Jurnal
Teknologi

Full Paper

SPATIAL CHARACTERIZATION AND IDENTIFICATION SOURCES OF POLLUTION USING MULTIVARIATE ANALYSIS AT TERENGGANU RIVER BASIN, MALAYSIA

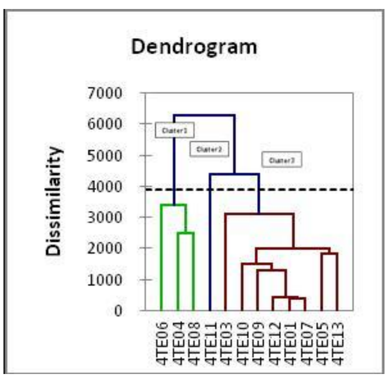
Article history
Received
5 February 2015
Received in revised form
1 September 2015
Accepted
1 October 2015

Ahmad Firdaus Kamaruddin[a], Mohd Ekhwan Toriman[a*], Hafizan Juahir[a], Sharifuddin Md Zain[b], Mohd Nordin Abdul Rahman[c], Mohd Khairul Amri Kamarudin[a], Azman Azid[a]

[a]East Coast Environmental Research Institute (ESERI), Universiti Sultan Zainal Abidin, Gong Badak Campus, 21300 Kuala Terengganu, Terengganu, Malaysia
[b]Department of Chemistry, Faculty of Science, University of Malaya, 50603 Kuala Lumpur, Malaysia
[c]Information Technology Centre, Universiti Sultan Zainal Abidin, Gong Badak Campus, 21300 Kuala Terengganu, Terengganu, Malaysia

*Corresponding author
ekhwan@unisza.edu.my

Graphical abstract



Abstract

The spatial water quality data (281x22) obtained from 12 sampling stations located along the Terengganu River and its main tributaries were evaluated with environmetric methods. Principal component analysis was used to investigate the origin of each variable due to land use and human activities based on the three clustered regions obtained from the hierarchical agglomerative cluster analysis. Six principal components (PCs) were obtained, where six varimax factor (VF) of values more than 0.70 that considered strong loading are discussed. The possible pollution sources identified are of anthropogenic sources, mainly municipal waste, surface runoff, agricultural runoff, organic pollution and urban storm runoff. As a conclusion, the application of environmetric methods could reveal important information on the spatial variability of a large and complex river water quality data in order to control pollution sources.

Keywords: Water quality, principle component analysis, multivariate analysis; Terengganu River, chemometric

Abstrak

Data kualiti air spasial (281x22) yang diperoleh daripada 12 stesen persampelan yang terletak di sepanjang Sungai Terengganu dan cawangan-cawangan utamanya dinilai menggunakan kaedah environmetrik. Analisis komponen prinsipal yang digunakan bertujuan untuk menyelidiki asal-usul bagi setiap pemboleh ubah hasil daripada guna tanah dan aktiviti-aktiviti manusia berdasarkan tiga kawasan kluster yang diperoleh melalui analisis kluster agglomeratif hierarkikal. Enam komponen prinsipal (PC) yang diperoleh, di mana enam faktor varimax (VF) yang mempunyai nilai melebihi 0.70 sahaja dikira sebagai faktor terkuat dan dibincangkan. Antara kemungkinan sumber pencemaran yang dikenalpasti adalah berpunca daripada antropogenik, terutamanya bahan buangan perbandaran, air larian permukaan, air larian pertanian, pencemaran organik dan air larian ribut bandar. Sebagai konklusi, penggunaan kaedah environmetrik mampu untuk mendedahkan informasi penting ke atas keragaman spasial bagi data kualiti air sungai yang kompleks dalam usaha untuk mengawal sumber pencemaran.

Kata kunci: Kualiti air, analisis komponen prinsipal, analisis multivariate, Sungai Terengganu, kemometrik

1.0 INTRODUCTION

River is one of world's most magnificent and highly valuable heritages that are unsurpassed. With approximately 0.0002%of the total water on earth [1], river water continues to be able to sustain the needs of water for all human beings and other organisms for centuries. Besides being the main source of life and living, the river also plays as an important asset for economic resources, including as an important mean of transportation in some countries.

Due to the important role of water and its contribution, water quality and its pollutions should be given more serious attention and priority by all. Water quality is also responsible for controlling the health and the state of the disease of our flora and fauna [2]. The municipal and industrial wastewater discharge, land use, eroded soils and atmospheric pollution are among common major factor of human activities in evaluating the quality of our water bodies. Water pollution is an alteration that makes it unsafe and disrupts the aquatic ecosystem. It may affect the surface water (rivers, lakes) and groundwater. It occurs mainly in surface water by chemical pollution or by viruses and pathogenic bacteria. A stream is considered polluted when the composition or condition of its waters are directly or indirectly modified due to human activity to such an extent that they lend themselves less easily all the uses to which they could be used in their natural state or part of them.

Chemical pollution caused by discharge into rivers mainly comes from organic and inorganic materials. For organic, it may source from degradable and nondegradable substances. The most dangerous among these is the non-biodegradable materials (plastics, etc.) which will accumulate and remain toxic across the time. These advanced products do not break down itself naturally to be dissolved with the surrounding land or soil. It requires a very high cost technology and involves other dangerous chemical in order to degrade them [3]. As for inorganic compounds, they mainly come from the extraction of mineral fuels, minerals, building materials, cleaning workshops, processing and packaging of these minerals or materials.

Rain also plays a very important role in contributing to urban pollution where it washes the polluted atmosphere of vehicle fumes, boiler, factories and industrial sites. The rain gets all the pollution that dispersed into the air when it starts to rain. It also washes industrial and urban soils such as factories, highways, parking lots, airports that contaminated with pollutants such as hydrocarbons, lead and zinc which finally ends up in the river. The major pollutions that affect the river in Malaysia attributed from sewage disposal, small- and medium-sized industry effluents, land clearing, logging and other earthwork activities. Up to 42% of suspended solids (SS) were contributed by poorly planned and uncontrolled land clearance activities alone, while 30% with biological oxygen demand (BOD) from industrial discharges, and 28% of ammoniacal nitrogen (AN) due to animal farming activities and domestic sewage disposal [4].

The chemometric techniques were used to conduct spatial analysis in evaluating the most significant variables that affect the health of the river [4]. In the environmental study, the application of different multivariate analysis such as cluster analysis (CA) and principal component analysis (PCA) are part of the environmetric technique that often been used to deal with complex and huge amount of data sets. These techniques assist in transforming the complex data matrices into an easier way to interpret and therefore offer a better understanding of the water quality and ecological status of the studied systems. This will ease the identification process of the possible factors that influence the water systems and offer a valuable tool for reliable management of water resources as well as rapid solutions to pollution problems. It also can be used to estimate the relationship among the factors and therefore it can provide valuable insight into the sources of pollution [5-7]. The objective of this study is to determine the possible main contributor of pollution of the Terengganu River by implementing multivariate analysis. By highlighting the most significant water quality parameters that affect the river, the pollutant agent can be identified and therefore the source of pollution can be tracked.

2.0 METHODOLOGY

Study Area

Terengganu River basin is situated in Terengganu State which located on the East Coast of Malaysia. Its main upstream originated from Kenyir Lake in northeast Malaysia and it flows through Kuala Terengganu, the state capital of Terengganu. As a major river system in the state, Terengganu River consists of several main tributaries, namely Pueh, Berang, Telemong and Nerus. All of these rivers flow into the Terengganu River and lastly, it flows into the South China Sea. 12

selected water quality monitoring stations are illustrated in Figure 1. The stations are manned by the Department of Environment (DOE), Ministry of Natural Resource and Environment Malaysia.

Figure 1 Map of the 12 sampling station locations

Data

There is a total of 281 samples with 22 water quality parameters derived from those stations along the main Terengganu River and its main tributaries ranging from the year 2006 until 2010. The parameters are dissolved oxygen (DO), biochemical oxygen demand (BOD), electrical conductivity (EC), chemical oxygen demand (COD), ammoniacal nitrogen (AN), pH, suspended solid (SS), temperature (T), salinity (Sal), turbidity (Tur), dissolved solid (DS), nitrate *(*NO^{-3}*)*, phosphate *(*PO_4^{3-}*)*, arsenic (As), mercury (Hg), cadmium (Cd), chromium (Cr), lead (Pb), zinc (Zn), *Escherichia coli (E. coli)* and coliform.

Data Pre-treatment

Firstly, when conducting a secondary data, the whole data will be rearranged according to the stations and year of monitoring. Then pre-treatment of data is done in order to trace any void and missing data, unreadable symbols and numerals, and any value that is below or above the detection limit. Thus, it will increase the effectiveness of multivariate analysis and to overcome any error faced while implementing XLSTAT add-in software to the data. Approximation using an average of nearest neighboring data was performed while some data that are closed or under the limitation of measurement which can be identified with the mathematical figure (≤) in front of the parameter value were set to half of its detection limit to make it legitimate for XLSTAT to analyze the data.

Hierarchical Agglomerative Clustering Analysis

In order to classify objects based on its characteristics individually and group them together according to their similarity with respect to a predetermined selection criterion, cluster analysis (CA) is used for the primary purpose. The obtained results will groups objects into both homogeneity and heterogeneity class at the same time. Homogeneity refers to the similar character within the cluster while heterogeneity refers to differences between the clusters. Hierarchical agglomerative clustering analysis (HACA) is the most common applied approach which provides intuitive similarity relationships between any one sample and the entire data set, and is typically illustrated by a dendrogram. The dendrogram provides a visual summary of the clustering process by presenting a simplified picture of the groups and their proximity, with a drastic reduction in dimensionality of the original data [4].

In this study, HACA was performed on the normalized new data set by means of the Ward's method which consists of the mean values of every station, where it uses variance analysis 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. For measuring the similarity, a squared Euclidean distance is used where it gives the similarity between two samples and a distance can be represented by the difference between analytical values from the samples. The spatial variability of water quality in the whole river basin was determined from HACA, using the linkage distance, reported as Dlink/Dmax, 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 [7, 13, 14].

Principal Component Analysis

Regarded as being the most powerful technique for pattern recognition, PCA can be used to explain the variance of a large set of interrelated variables and transforming it into a smaller set of uncorrelated variables called principal components [8]. PCA target is to reveal a more concealed set of factors that accounts the major pattern across all the original variables [9]. This technique highlights the most significant parameters due to spatial and temporal variations by rendering data reduction that describes the whole data set and excluding the less significance parameters with minimum loss of original information [10-14]. PCA is also sensitive to outliers, missing data, and poor linear correlation between variables due to inadequate assigned variables [15].

The PC's score of a^{th} PC and b^{th} sample is expressed as: $y_{ab} = z_{a1}x_{1b} + z_{a2}x_{2b} + z_{a3}x_{3b} + \ldots + z_{am}x_{mb}$ (1)
Where z is the loading and x is the measured value of a water quality parameter, *a* is the component number, *b* is the sample number, and *m* is the total number of variables. These PCs were subjected to varimax rotation (raw) generating VFs. Rotations were done in order to maximize the number of factors in common to the number of variables used in the study. Thus, it reduces the large scope of the data and eventually identifies the most significant new variables

[16]. In this study, only varimax factors with values more than 0.70 will be discussed. Variables with loadings greater than 0.7 are considered strong, while variables with loadings from 0.7 to 0.5 are moderate and variables with loadings lower than 0.5 are considered a weak variables [4].

3.0 RESULTS AND DISCUSSION

After going through the data pre-treatment, HACA was used to detect the similarity groups between the monitoring sites. It yielded a dendrogram (Figure 2), grouping all 12 monitoring stations of the basin into three statistically significant clusters at (Dlink/Dmax)*100<4000. Three statistically significant clusters formed are: Cluster 1 corresponds to sampling sites 4TE04, 4TE06 and 4TE08 which represent low polluted sites. These three sites are situated in the upstream of the river. Cluster 2 corresponds to sampling site 4TE11, which is a high polluted site. It is located in the most developing and highly occupied area. Meanwhile 4TE01, 4TE03, 4TE06, 4TE07, 4TE09, 4TE10, 4TE12 and 4TE13 are grouped together in cluster 3 with medium polluted sites. These eight remaining sites are mainly situated in the moderately occupied and mostly agricultural area (oil palm and rubber tree plantation). The obtained clusters seem quite convincing where they are grouped according to their similar characteristics and background features that affected by similar sources based on 22 selected water quality parameters.

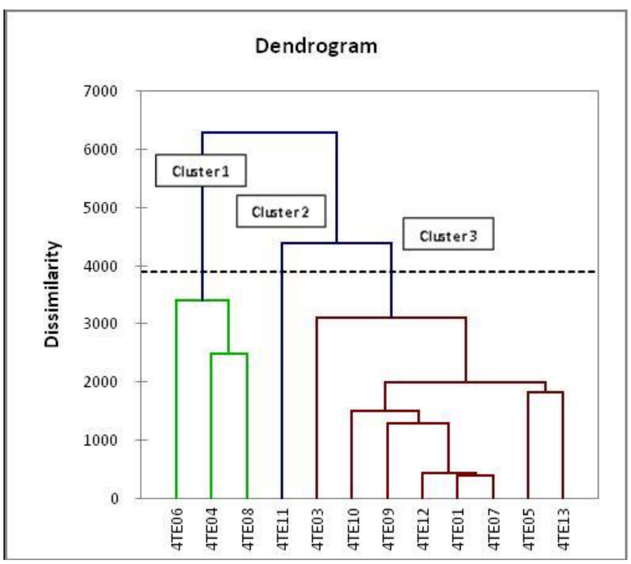


Figure 2 Dendogram showing different clusters of sampling sites located on the Terengganu River Basin based on water quality parameters

PCA with varimax rotation was employed on the data [13]. Six PCs were obtained with eigenvalues >1 summing almost 65.75% of the total variance in the water dataset (Table 1). Eigenvalues of 1.0 or greater are considered significant [17]. Six Varimax Factors (VFs) which equal to the number of PCs were achieved after varimax rotation (Table 1 and Figure 3). Eigenvalues and the corresponding factors were sorted by descending order.

Table 1 Factor loadings after Varimax rotation

	VF1	VF2	VF3	VF4	VF5	VF6
DO (% Sat)	-0.251	0.006	**0.890**	-0.086	-0.120	0.083
DO (mg/l)	-0.291	0.009	**0.887**	-0.074	-0.110	0.140
BOD (mg/l)	0.035	-0.017	-0.087	**0.964**	-0.011	0.026
COD (mg/l)	0.116	0.049	-0.030	**0.961**	0.052	0.071
SS (mg/l)	0.069	0.415	0.085	0.141	-0.048	0.604
pH (unit)	-0.414	-0.051	0.494	0.067	0.331	-0.197
NH3-NL (mg/l)	0.479	-0.017	-0.312	0.270	0.427	-0.059
TEMP (Deg C)	0.475	-0.026	-0.309	-0.059	-0.049	-0.357
COND (uS)	**0.890**	-0.008	-0.276	0.081	0.083	0.005
SAL (ppt)	**0.895**	-0.009	-0.270	0.084	0.070	0.012
TUR (NTU)	0.000	0.423	0.102	0.309	-0.055	0.644
DS (mg/l)	**0.919**	-0.007	-0.207	0.032	0.011	-0.010
NO3 (mg/l)	-0.056	-0.128	0.194	-0.027	0.049	**0.741**
PO4 (mg/l)	0.048	0.025	-0.377	0.019	0.516	0.091
As (mg/l)	0.231	-0.050	-0.222	-0.054	0.660	0.047
Hg (mg/l)	-0.113	-0.218	-0.230	-0.093	-0.145	0.330
Cd (mg/l)	0.212	0.016	0.072	0.080	-0.101	-0.210
Cr (mg/l)	0.034	0.041	-0.053	0.067	**0.751**	-0.053
Pb (mg/l)	0.000	0.000	0.000	0.000	0.000	0.000
Zn (mg/l)	**0.818**	-0.084	0.144	0.043	0.028	-0.054
E-coli (cfu/100ml)	-0.058	**0.845**	-0.025	0.036	0.015	-0.006
Coliform (cfu/100ml)	-0.048	**0.793**	0.008	-0.035	0.002	0.132
Eigenvalue	5.329	2.544	1.778	1.641	1.304	1.211
Variability (%)	25.376	12.114	8.467	7.816	6.207	5.767
Cumulative %	25.376	37.490	45.957	53.773	59.980	65.748

*Values in bold are considered strong loadings

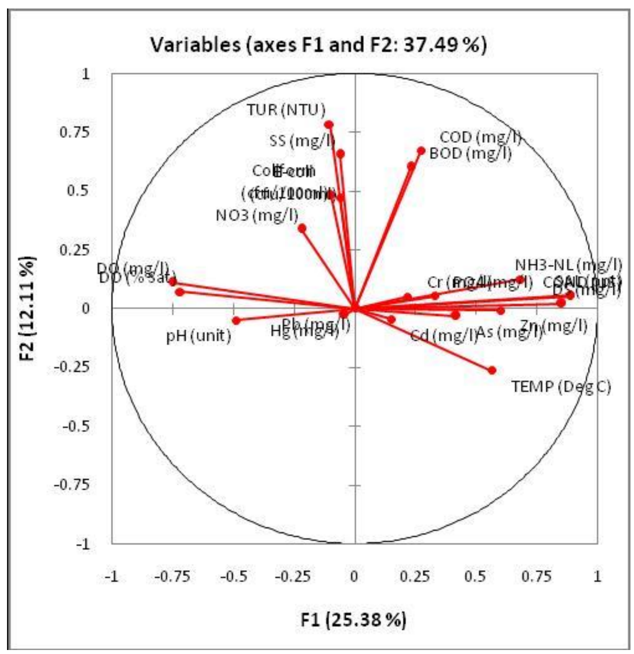


Figure 3 Factor loadings of variables after Varimax rotation

According to the obtained Varimax Factor, VF1 represent 25.38% of the total variability, which has strong positive loadings on EC, SAL, DS and Zn. It can be interpreted as the soil erosion, especially due to sand mining activities observed along the river and high dissolved salt level [7], especially sodium since the salinity level is high. As for Zn, by considering the large number of houses and buildings along the river that use metallic roofs coated with zinc, when in contact with acid rainwater and smog, these could readily mobilize zinc into the atmosphere and waterways [4].

The VF2 with 12.11% of the variance shows strong positive loading of *E-coli* and coliform, which represents fecal waste according to its factor [19]. *E. coli* is strongly related to raw and municipal sewage mainly from domestic and poultry farm, surface runoff and discharge from wastewater treatment plants [4]. This can be clearly observed where there are poultry farms (chicken, cow and goat) in a few parts of the Terengganu river side, and also there are many aquaculture freshwater fish farms along the river all the way from the upstream to the most downstream. The strong loading of VF3 with 8.47% of variance is due to the reason that the concentration of DO is affected by factors such as flow of the river, present of sources of organic pollution, temperature of the water and assimilative capacity of the river [19]. The Terengganu river flow rate is moderately slow and some layer of oil can be seen on some part of the surface river water where the sand mining activities are actively operating due to the waste oil released by the sand mining machine.

The VF4 strong factor loading with 7.82% variance shows two parameters that accounts for BOD and COD represent the anthropogenic input typically from organic pollution. VF explained the high levels of dissolved organic matter and biological organic matter, possibly comes from runoff of solids or waste disposal activities [20]. VF5 has a strong positive loading only on one parameter, Cr with 6.21% of the variance. This may attributed from urban storm runoff due to urban activities [19]. The VF6 strong loading with 5.77% of the variance for NO^{-}_{3} is possibly due to agricultural runoff, since nitrogen based fertilizers are still commonly used in all oil palm and rubber tree plantation along the Terengganu river. The positive correlation of NO^{-}_{3} and agricultural is consistent with many other studies [4].

4.0 CONCLUSION

In this study, different multivariate techniques applied to evaluate spatial variations in surface water quality of the Terengganu river basin consist of 12 monitoring stations have been grouped into three clusters of similar water quality characteristics. PCA managed to extract and identify the main possible factors that contribute to variation in river water quality based on point sources and non-point sources. These apportionment results are believed to be very useful for the local authorities in order to effectively manage the source of pollution of the examined area and also as a reference for further research in the future.

Acknowledgement

The authors would like to acknowledge the Department of Environment, Ministry of Natural Resources and Environment Malaysia for their permission to utilize the water quality data and the financial support of Ministry of Higher Education Malaysia for funding this project (FRGS Project number: RR061).

References

[1] United States Geological Survey. The World's Water. [Online]. From: http://water.usgs.gov/edu/earthwherewater.html. [Accessed on 19 November 2014].
[2] Kazi, T. G., Arain, M. B., Jamali, M. K., Jalbani, N., Afridi, H. I., Sarfaz, R. A., Baig, J. A., Shahazi, A. Q. 2008. *Ecotoxicology and Environmental Safety*. *72*(2): 301-309.
[3] Thompson, R. C., Moore, C. J., Saal, F. S.v., Swan, S. H. 2009. *Philos Trans R Soc Lond B Biol Sci.* 364(1526): 2153-2166.
[4] Juahir, H., Zain, S. M., Yusoff, M. K., Hanidza, T. I. T., Armi, A. S. M., Toriman, M. E., Mokhtar, M. 2010. *Environmental Monitoring and Assessment.* 173(1-4): 625-641.
[5] Morales, M. M., Marti, P., Llopis, A., Campos, L., Sagrado, J. 1999. *Analytica Chimica Acta*. 394: 109-117.
[6] Reghunath, R., Murthy, T. R. J., Raghavan, B. R. 2002. *Water Research*. 36(1): 2437-42.
[7] Alberto, W. D., Pilar, D. M., Valeria, A. M., Fabiana, P. S., Cecilia, H. A., Maria, B. 2001. *Water Research.* 35(12): 2881-94.
[8] Hopke, P. K. 1985. *Receptor Modeling in Environmental Chemistry*. New York: Wiley.
[9] Saim, N., Osman, R., Sapian, D. R. S. A., Jaafar, M. Z., Juahir, H., Abdullah, M. P., Ghani, F. A. 2009. *Water Research.* 43(20): 5023-5030.
[10] Vega, M., Pardo R., Barrado, E., Deban, L. 1988. *Water Research.* 32(12): 3581-3592.
[11] Helena, B., Pardo, R., Vega, M., Barrado, E., Fernandez, J.M., Fernandez, L. 2000. *Water Research.* 34(3): 807-816.
[12] Krishna, A. K., Satyanarayanan M., Govil, P. K. 2009. *Journal of Hazardous Materials*. 167(1-3): 366-373.
[13] Singh, K. P., Malik, A., Mohan, D., Sinha, S., Singh, V. K. 2004. *Water Research.* 38(18): 3980-3992.
[14] Singh, K. P., Malik, A., Mohan, D., Sinha, S., Singh, V. K. 2005. *Analytica Chimica Acta.* 532(1): 15-25.
[15] Sarbu, C., Pop, H. F. 2005. *Talanta*. 65(5): 1215-1220.
[16] Nasir, M. F. M., Zali, M. A., Juahir, H., Hussain, H., Zain, S. M., Ramli, N. 2012. *Iranian Journal of Environmental Health Science & Engineering*. 9(1): 18.
[17] Kim, J.-O., Mueller, C. W. 1987. Introduction to Factor Analysis: What It Is and How To Do It. Series: Quantitative Applications In The Social Sciences. Newbury Park: Sage Publications.
[18] Song, M. W., Huang, P., Li, F., Zhang, H., Xie, K. Z., Wang, X. H. and He G. X. 2011. *Environmental Monitoring Assessment.* 172(1-4): 589-603.
[19] Saad, F. N. M., Rahman, N. N. N. A, Kadir, M. O. A., Omar, F. M. 2008. Project Report: Identification of Pollution Sources within the Sungai Pinang River Basin, University Science of Malaysia.
[20] Yeung, I. M. H. 1999. *Environmental Monitoring and Assessment*. 59(3): 331-342.