNstage,there are Nelements for classification.

The nearest two elements in a group will be found and joined together to form a new element. The distance between this new element and the remaining elements will be calculated and classification will be performed at the very condition of the first stage, but with N-1 elements. Again, the two nearest searched elements will be joined together and the new distances will be calculated. This process continues until only one single element is found (Karimian, 2006).

Cluster analysis is a general title for a series of mathematical methods, which is used for finding the similarities between the objects in a set.The purpose of formation of clusters or categories is to put objects with less variance or variation, compared to that between categories, in each category.In Cluster analysis, usually P parameters on N stations are measured and an N*P matrix is formed.

Results and discussion

To classify water quality of the sampling stations and determine pollution sources, the cluster analysis was conducted through Ward method using Euclidean distance based on the standardized mean of 13 measured parameters. According to the furthest Euclidean distance Dendrogram cut divided stations into three groups(Laurie et al, 2005). Figure 2 shows Dendrogramof clusteranalysisof stationsbased on the measured parameter.

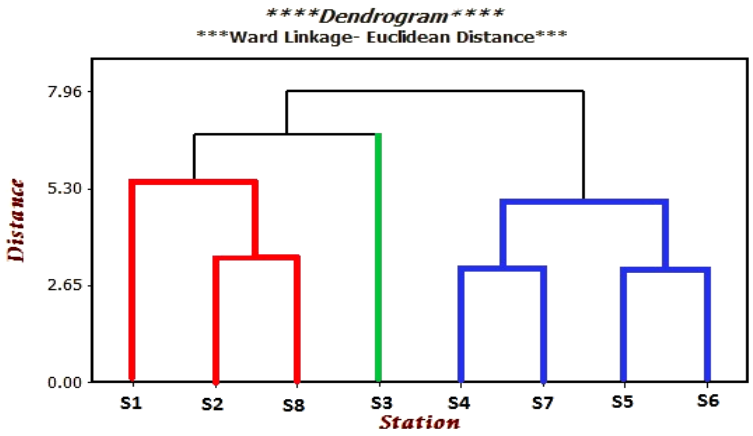

Fig. 2 Dendrogram of cluster analysis of sampling sites for quality characteristics of the surface water of the studied basin.

The first group included the stations 1, 2, 8, in which qualitativechangesofwater were mainlyaffected by pollutants from residential centers, sewageof food companies.The second group included three stations where water quality wasaffected by pollution sourcesof agricultural activitiesandsewage of manufacturing companies and commercial andresidential complexes.The third group included the sites 4,5,6,7 in the region with activitiesoftheindustrial town establishment(where most industries include paint, tannery units and oil refinery), where includes municipal wastewater and industrial slaughterhouses. Thus, their classification is affected by human and municipal sewage and wastewater treatment plant effluent and industrial

slaughterhouses which have the greatest difference with allthe stations. Thus, the difference between groups indicates the difference in pollution sources. The differences between the groups suggest that the stations in each clusterhad no significant difference in terms of the assessed parameters.However, table 4 shows thatthere is a significant difference between the clusters in terms of most assessed properties at 1% and 5% probability levels.

It suggests that selecting a station from each group in the monitoring network provides appropriate information for rapid assessment of water quality (Shingh, 2004). Among the first group stations, station 1 has the highest quality and is located is in a sub-cluster. This is due to nearness of the station to thesource and getting further from the source.It is shownthat stations 4, 5, 6, 7 are High Pollution stations(HP) which are distinguished from other stations by the amount of pollution and are at the greatest distance from the others.The second category into which only station 3 falls shows Middle Pollution(MP)and the rest of the stations are in Low Pollution (LP) group.As the results in table 4 shows the parameterBOD in the first cluster has not changed so much, but the changes in the second cluster are tangible and in the third cluster BOD has decreased. The changes in the second cluster are caused by entrance of the wastewaterscontaining organic materials such as urban wastewater and milk factory. MehranRoud, one of the minor branches ofTalkherud, absorbs Tabriz urban and household sewage and wastewater when passing throughTabriz and after connecting to Talkherud; it conducts the effluents to Talkherud and thereby, reduces the river's water quality within the scope of this station.Also, the reduction ofDO in clusters 2 and 3 results from the passage of this basinalong the city of Tabriz, and its location aturban areas, and entrance ofwastewaters of industrial complexes located in the West Tabriz, including partially purified wastewaters of slaughterhouse,pasteurized milk factory as well as refineries, petrochemicals, power plants, TractorManufacturing Companies and Machine Manufacturing Companies, etc and highly contaminated wastewaters ofmore than 300 units of tanning in Charmshahr cityThe increased value of TDS in thethird cluster is caused by river crossing the salty and gypsum lands and dissolving them in the river water and entering the soil minerals and particles by drainers.While due to the absence of such pollution sources in the stations in the nextcluster, there is no significant increase in the total amount of their dissolved solids.Parameter TSS is another parameter of water pollution index the significant changes of which are seen in the third cluster stations. The main reason for the increase is location of these stations at the scope of the irrigation and drainage network of Tabriz plain and entrance of soil minerals and particles into the river. The study of changes in two parameters of totalcoliform and fecal coliform in thestations of triple clusters shows thatthese changes in all threeclusters are caused by the entrance of urban sewage to the river, and they will significantly decrease when the river passes through gypsum and salty lands which leads to the natural disinfection.

Increase or decrease of alkalinity in different stations located at the clusters is caused by the calcareous nature of longitudinal part of the river bed in some parts andalso entrance of wastewatersthat somehow increase alkalinity. However, low alkalinity in the first cluster could be a good indicator to show the lowness of parameter in the upstream and source.

The important point about station 8 is that it is the last sampling station and it should basically fall into high pollution cluster, while conversely, it falls into the cluster of low pollution stations,which indicates the river's high assimilationpower on theway of station 8.So that due to the low temperature of this station andhigh river flow at that station, oxygen uptake rises and thereby, increase of the oxygen dissolved in waterincreases the activityof microorganisms is water. Also, high river flow at this stationleads to fast carriage of pollutants,appropriate mixture conditions and short water residence timeand therebyreduces the ecological destructive

effects.On the other hand, high vegetation in this station and photosynthesis, also turbulent flows, leads to increasetheentrance ofoxygen to the water and the river's assimilationpower.

Table 4. Mean, mean absolute deviation and standard deviation in the three clusters obtained from cluster analysis for assessed parameters.

Cluster	Statistical parameters	Temp.	pH	Tpo_4	Tpo_3	BOD	DO	TDS	Turb.	TSS	COD	Tcoli.	Fcoli.	AL.
1	χ	8.66	8.07	1.60	2.87	19.21	9	462.33	18.66	103.32	58	643.33	575.83	98.33
	$\bar{\chi}_k$ – ??	0.16	0.03	-0.02	0.36	1.85	0.73	-2030.1	-8.08	-63.500	-58.38	-177.92	-3.85	-28.6
2	$\bar{\chi}$	-7.34	7.80	2.66	1.99	28.30	6.50	1400	22	120	170	1130	620	151
	$\bar{\chi}_k$ – ??	1.42	-0.31	1.02	-0.52	10.94	-1.76	-1092.5	7	-46.380	53.62	308.750	40.31	24
3	$\bar{\chi}$	10.24	8.21	1.40	2.36	13.23	8.15	4288.20	34	225.250	146.75	877.5	572.50	142.5
	$\bar{\chi}_k$ – ??	1.73	0.10	-0.23	-0.14	-4.12	-0.11	1795.70	7.25	58.87	30.37	56.250	-7.19	15.50
Mean	Total	8.50	8.11	1.36	2.51	17.35	8.26	2492.50	26.75	166.38	116.38	821.250	579.69	127

Conclusion

The use of cluster analysis in the qualitative analysis of Talkherud water highly contributes to the determination of the river water pollution in the studied stations.According to the survey results, in the areas where stations are very similar, similar stations can be eliminated and in the areas where the stations are highly different from each other, new stations should be established.The analysis of results shows that the amount of pollution is generally increased from upstream to downstream of Talkherud.The study results are consistent with the results of the study Zhang *et al.* conducted in 2009, the method of clustering stations in both studies isWard method, with the difference that, the number of sampling parameters in both studies is different. In comparison withthe results of this study, the results of the study of Hashemi *et al.* in 2005 clearly reveals the advantages of cluster analysisIn this paper, the stations are divided into three categories of high pollution; middle pollution and low pollution which indicate the exact condition of pollution in the studied stations as well as location of similar stations in a cluster in terms of amount and value of pollution.

According to the results of the physical-chemical parameters and the results ofcluster analysis, industrial and agricultural activities and the location of Talkherud Basin in the industrial and urban area are the main pollution sources of river. The existence of population centers, industries and industrial and agricultural centers marge the river and location of the region, water quality in this region is very important. Therefore, it is recommended that creating treatment plant for municipal and industrial sewage, their direct discharge to the river will be

prevented.The results obtained from this study in the classification of the studied stations on the river using cluster analysis can be used for other water sourcesThe overall results of this study indicate the usefulness of the multivariate statistical techniques in interpreting multiple datasets, qualitative evaluation of water and identification of pollution sources and causes.

References

Karimian A. 2006. Statistical analysis and determination of qualitative zones of Zohre River water using cluster analysis of similarities.MS thesis.Water and Wastewater. Khuzestan Science and Research University.

Janon Richard A, Wichern Dean W, Translated By, Niroomand A. 2000. applied multivariate statistical analysis. Mashhad, Mashhad Ferdowsi University Press.

Ghods Niroo Consulting Engineers. 2006.The project of the quality of water sources of Talkheroud Basin.Vol. III. Final hydrology report, East Azerbaijan Regional Water Association.

Zhang Qi, shiX, HuangB, Yu DS, Oborn I, BlombackA, Wang HJ, PagellaTF, Sinclair F.2007. Surface water quality of factory-based and vegetable-based Peri-urban areas in the Uangtze River delta region China, Catena69, 57-64. http://dx.doi.org/10.1016/j.catena.2006.04.012

Shingh KP, Malik A, MohanD, SinhaS. 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. http://dx.doi.org/10.1016/j.watres.2004.06.011

Boyacioglu HU, Boyacioglu H. 2007. Water pollution sources assessmentg by multivariate statistical methods in the Tahtali Basin, Turkey. Environmental Geology54, 275-282.

Zhang QI, LiZhongwu, ZengGuangmingLi, JianbingFang Y, Qingshui Y. 2009.Assessment of surface water quality using multivariatestatistical techniques in red soil hilly region, a case studyof Xiangjiang watershed, China. Environmental Monitoring and Assessment152,123–131. http://dx.doi.org/10.1007/s10661-008-0301-y

Fataei E, Monavari M, Hasani AH, Mirbagheri SA, Karbasi AH. 2010. Heavy metal and agricultural toxic monitoring in Garasou River in Iran for water quality assessment using multivariate statistical methods. Asian Journal of Chemistry 4,2991-3000.

Mckenna JE. 2003. An enhanced cluster analysis program with bootstrap significance testing for ecological community analysis. Environmental Modelling& Software 18(3),205-220. http://dx.doi.org/10.1016/S1364-8152(02)00094-4

Jolliffe IT. 1986. Principal Component Analysis.Springer-Verlag, p. 271.

Mohammadi SA, Prasanna BM.2003. Analysis of genetic diversity in crop plant salient statistical tools and considerations.Crop Science 43, 1235-1248. http://dx.doi.org/10.2135/cropsci2003.1235

Vega M, Pardo R, Barrado E, DebanL.1998. Water Research 32: 3581.

Otto M, Kellner R, Mermet JM, Widmer HM. 1998. Multivariate methods (Eds.). Analytical Chemistry. Wiley-VCH.Weinheim.

Laurie K, Manly BFJ.2005.Multivariate Statistical Methods. A Primer. Chapman. Hall.crc, 241 p.

Sundaray SK, Panda UC.2006. Multivariate statistical teqniques for the evaluation of spatial and temporal variations in water quality of the Mahanadi riverestuarine system, India. Journal Environ Geochem Health. waste water.19th edituion. USA.