Investigation of Cluster Analysis in Surface Water in Yesilirmak River

Nurgul Ozbay
Engineering Faculty,
Bilecik University, Bilecik, TURKEY
nurgul.ozbay@bilecik.edu.tr

Suheyla Yerel
Bozuyuk Vocational School,
Bilecik University, Bilecik, TURKEY
suheyla.yerel@bilecik.edu.tr

Huseyin Ankara
Department of Mining Engineering,
Eskisehir Osmangazi University, Eskisehir, TURKEY
hankara@ogu.edu.tr

Abstract: The main aim of this study is focused on surface water quality classification of the Yesilirmak River (Turkey) and evaluation of pollution dataset obtained by the monitoring stations. The study shows the application of selected statistical technique to the pollution monitoring dataset, namely, cluster analysis. Cluster analysis is an exploratory data analysis tool for solving classifications problems. Its objective is to sort cases into clusters so that degree of association is strongly members of the same cluster and weak between members of different clusters. The analysis of the monitoring stations identified two clusters. It was concluded that agricultural pollution strongly effected Stations II and Station III. Finally, it was believed to help surface water management to water quality issues and determine priorities to improve surface water quality.

1. Introduction

The surface water quality is a matter of serious concern today. Rivers, due to their role in carrying off the municipal and industrial wastewater and runoff from agricultural land in their vast drainage basins, are among the most vulnerable water bodies to pollution. The surface water quality in a region is largely determined both by the natural process and the anthropogenic influence of water quality (Carpenter et al., 1998, Singh et al., 2005; Yerel, 2009). The particular problem in the case of water quality monitoring has a complexity associated with analyzing the large number of measured variables. The data sets contain rich information about the behavior of the water resources.

The classification and interpretation of monitoring stations are the most important steps in the assessment of surface water quality. In order to determine the data structure, to classify and model the data sets, to reveal time trends and to identify the contribution of pollution etc. cluster analysis should be applied. Some applications of the cluster analysis have also been carried out. Muri (2004) has investigated basic physical and chemical characteristics of water in lakes using cluster analysis. Although the water quality has deteriorated in some lakes, most of the lakes are still in a good condition. Boyacioglu and Boyacioglu (2008) suggested that cluster analysis was applied to assess water quality. In their study, cluster analysis can be used to understand complex nature of water quality issues and determine priorities to improve water quality.

The aim of this study was to examine whether or not the monitoring stations were similar by using single linkage cluster analysis.

2. Material and Methods

2.1. Dataset

Surface water quality dataset covers a year and contains the values of selected pollution indicators for three monitoring stations from the Yesilirmak River in Turkey. Coordinates of the monitoring stations were

depicted in Tab. 1 and selected pollution indicators were given in Tab. 2, respectively.

Station No	X	Y
Station I	299530	4470025
Station II	287150	4468680
Station III	271125	4463355

Table 1. Coordinates of the monitoring stations

Parameter	Symbol	Units
Dissolved oxygen	DO	mg/l
Chloride	Cl^-	mg/l
Sulfate	SO_4^{-2}	mg/l
Ammonium	$NH_4^+ - N$	mg/l
Nitrite nitrogen	$NO_2^- - N$	mg/l
Nitrates	$NO_3^- - N$	mg/l
Total phosphorus	P-tot	mg/l
Total Dissolved Solid	TDS	mg/l

Table 2. Selected pollution indicators

2.2. Cluster analysis

Cluster analysis is an exploratory data analysis tool for solving classification problems. Its objective is to sort cases into groups or clusters, so that the degree of association is strong between members of the same cluster and weak between members of different clusters. Each cluster thus describes, in terms of the data collected, the class to which its members belong; and this description may be abstracted through use from the particular to the general class type (Einax et al., 1998; Kowalkowski et al., 2006). It is evident that the cluster analysis is useful in offering reliable classification of surface water in the whole region and would make possible to design a future spatial sampling strategy in an optimal manner. Thus, the number of observation stations in the monitoring network will be reduced, hence cost without loosing any significance of the outcome (Singh et al., 2005).

In this case of cluster analysis, the similarities-dissimilarities are quantified through Euclidean distance measurements, the distance between two objects, i and j, is given as;

$$d_{ij}^2 = \sum_{k=1}^{m} \left(z_{ik} - z_{jk} \right)^2 \qquad (1)$$

where d_{ij}^2 donates the Euclidean distance, z_{ik} and z_{jk} are the values of variable k for object i and j, respectively, and m is the number of variables (Kowalkowski et al., 2006; Yerel, 2009). Euclidean distance and the Single linkage cluster method were used to obtain dendrograms.

3. Application of cluster analysis to monitoring stations

Cluster analysis organizes sampling entities into discrete groups, such that within-group similarity is maximized and among-group similarity is minimized according to some objective criteria (McGarial et al., 2000). In this study monitoring stations classification was performed by the use of single linkage cluster method. Two major clusters were formed by treating all the by clustering. The dendrogram of the monitoring stations model resulting from the single linkage cluster method of measured surface water quality dataset is presented in the fig. 2.

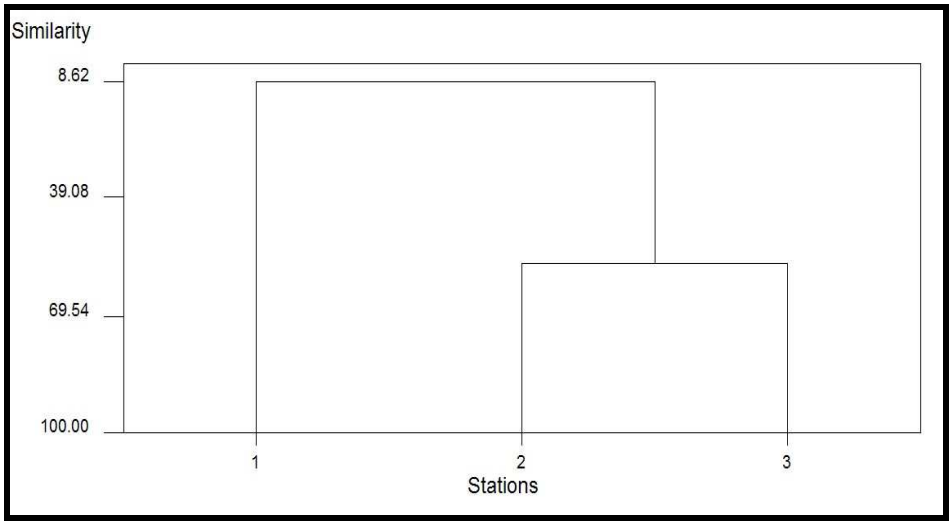

Figure 2 Dendrogram of the single linkage cluster method

The dendrogram shows that all the monitoring stations may be generally grouped into two clusters. Cluster 1 correspond to Station I. Cluster 2 corresponds to Stations II and III. The classification to those clusters varies with the significance level. It is shows that Cluster 1 is characterized by the biggest Euclidean distance to the Cluster 2.

The dataset of the surface water quality parameters were to compare the aspects of the variation in surface water samples collected from three monitoring stations as shown in fig. 3. Among the mean concentrations, all parameters were found very high at monitoring stations II and III.

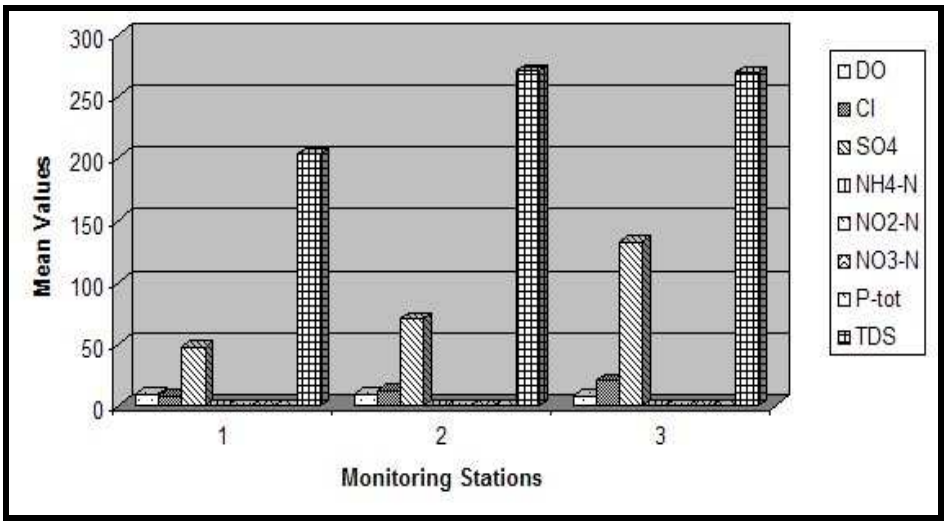

Figure 3 Water quality parameters mean values at Yesilirmak River

4. Conclusion

In this study, cluster analysis were applied to dataset obtain from Yesilirmak River in Turkey. This analysis is important to intercept misinterpretation of monitoring stations dataset due to uncertainties. Cluster analysis grouped three monitoring stations into two clusters of similar water quality characteristics. Based on the above results, it was concluded that agricultural pollution strongly affected Cluster 2. Thus, this study show that usefulness of cluster analysis in water quality assessment, determination of pollution sources with a view to get better information about the monitoring stations.

5. References

1. Boyacioglu, H., & Boyacioglu, H. (2008). Water Pollution Source Assessment by Multivariate Statistical Methods in the Tahtali Basin, Turkey. *Environmental Geology*. 54, 275-282.

2. Einax, J.W., Truckenbrodt, D., & Kampe, O. (1998). River pollution data interpreted by means of chemometric methods. *Microchem. J.*, 58, 315-324.

3. Carpenter, S., Caraco, N. F., Correll, D. L., Howarth, R. W., Sharpley, A. N., & Smith V. H. (1998). Nonpoint pollution of surface waters with phosphorus and nitrogen. *Ecol. Appl*,. 8(3),559-568.

4. Kowalkowski, T., Zbytniewski, R., Szpejna, J., & Buszewski, B. (2006). Application of chemometrics in river water classification. *Water Research*, 40, 744-752.

5. McGarial, K., Cushman, S., & Stafford, S. (2000). *Multivariate statistics for wildlife& ecology research*, Springer, New York.

6. Muri, G. (2004). Physico-Chemical Characteristics of Lake Water in 14 Slovenian Mountain Lakes. *Acta Chim. Slov*. 51, 257-272.

7. Singh K.P., Malik A. & Sinha, S. (2005). Water quality assessment and apportionment of pollution sources of Gomti river (India) using multivariate statistical techniques—a case study. *Analytica Chimica Acta*, Vol. 538.

8. Yerel, S., (2009). Assessment of surface water quality using multivariate statistical analysis techniques: A case study from Tahtali dam, Turkey, *Asian Journal of Chemistry*, 21, 4054- 4062.