World Journal of Fish and Marine Sciences 3 (5): 366-370, 2011
ISSN 2078-4589

Evaluation of Surface Water Quality Using Cluster Analysis: A Case Study

[1]E. Fataei and [2]S. Shiralipoor

[1]Department of Environmental Engineering, Ardabil Branch,
Islamic Azad University, Ardabil branch, Ardabil, Iran
[2]Young Researchers Club, Islamic Azad University-Ardabil Branch, Ardabil, Iran

Abstract: In This study, multivariate statistical method including cluster analysis was applied to surface water quality data sets obtained from the Herou river, Iran. Cluster analysis (CA) generated two clusters. Based on CA analysis the stations were divided to twp groups of highly polluted (HP) and less polluted (LP) stations. Based on the locations of the sites consisted by each cluster and variable concentrations at these stations, it was concluded that residential wastes and agricultural drainage are the main factors responsible for deteriorating the quality of water in the Herou river. Extracted clustered information can be used in reducing the number of sampling sites on the river without missing much information. This method is believed to assist water managers to understand complex nature of water quality issues and determine priorities to improve water quality.

Key words: Pollutant sources · Multivariate statistical · Surface water resources · Herou River · Iran

INTRODUCTION

The application of different multivariate statistical techniques including cluster analysis (CA), Discriminant analysis (DA), principal component analysis (PCA) and factor analysis (FA), source apportionment by multiple linear regression on absolute principal component sores (APCS-MLR) for interpretation of the complex databases offers a better understanding of water quality in the study region. These techniques also permit identification of the possible factors/sources that are responsible for the variations in water quality and influence the water system and in apportionment of the sources, which, thus offers valuable tool for developing appropriate strategies for effective management of the water resources [1-7].

The multivariate statistical techniques such as cluster CA, FA, PCA and DA have widely been used as unbiased methods in analysis of water -quality data for drawing meaningful information [1, 4, 8-11]. Compared with these methods, traditional multivariable statistical methods such as CA, DA and PCA have become widely accepted in water quality assessment and sources apportionment of river water in the last decade [1-2, 4, 12-15].

Cluster analysis helps in grouping objects (cases) into classes (clusters) on the basis of similarities within a class and dissimilarities between different classes. The class characteristics are not known in advance but may be determined from the analysis. The results of CA help in interpreting the data and indicate patterns [1]. CA divides a large number of objects into smaller number of homogenous groups on the basis of their correlation structure. Within group similarity is maximized and amoug-group similarity is minimized according to some objective criteria [16, 17]. Important characteristics of cluster analysis are: 1) organizes observations into discrete classes or groups such that within group similarity is maximized and among-group similarity is minimized according to some objective criteria. 2) assesses relationships within a single set of variables; no attempt is made to define the relationship between a set of independent variables and one or more dependent variables, etc. [17]. Typical cluseringactivity involves following stes(18-20): definition of observation similarity, Clustering or grouping and Data abstracition.

Threfore, this study the efficiency of cluster analysis of multivariate statistical techniques was applied to evaluate in water quality data matrix of the Herou river (Iran) without loosing important information.

MATERIALS AND METHODS

Study Area: The Herou river is one of the branches of the Ghezelozan river in the southwest side of Caspian sea, in Ardabil Province, northwest of Iran. Herou basin had been developed in the cities of Khalkhal and Kivi. Herou river is one of the most important water resources in this

Corresponding Author: E. Fataei, Department of Environmental Engineering, Ardabil Branch,
Islamic Azad University, Ardabil Branch, Ardabil, Iran.

Table 1: Sampling stations in Herou river basin of Khalkhal, Iran

Stations number	Station name	X (UTM)	Y (UTM)	Elevation
1	Khojin	285269	4162395	1868
2	Khalkhal entrance	280380	4171450	1766
3	Bridge of Kivi Mineral water	271404	4175258	1530
4	Kivi existence	265660	4173092	1489

Table 2: Water quality parameters, units and methods of analysis

Parameters	Units	Abbreviations	Analytical methods
Q	m^3/s	Discharge	Hydrometric
WT	°C	Water Temperature	Mercury thermometer
pH	pH unit	pH	pH-meter
Turb.	NTU	Turbidity	Turb. -meter
DO	mgl^{-1}	Dissolved oxygen	Winkler azide method
Ec	$\mu S\ cm^{-1}$	Electrical conductivity	Electrometric
TDS	mgl^{-1}	Total dissolved solids	Gravimetric
HCO_3	mgl^{-1}	bicarbonate	Titrimetric
SO_4^{2-}	mgl^{-1}	Sulphate	Spectrophotometric
Cl^-	mgl^{-1}	Chloride	Spectrophotometric
Ca^{2+}	mgl^{-1}	Calcium	Flame AAS
Mg^{2+}	mgl^{-1}	Magnesium	Flame AAS
Na	mgl^{-1}	Sodium	Flame photometer
TH	$CaCo3mgl^{-1}$	Total hardness	Titrimetric
NH_3	mgl^{-1}	Ammonical nitrogen	Spectrophotometric
NO_3^-	mgl^{-1}	Nitrate nitrogen	Spectrophotometric
PO_4^{3-}	mgl^{-1}	Phospate	Spectrophotometric
BOD	mgl^{-1}	Biochemical oxygen demand	Winkler azide method
COD	mgl^{-1}	Chemical oxygen demand	Dichromate reflex method
TColi.	MPN/100ml	Total coliform	Multiple tube method
FColi.	MPN/100ml	Feacal coliform	Multiple tube method

region that play a important role. Herou river originated from south altitudes of the Khalkhal in the region of Majareh and after crossing the Aznow valley and the Khalkhal city and make fertile lands flows into north. This river after passing through the Sangavar area finally discharged in the Ghezelozan river.

Since the river passes through two urban (Khalkhal and Kivi) and some rural areas as well as, vast farmlands, it is quite naturally exposed to various sources of the pollution. Various anthropogenic factors threat the overall river quality such as excess water consumption, sewage discharges from different urban and agricultural settlement.

Methods of Sampling and Analyzing Parameters: Four sampling station were chosen. The coordination of each sampling station was by means of GPS (Table 1).

The sampling process was carried out during 2008-2009 in Herou River basin in northwestern Iran, are analyzed with CA after pretreatment. Sampling and analysis was done upon 24 physical and chemical and microbiological parameters. These parameters were sampled monthly (Table 2).

Multivariate Statistical Methods- Cluster Analysis: Cluster analysis is a group of multivariate techniques whose primary purpose is to assemble objects based on the characteristics them possess. Cluster analysis 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 dedrogram (tree diagram) [21].

The dendrogram provides a visual summary of the clustering processes, presenting a picture of the groups and their proximity, with a dramatic reduction in dimensionality of the original data. The Euclidean distance usually gives the similarity between two samples and a distance can be represented by the difference between analytical values from the samples [22].

In this study, hierarchical agglomerative CA was performed on the normalized data set by means of the Ward's method, using squared Euclidean distances as a measure of similarity. The Ward's method uses an analysis of variance approach to evaluate the distances between clusters in an attempt to minimize the sum of squares (SS) of any two clusters that can be formed at each step. The spatial variability of water quality in the whole river basin was determined from CA, using the linkage distance, reported as D_{link}/D_{max}, which represents the quotient between the linkage distances for a particular case divided by the maximal linkage distance. The quotient is then multiplied by 100 as a way to standardize the linkage distance represented on the *y-axis* [2, 4, 6].

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 $MINITAB_{15}$.

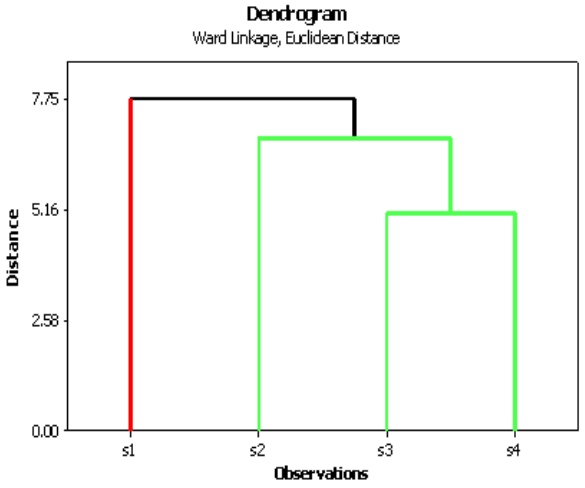

Fig. 1: Cluster analysis dendrogram of the sampling station for surface water quality assessment in Herou river basin (S_1, S_2, …, S_4 are stand for station 1, station 2, …., station 4 respectively)

RESULTS AND DISCUSSION

To classify the water quality in sampling stations and to determine the sources of the pollution, CA with Ward method, using Euclidean distance based on the standardized mean of the 18 measured parameters, were used. With regard to dendogram cross-section, the stations were divided into Two groups based on the farthest Euclidean distance [23]. Figure 1 represents cluster analysis dendrogram based on the measured parameters.

The first group included station S_1. The second group includes stations S_2, S_3 and S_4 where the water quality in these stations is mainly affected by residential pollutant sources, Kivi warm mineral waters wastes,, agricultural pollutants, sewage from Khalkhal sewage treatment plant, several slaughter houses and Khalkhal Nepan factory. Therefore the differences between the groups indicate the differences in the sources of pollution.

Table 3: Mean and variance for each evaluation parameters resulted from cluster analysis

Cluster	1		2			
Statistical parameters	O	$O_h - O_{..}$	O	O_h $O_{..}$	Mean	F
Q	0.31	-0.48	1.27	0.48	0.79	**
WT	8.96	-0.78	10.51	0.78	9.74	**
pH	8.37	0.09	8.19	-0.09	8.28	**
Turb.	14.67	-24.61	63.90	24.61	39.29	**
DO	8.03	0.03	8.09	0.03	8.06	**
Ec	464.95	-95.93	656.80	95.93	560.87	**
TDS	343.16	-20.22	383.60	20.22	363.38	**
HCO_3	206.15	-23.23	252.60	23.23	229.37	**
SO_4^{2-}	30.77	-22.84	76.45	22.84	53.61	**
Cl^-	32.00	-103.00	238.00	103.00	135.00	**
Ca^{2+}	43.55	-0.80	45.15	0.80	44.35	**
Mg^{2+}	10.01	-2.70	15.40	2.65	12.75	**
Na	42.84	-18.29	79.41	18.29	61.12	**
TH	150.05	-13.13	176.30	13.13	163.17	**
NH_3	0.16	-0.05	0.25	0.05	0.20	**
NO_3^-	5.89	-5.00	15.90	5.00	10.90	**
PO_4^{3-}	0.18	-0.10	0.39	0.01	0.38	**
BOD	1.64	-0.29	2.22	0.29	1.93	**
COD	7.04	-1.72	10.47	1.72	8.75	**
TColi.	206.24	45.67	114.90	-45.67	160.57	**
FColi.	34.92	1.46	32.00	-1.46	33.46	**

** Significant ($P < 0.01$)

Among these stations the best quality of water belongs to station 1. The reason for this is that this station is located headwater. As the stations' distance increases from the source of river, with the entrance of pollutants from different sources the quality of water decreases.

The dendrogram of Figure 1 show that stations S_2, S_3 and S_4 have the highest pollution level (HP). The station S_1 is 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 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 most studied parameters.

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

The results show the effect of different pollutant 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 the surface water resources quality assessment. The findings are in accordance with the findings of Zeng and Rasmussen in the lake Lanier, Georgia in USA and Reghunath et al., in the Karnataka river in India with regard to sampling stations clustering.

Using CA method, 4 sampling stations were divided into two clusters with similar qualitative features. The results obtained from groupings, like the findings of Shrestha, et al. in the Fuji river in Japan, Singh *et al*., in the Gomti river in India, Reghunath et al., in the Karnataka river in India and Fataei et al. in the Garasou river in Iran showed that the number of sampling stations and associated monitoring costs can be reduced without missing much information.

CONCLUSION

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 two clusters of similar characteristics pertaining to water quality characteristics and pollution sources. Therefore, the resulted pollution from residential wastes and agricultural drainage are the main factors responsible for deteriorating the quality of water in the Herou River. Also, Multivariate statistical technique is very useful in identification of the factors that can affect water quality. Naturally, this will help for better understanding of sources.

ACKNOWLEDGMENT

We would like to thank L. Tabrizi and F. Gaffari for technical cooperation.

REFERENCES

1. 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 Res., 32: 3581-3592.
2. 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 Res., 35: 2881-2894.
3. Liu, C.W., K.H. Lin and Y.M. Kuo, 2003. Sci. Total Environ. Application of factor analysis in the assessment of groundwater quality in a blackfoot disease area in Taiwan.The Science of The Total Environment, 313: 77-89.
4. Simeonov, V., J.A. Stratis, C. Samara, G. Zachariadis, D. Voutasa, A. Anthemidis, M. Sofoniou and T.H. Kouimtzis, 2003. Assessment of the surface water quality to northern Greece. Water Res., 37: 4119-4124.
5. Bengraine, K., T.F. Marhaba, J. Hazard and B. Mater, 2003. Using principal component analysis to monitor spatial and temporal changes in water quality. J. Hazardous Materials, 100: 179-195.
6. 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.
7. Fataei, E., M. Monavari, A.H. Hasani, S.A. Mirbagheri and A.R. Karbasi, 2010. Heavy metal and agricultural toxics monitoring in Garasou River in Iran for water quality assessment, Asian J. Chemistry, 4: 2991-3000.

8. Bengrain, K. and F.T. Marhaba, 2003. Using principal component analysis to monitor spatial and temporal changes in water quality. J. Hazard. Mater. B., 100: 179-195.
9. Voncina, D.B., D. Dobcnik, M. Novic and J. Zupan, 2002. Chemometric Characterization of the quality of river water. Anal. Chim. Acta, 462: 87-100.
10. Liu, C.W., K.H. Lin and Y.M. Kuo, 2003. Application of factor analysis in the assessment of groundwater quality in a blackfoot disease area in Taiwan. Sci. Tot. Environ., 313: 77-89.
11. Reghunath, R., T.R.S. Murthy and B.R. Raghavan, 2002. The utility of multivariate statistical techniques in hydrogeochemical studies: an example from Karnataka, India. Water Res., 36: 2437-2442.
12. Grande, J.A., J. Borrego, J.A. Morales and M.L. De La Torre, 2003. A description of show metal pollution occurs in the Tinto-Odiel rias (Huelva-Spain) through the application of cluster analysis. Marine Pollution Bulletin, 46: 475-480.
13. Pekey, H., D. Karakas and L.M. Bakog, 2004. Source apportionment of trace metals in surface waters of a polluted stram using multivariate statistical analyses. Marine Pollution Bulletin, 49: 809-818.
14. Kowalkowski, T., R. Zbytniewski, J. Szpejna and B. Buszewski, 2006. Application chemometrics in river watwr classification. Water Res., 40: 744-752.
15. 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.
16. Zeng, X. and T.C. Rasmussen, 2005. Multivariate ststistical characterization of water quality in lake Lanier, Georgia, USA. J. Environ. Qual., 34: 1980-1991.
17. McGarial, K., S. Cushman and S. Stafford, 2000. Multivariate statistics for wildlife and ecology research. Springer, New York
18. Jain, A.K., M.N. Murty and P.J. Flynn, 1999. Data clustering: a review. ACM Computing Surveys, 31(3).
19. Boyacioglu, Hu. and H. Boyacioglu, 2007. Water pollution sources assessment by multivariate statistical methods in the Tahtali Basin, Turkey. Evniron. Geol., 54: 275-282.
20. Mihailov, G., V. Simeonov, N. Nikolov and G. Mirinchev, 2005. Multivariate statistical assessment of the poiiutoin sources along the stream of Kamchia River. Bulqaria Water Sci. Technol., 51(11): 37-43.
21. McKenna, J.E. Jr., 2003. An enhanced cluster analysis program with bootstrap significance testing for ecological community analysis, Environmental Modelling and Software, 18(3): 205-220.
2. Otto, M., 1998. Multivariate methods. In: R. Kellner, J.M. Mermet, M. Otto and H.M. Widmer, (Eds.), Analytical Chemistry. Wiley-VCH, Weinheim.
23. Laurie Kelly and Bryan F.J. Manly, 2005. Multivariate Statistical Methods: A Primer. Chapman & Hall/Crc., pp: 214.