Available online at www.sciencedirect.com

ScienceDirect

Procedia Environmental Sciences 30 (2015) 79 – 84

Procedia Environmental Sciences

International Conference on Environmental Forensics 2015 (iENFORCE2015)

Classification of river water quality using multivariate analysis

Shah Christirani Azhar[a,b], Ahmad Zaharin Aris[a,b]*, Mohd Kamil Yusoff[a], Mohammad Firuz Ramli[a,b], Hafizan Juahir[c]

[a]*Department of Environmental Sciences, Faculty of Environmental Studies, Universiti Putra Malaysia,43400 UPM Serdang, Malaysia*
[b]*Environmental Forensics Research Centre, Faculty of Environmental Studies, Universiti Putra Malaysia,43400 UPM Serdang, Malaysia*
[c]*East Coast Environmental Research Institute, Universiti Sultan Zainal Abidin,Gong Badak Campus, 21300 Kuala Terengganu, Malaysia*

Abstract

The classification of river water quality is a useful way of reporting the water quality status of a river to control water pollution in monitored regions. The main objective of this study is to classify the water quality of the Muda River basin (Malaysia) using nine monitoring stations. This study utilised multivariate analysis of cluster analysis (CA), principal component analysis (PCA) and discriminant analysis (DA). CA and PCA identified two different clusters (classes) that reflect the different water quality characteristics of the water systems. DA validated these clusters and produced a discriminant function (DF) that can predict the cluster membership of new samples. The classification generated by the multivariate analysis is consistent with those made by the Department of Environment (DOE). This study demonstrated that multivariate statistical techniques are effective for river water classification.

Peer-review under responsibility of organizing committee of Environmental Forensics Research Centre, Faculty of Environmental Studies, Universiti Putra Malaysia.

Keywords: Classification; multivariate analysis; river water quality; Muda River basin

1. Introduction

Rivers are systems that carry a significant load of materials in dissolved and particulate phases from both natural and anthropogenic sources in one direction [1]. Human activities, use of agricultural chemicals, and land use changes are the major factors that influence surface water quality [2, 3].

* Corresponding author. Tel.: +6-03-8945-7455; fax: +6-03-8943-8109
E-mail address: zaharin@upm.edu.my

Peer-review under responsibility of organizing committee of Environmental Forensics Research Centre, Faculty of Environmental Studies, Universiti Putra Malaysia.
doi:10.1016/j.proenv.2015.10.014

Located in northwest Malaysia, the Muda River basin is affected by organic pollutants from industrial effluents and excess nutrients resulting from agricultural runoff. With an increased understanding of the importance of fresh water systems for public benefit and to aquatic life, the classification of water qualities for effective management options is becoming a concern [4]. For this reason, the classification of river water quality is a useful way of reporting the water quality status of a river as well as identifying the most significant strategies to control water pollution in monitored regions.

In Malaysia, the water quality index (WQI) was developed as a basis for categorising the level of pollution into five classes (Class I–Class V) as stipulated in the National Water Quality Standards for Malaysia (NWQS). This procedure in terms of classification principles described as supervised pattern recognition method [5]. As a result of WQI calculation, the classes of the membership of an object, which in this case are the monitoring stations, are known in advance. Consequently, this approach reduces some variability of the considered set of objects. The supervised pattern recognition may deform the natural pattern of objects' similarities [6].

Grouping of data (objects or variables) is also possible by means of unsupervised methods [5]. Unsupervised methods identify the natural clustering pattern and group monitoring stations on the basis of similarities between the samples. The most common unsupervised methods of multivariate analysis for classification are cluster analysis (CA) and principal component analysis (PCA) [4]. The discriminant analysis (DA) is used to confirm the groups found by means of the CA and PCA. In environmental studies, the application of multivariate analysis to complex datasets has been of great scientific interest over the last few years [7].

In this study, the large data matrix generated during the ten-year (1998–2007) monitoring programme is subjected to different multivariate statistical techniques from nine monitoring stations. The objective of this study is to classify the water quality of the Muda River basin.

2. Materials and methods

2.1. Study area

The Muda River has a total catchment area of approximately 4210 km^2. It lies within latitudes 100^{o}20'33.05"E and 100^{o}56'17"E and longitudes 5^{o}24"56.46"N and 6^{o}10'51.25"N. The catchment is illustrated in Fig. 1. The length of the Muda River is about 180 km, and it is the longest river in Kedah. It flows towards the southern area of the state. This river is the main source of water for domestic use and a source of irrigation for rice cultivation in this catchment area. The land in the basin is mainly used for agriculture (such as paddy, oil palm, and rubber) and forestry.

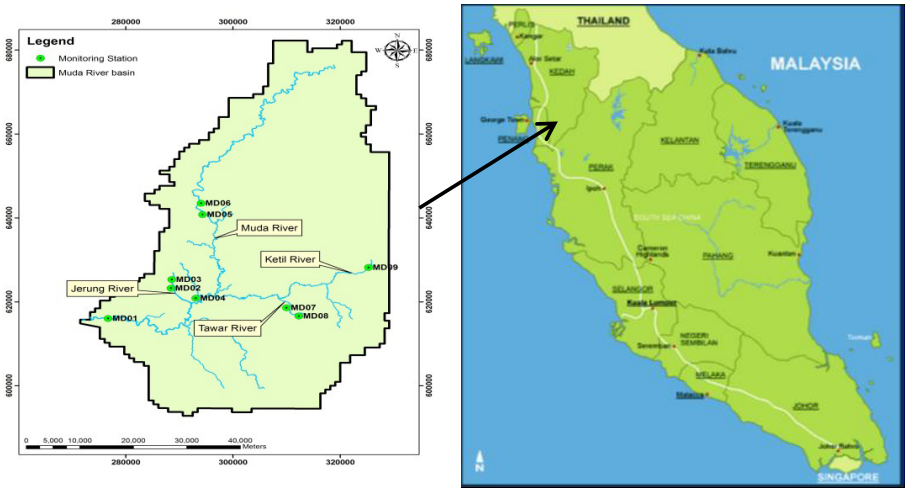

Fig. 1. Location of monitoring stations in the Muda River basin

2.2. The water quality data

The nine monitoring stations (Fig. 1) cover most of the areas of the Muda River basin. The archive data set covers the period from 1998 to 2007 and is provided by the Department of Environment (DOE), Malaysia. The data set comprises six water quality variables: dissolved oxygen (DO), biochemical oxygen demand (BOD), chemical oxygen demand (COD), suspended solids (SS), and pH and ammonia nitrogen (NH_3-N). The data matrix used for classification has the dimension of 9 (monitoring stations) X 6 (variables).

2.3. Statistical procedures

The river water quality data sets were interpreted by integrating three multivariate statistical techniques: CA, PCA, and DA. CA applied to the water quality data set to determine the group monitoring stations for the study regions. In this study, hierarchical CA used Ward's method with squared Euclidean distances as a measure of similarity [8]. PCA was used to quantify the significance of variables that explain the observed groupings and patterns of the inherent properties of the monitoring stations. New orthogonal variables (factors) explained by a reduced set of calculated factors are called principal components (PCs) [9]. The DA was performed to confirm the groups identified by means of the CA. The DA technique builds up a discriminant function (DF) for each group to predict the cluster membership of new cases by means of the cluster centroids.

3. Results and discussion

A dendrogram of the location pattern resulting from the CA is presented in Fig. 2. Using the CA, the monitoring stations of a different character (separate pollution state) may be explicitly distinguished. The dendrogram shows that all the monitoring stations may be grouped into two main clusters (groups). Cluster I is formed by stations MD01 and MD03–MD09 and, Cluster II by station MD02.

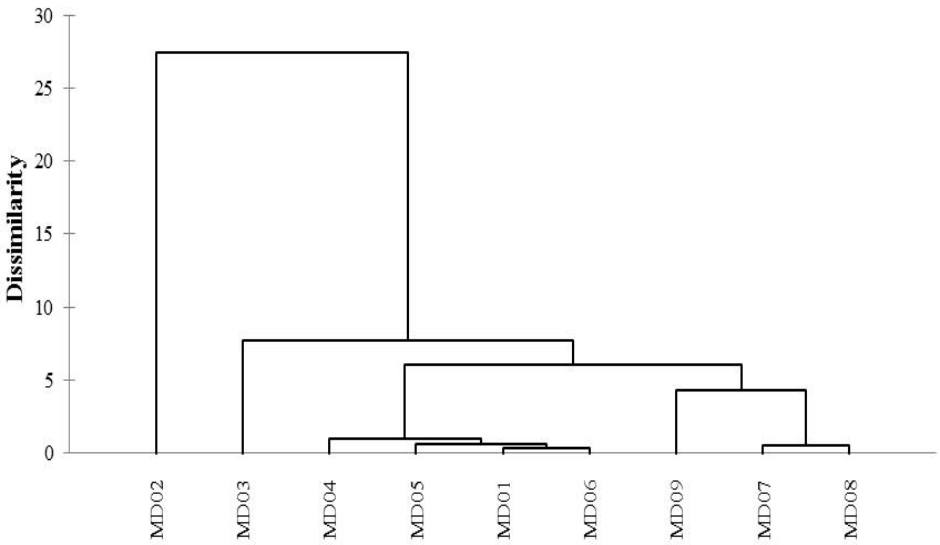

Fig. 2. Dendrogram showing spatial similarities of monitoring stations produced by cluster analysis

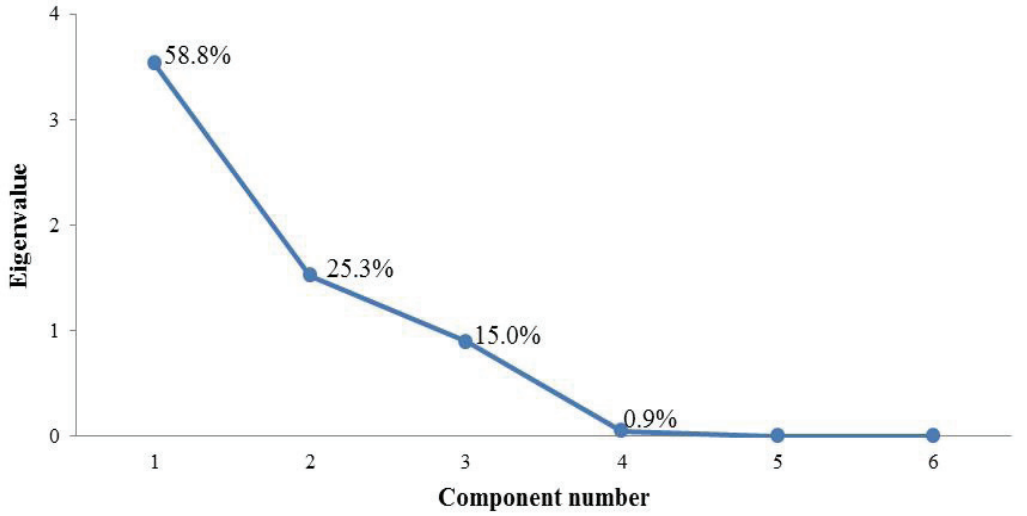

Fig. 3. “Scree-plot” for the principal component model of the monitoring data

A scree-plot (Fig. 3) for the PCA shows that up to 84.1% of the original mean dataset variability is on the first two variables (components). Thus, all information about pollution in the nine monitoring stations collected in the original six variables can be performed in the reduced space. According to the “eigenvalue-one” criterion [10,11], only the PCs with eigenvalues greater than one are considered important. This criterion is based on the fact that the average eigenvalue of the auto-scaled data is just one. These results of PCA are in good agreement with the CA results when two factors are taken into account in the classification of monitoring stations (Fig. 4).

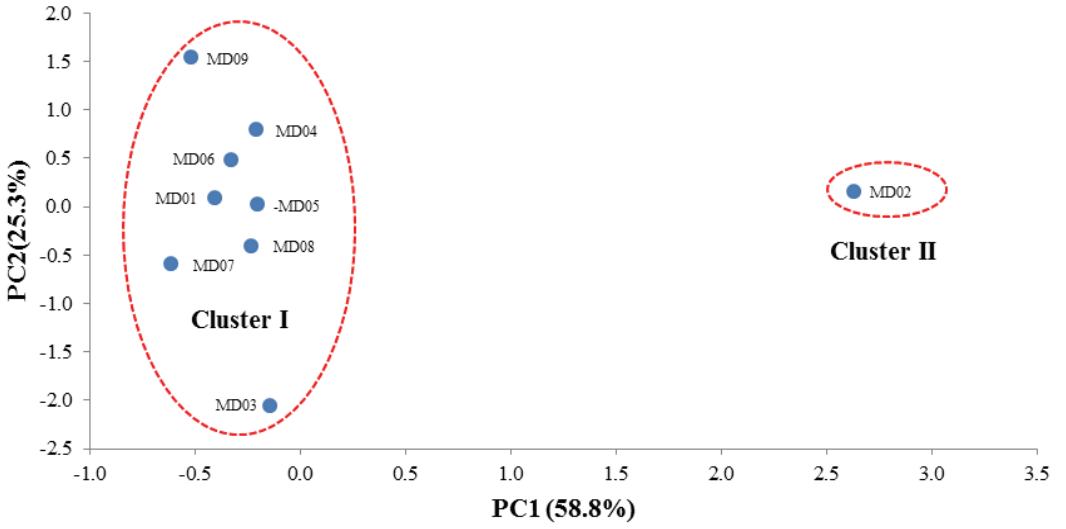

Fig. 4. Principal component scores for mean monitoring data

The two spatial clusters in the data model confirmed by DA that runs on the raw WQ data in the stepwise "mode". Since the monitoring station clusters had two categories, one DF was obtained. The Wilk's lambda values reveal that only ammonia nitrogen satisfies this condition that gives good classification accuracy of 100%. The major features of this function are an eigenvalue of 160.358, a canonical correlation of 0.997 and being significant (Wilk's lambda = 0.006, DF = 1, $p < 0.05$). This DF was developed using the unstandardised DF coefficients and Equation (1). This function can be used to predict cluster membership of new cases through the cluster centroids. For a new sample or a sample of unknown cluster membership, the DF is calculated and compared to the closest centroid. In this study, the centroids of Cluster I and Cluster II were 3.948 and 31.588, respectively.

$$DF = -4.672 + 2.409\ NH_3\text{-}N \quad (1)$$

Fig. 5 shows the classification of the WQI for each station from 1998 to 2007 by the DOE, Malaysia. The nine monitoring stations can be divided into two classes of water quality status, namely Class II and Class III. The monitoring stations in Class II consists of monitoring stations in Cluster I (MD01, MD03–MD09), while the station with Class III consists of monitoring station in Cluster II (MD02), as obtained from CA. This finding indicates that the status of water quality monitoring stations in Cluster I (Class II) is better than Cluster II (Class III). River pollution is higher at station MD02 due to the location of this station adjacent to the rubber plant. Effluent from the rubber industry consists of a complex mixture of chemicals [12] which contribute to water pollution.

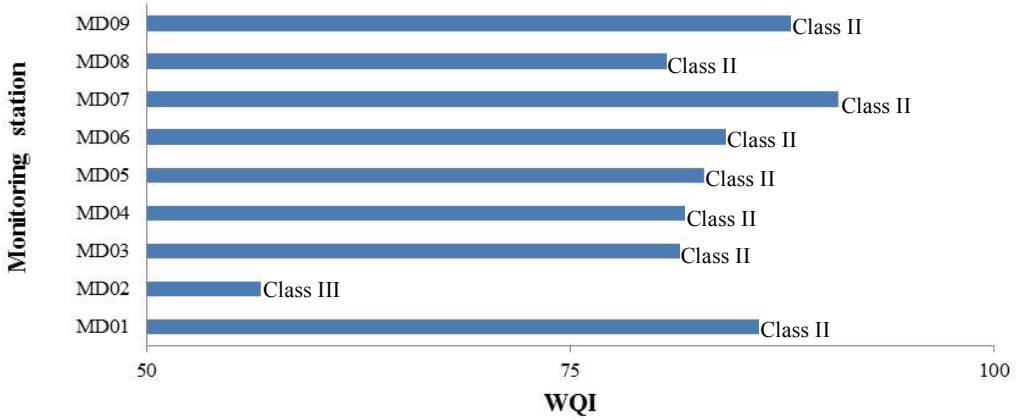

Fig. 5. Classification of the mean of the WQI for each station from 1998 to 2007

4. Conclusion

This study employed three multivariate statistical techniques (CA, PCA, and DA) to classify nine monitoring stations located on the river into groups of similar water quality characteristics based on six selected water quality variables. CA and PCA grouped the nine monitoring locations into two clusters based on similarities in water quality characteristics. DA confirmed the clusters identified by CA and PCA and produced a DF that used one water quality variable (NH_3-N) to distinguish between these clusters. This function exhibited a correct classification efficiency of 100%. The multivariate analysis confirmed the measurement results of the water quality status by the DOE in the sense of water classification.

Acknowledgements

This study was supported by Research Universities Grant Scheme (RUGS) Project No: 03-02-12-1744RU awarded by Universiti Putra Malaysia. The authors would also like to thank the Department of the Environment (DOE) for their permission to use their water quality data for this study.

References

1. Zhang ZI, Tao F, Du J, Shi P, Yu D, Meng Y,Sun Y.Surface water quality and its control in a river with intensive human impacts-a case study of the Xiangjiang River, China. *Environ Manage* 2010; **91**(12):2483-2490.
2. Zhang QL, Shi XZ, Huang B, Yu DS, Öbornc I, Blombäckc K, Wang HJ, Pagella TF, Sinclair FL. Surface water quality of factory-based and vegetable-based peri-urban areas in the Yangtze River delta region, China. *Catena* 2007; **69** (1): 57–64.
3. Hussain M, Ahmed SM, Abderrahman W. Cluster analysis and quality assessment of logged water at an irrigation project, eastern Saudi Arabia. *Journal of Environmental Management* 2008; **86**:297–307.
4. Kannel PR, Lee S, Kanel SR, Khan SP. Chemometric application in classification and assessment of monitoring locations of an urban river system. *Analytica Chimica Acta* 2007; **582**(2): 390-399.
5. Einax JW, Zwanziger HW, Geis S. *Chemometrics in Environmental Analysis*. Wiley, Weinheim; 1997.
6. Kowalkowski T, Zbytniewski R, Szpejna J, Buszewski B. Application of chemometrics in river water classification. *Water Research* 2006; **40** (4): 744–752.
7. Dominick D, Juahir H, Latif MT, Zain SM, Aris AZ. Spatial assessment of air quality patterns in Malaysia using multivariate analysis. *Atmospheric Environment* 2012; **60**: 172–181.
8. Zhaoa Y, Xia XH, Yang ZF, Wang F. Assessment of water quality in Baiyangdian Lake using multivariate statistical techniques. *Procedia Environmental Sciences* 2012; **13**: 1213 – 1226.
9. Osei J, Nyame F, Armah T, Osae S , Dampare S , Fianko J, Adomako D, Bentil N. Application of Multivariate Analysis for Identification of Pollution Sources in the Densu Delta Wetland in the Vicinity of a Landfill Site in Ghana. *Journal of Water Resource and Protection* 2010; **2**(12):1020-1029.
10. Pathak H, Limaye S.N. Study of seasonal variation in groundwater quality of sagar city (India) by principal component analysis. *J. Chem*. 2011; **8**: 2000–2009.
11. Arslan H. Application of multivariate statistical techniques in the assessment of groundwater quality in seawater intrusion area in Bafra Plain, Turkey. *Environ. Monit. Assess*. 2013; **185**:2439–2452.
12. Arimoro FO. Impact of rubber effluent discharges on the water quality and macroinvertebrate community assemblages in a forest stream in the Niger Delta. *Chemosphere* 2009; **77**(3): 440–449