WATER QUALITY PATTERN RECOGNITION STUDY OF MUDA RIVER BASIN USING MULTIVARIATE ANALYSIS

Shah Christirani bt Azhar*, Kamil Mohd Yusof and Hafizan Juahir

PhD (GS23343)
2nd Semester

Introduction

River water quality has become the matter of serious concern because of its effect on human health and aquatic ecosystems. Pollution is the result of the changes in the environment. These changes are due to human socio-economic activities that bring negative effect to humans and other living things around them. Pollution sources can be classified as nonpoint source pollution (NPS) and point source pollution (PS)**.** Nonpoint source pollution is water pollution affecting a water body from diffuse sources, such as polluted runoff from agricultural areas draining into a river, or wind-borne debris blowing out to sea. Nonpoint source pollution can be contrasted with point source pollution, where discharges occur to a body of water at a single location, such as discharges from a chemical factory, urban runoff from a roadway storm drain, or from ships at sea.

Objectives

This study is conducted using secondary data sets of water quality in order (1) to classify sampling stations base on water quality data ,(2) to evaluate spatial variations of river water quality and (3) to determine sources of pollution.

Research Methodology

(a) Study Area: Muda River Basin

Muda River basin which is located within the boundary of Kedah and Pulau Pinang with a catchment area of 4,210 km^2 and 180 km length. Muda River basin is supply water for agricultural, industrial, and domestic for both Pulau Pinang and Kedah.

(b) Data Sets

Secondary data on water quality for Muda River Basin between 2002 and 2006 are obtained from the Department of Environment, Malaysia. 30 water quality parameters are selected in this study . They are Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand(COD), Conductivity (Cond), pH, Temperature(Temp), Suspended Solids (SS), Total Solid (TS), Ammoniacal Nitrogen(NH_3-N) , Salinity(Sal), Turbidity(Tur), Nitrate(NO_3), Chlorine (Cl), Phosphate (PO_4), Zink (Zn), Calcium (Ca), Lead (Pb), Cadmium (Cd), Cromiun (Cr), Arsenic (As), Mercury (Hg) Ferum (Fe), Kalium (K), Magnesium(Mg), Natrium (Na), Dissolved Solid (DS), Total Solid (TS),Organic group(OG) methylene blue active substances (MBAS), *E.coli* and Coliform(Col).

(c) Statistical procedures

Cluster analysis (CA) is an exploratory data analysis tool for solving classification problems. The 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. In this study CA is employed to investigate the grouping of sampling sites (spatial) within the study regions (Einax et al.,1998).

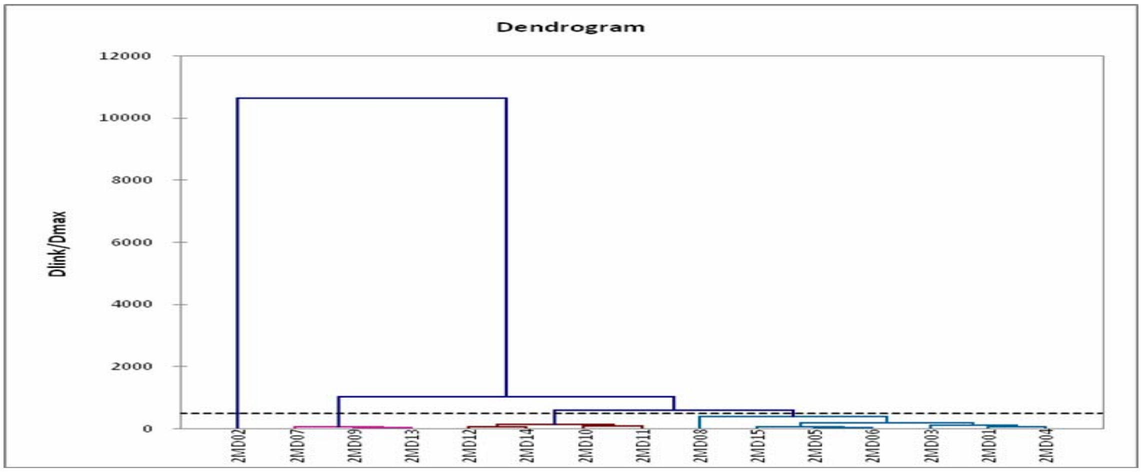
Discriminant analysis (DA) is to construct the best discriminant functions for each group. Groups of spatial data (determined from CA) are selected. Forward/backward stepwise modes are used to construct DFs to evaluate both spatial variations in the river water quality.

Principle Component Analysis (PCA) and Factor Analysis provide information on the most meaningful parameters due to spatial and temporal variations which describes the whole data set. This procedures allowing data reduction with minimum loss of original information

Result and Discussion

Cluster analysis was performed on the water quality data set to evaluate spatial variation among sampling sites. This analysis resulted in the grouping of sampling stations into 4 clusters/groups (Figure1). Cluster 1 (Stesen 2MD04, 2MD01, 2MD03,2MD06,2MD05, 2MD015 dan 2MD08), cluster 2 (Stesen 2MD02), cluster 3 (Stesen 2MD013, 2MD09 dan 2MD07) dan cluster 4 (Stesen 2MD011, 2MD010, 2MD014 dan 2MD012). The clustering procedure generated 4 clusters/groups in a very covincing way, as the sites in these groups have similar characteristics and natural backgrounds. This result implies that for rapid assessment of water quality, only one station in each cluster is needed to represent a reasonably accurate spatial assessment of the water quality for the whole network and can be used to design future spatial sampling strategies in an optimal manner.

Figure 1: Dendogram showing different clusters of sampling sites located at Muda River basin on water quality parameters



To study the spatial variation among different stream regions, DA was applied on the raw data post gruping of the Muda River Basin into four clusters/groups defined by CA. Groups (Cluster 1, Cluster 2 , Cluster 3 and Cluster 4) were treated as dependent variables, while water quality parameters was treated as independent variables. DA was carried out via standard, forward stepwise and backward stepwise methods. The accuracy of spatial classification using standard, forward stepwise and backward stepwise method mode were 84.58% (23 discriminant variables), 77.92% (17 discriminant variables) and 82.08% (20 discriminant variables) respectively (Table 1). Using backward stepwise mode - DO, BOD, COD, SS, pH, NH_3-N, Temperature, Conductivityi,Tur, DS, TS, Cl, PO_4, As, Ca, K, Mg, OG, MBAS and *E-coli* were found to be the significant variables. These indicates that these parameters have high variation in terms of their spatial variation.

Table 1: Classification matrix by DA for spatial variations in Muda River Basin

Sampling regions	% Correct	Regions assigned by DA			
		Cluster 1	Cluster 2	Cluster 3	Cluster 4
Standard mode					
Cluster 1	93.85%	122	0	3	5
Cluster 2	85.00%	3	17	0	0
Cluster 3	92.86%	3	0	39	0
Cluster 4	52.08%	19	0	4	25
Total	84.58%	147	17	46	30
Forward stepwise mode					
Cluster 1	90.00%	117	0	3	10
Cluster 2	80.00%	4	16	0	0
Cluster 3	85.71%	5	0	36	1
Cluster 4	37.50%	26	0	4	18
Total	77.92%	152	16	43	29
Backward stepwise mode					
Cluster 1	92.31%	120	0	6	4
Cluster 2	80.00%	4	16	0	0
Cluster 3	88.10%	5	0	37	0
Cluster 4	50.00%	20	0	4	24
Total	82.08%	149	16	47	28

PCA was employed on the dataset to compare the compositional pattern between the examined water parameters. Two PCs were obtained with eigen values larger than 1 summing 40.24% of the total variance in the dataset respectively. Corresponding varimax factors (VFs), variables loading and variable explained are presented in Table 2. VF1 accounts for 30.65% of the total variance showing strong positive loading on BOD , COD, NH_3-N, Cond, Sal,TS, PO_4, K, Mg and Na. This factor contain chemical parameter that are responsible for the water hardness (Mg) and those attributed to products from anthropogenic activities (DS and TS). DS and TS can be identified to originate from both wastewater treatment plants (PS) and NPS pollution sources (USGS,1999; Ha dan Bae, 2001). The presence of K, Mg and Na increase conductivity and salinity values (Zampella,1995; barnes et al.,1981; Dahlgren dan Singer,1994). Salinity and phosphate has its origin in soils due to the use of phosphate fertilizers in this region. Arheimer and Liden (2000) and Hill (1981), in their studies, conclude that agricultural land use strongly influences stream phosphorous.VF2, explaining 9.59% of the total variance and has strong positive loading on SS, Tur and Fe. This factor, within the region can be attributed to runoffs from fields with high load of soil and waste disposal activities. The presence of Fe basically represents the metal originating from industrial effluents.

Significance of finding

The final results would be helpful for Malaysia Government to optimize the river water monitoring plan and reduce the cost of water quality research. It will also gives better evaluation of the water quality in a monitored region and provides a valuable information to develop assessment strategies.

Table 2: Loading of environmental variables on the Varimax rotated PCs for water quality data collected from Muda River Basin (2002-2006)

Parameter	VF1	VF2
DO	-0.471	-0.044
BOD	**0.914**	0.029
COD	**0.816**	0.132
SS	0.082	**0.817**
pH	0.253	-0.078
NH_3-N	**0.813**	0.033
TEMP	0.172	0.236
COND	**0.893**	-0.013
SAL	**0.898**	-0.015
TUR	0.046	**0.885**
DS	0.550	-0.256
TS	**0.839**	0.427
NO_3	0.006	0.036
Cl	0.334	-0.315
PO_4	**0.758**	0.061
As	0.364	-0.053
Hg	0.098	-0.041
Cd	-0.035	0.000
Cr	0.639	-0.091
Pb	-0.048	-0.087
Zn	0.091	0.344
Ca	0.309	-0.319
Fe	-0.079	0.730
K	**0.956**	-0.025
Mg	**0.890**	-0.185
Na	**0.752**	-0.199
OG	0.360	-0.072
MBAS	0.000	0.000
E-coli	0.139	-0.059
Coliform	0.146	-0.194
Variability (%)	30.648	9.589
Cumulative(%)	30.648	40.237
Eigenvalue	8.892	2.777