Available online at www.sciencedirect.com
ScienceDirect
Procedia Environmental Sciences 30 (2015) 73 – 78
Procedia Environmental Sciences

International Conference on Environmental Forensics 2015 (iENFORCE2015)

Characterization of water quality conditions in the Klang River Basin, Malaysia using self organizing map and K-means algorithm

Sharifah Mohd. Sharif[a,*], Faradiella Mohd. Kusin[a,b], Zulfa Hanan Asha'ari[a], Ahmad Zaharin Aris[a,b]

[a]*Department of Environmental Sciences, Faculty of Environmental Studies, Universiti Putra Malaysia, 43400 Serdang, Malaysia*
[b]*Environmental Forensics Research Centre, Faculty of Environmental Studies, Universiti Putra Malaysia, 43400 UPM Serdang, Malaysia*

Abstract

This study aimed to determine the spatiotemporal pattern of the water quality data and identifying the sources of pollution in the Klang River Basin. The self organizing map (SOM) combined with the K-means algorithm arranged the data based on the relationships of 25 variables. The data from 2006 to 2009 for 30 monitoring stations were classified into six clusters. Water pollution in this river basin originated primarily from urban runoff, construction sites, faulty septic systems and industrial activities. The application of machine learning approaches is highly recommended to extract valuable information from the data for a holistic river basin management.

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

Keywords: Water quality; spatiotemporal pattern; pollution sources; machine learning; self organizing map; K-means

1. Introduction

A great concern of protecting aquatic ecosystems has significantly increased in order to conserve the availability of the resources through sustainable development [1]. The Klang River Basin is considered as the most developed area in Malaysia with the highest rate of urban growth. The availability of clean water is decreasing due to

* Corresponding author. Tel.: +603-89468596; fax: +60-89467463
E-mail address: sharifahmohdsharif@gmail.com

Peer-review under responsibility of organizing committee of Environmental Forensics Research Centre, Faculty of Environmental Studies, Universiti Putra Malaysia.
doi:10.1016/j.proenv.2015.10.013

environmental degradation [2]. The challenges in river basin management can be eased by identifying the sources of pollution and trend of water quality conditions. Machine learning techniques are useful to solve real life problems involving the field of environmental sciences through data analysis, modelling and visualization [3]. Only linear structures can be correctly extracted from the data using the linear techniques, whereas water quality assessment is complex and fuzzy that needs correct methods to solve the non-linear relationship between assessment factor and quality grade [1].

Self organizing map (SOM) is a type of unsupervised Artificial Neural Network (ANN). The component maps of SOM are capable of visualizing many of the additional non-linear relationships between the variables that cannot be expressed in the commonly used Principle Component Analysis (PCA) [1,4]. K-means clustering is often utilized after the process of the SOM network[4]. In this study, water quality data measured in the Klang River Basin were analyzed using the SOM coupled with K-means algorithm to determine the spatiotemporal pattern of the water quality data and to identify the sources of pollution. Besides being able to detect the characteristics of water quality conditions, optimized sampling and monitoring strategies are also possible for decision makers to manage river basins effectively and economically.

2. Methodology

2.1. Study area and data used

Klang River Basin covers an area of 1288 km^2. Klang River (3° 13′ 01.33″, 101° 40′ 54.92″) is approximately 120 km length. The main river and its tributaries flow through the Federal Territory of Kuala Lumpur and part of the state of Selangor and eventually into the Straits of Malacca. The upper catchment that comprises the Selangor districts of Gombak and Hulu Langat is mountainous and still covered by tropical forest. The middle catchment is the populous urban areas within Kuala Lumpur. The lower catchment includes the Selangor districts of Petaling and Klang[2]. Half of the river basin has been developed for residential, infrastructure and utilities, 20% are permanent forest reserves, and other land uses include institutional, open space, community facilities, commercial, industrial and squatter. The rainfall peaks in April and November to December. The months with the lowest humidity are June, July and September. The average annual temperature ranges between 29 and 32^{0}C. April to June recorded the maximum temperatures [5].

About 25 water quality variables were obtained from the Department of Environment (DOE), Malaysia encompassing dissolved oxygen (DO), biochemical oxygen demand (BOD), chemical oxygen demand (COD), ammoniacal nitrogen (NH_3-N), suspended solids (SS), dissolved solids (DS), pH, temperature, conductivity, salinity, turbidity, total solids (TS), nitrate (NO_3^-), chloride (Cl^-), phosphate (PO_4^-), arsenic (As), chromium (Cr), zinc (Zn), calcium (Ca), iron (Fe), potassium (K), magnesium (Mg), sodium (Na), *Escherichia coli* (*E. coli*) and coliform. The data (bimonthly from January 2006 to November 2009) were collected from 30 monitoring stations.

2.2. Two-level clustering

A total of 720 entries (6 months x 4 years x 30 monitoring stations) for each water quality variable were gathered. After data pre-treatment, a two-level clustering approach was utilized. Creating a network using SOM in the first level to determine the relationships between the multiple water quality variables and to reduce them to a few independent variables. Observations with parallel water quality were then clustered in the second level by employing K-means algorithm. All statistical and mathematical calculations were made using the add-ins applications for Microsoft Excel 2010. The SOM and K-means clustering were applied using GeoSOM version 1.0 software, while the group characterization was applied using TANAGRA version 1.4.41 software. Attained

3. Results and discussion

3.1. Self organizing map (SOM) interpretations

The SOM maps were formed using an 11 x 12 hexagonal architecture with a Gaussian neighbourhood function based on the optimum number of neurons (m = 134). The learning algorithm was run consecutively by having variables act as input vectors in order to classify the observations/sampling events (monitoring stations – sample period) based on similar water quality characteristics. The training stage was selected with parameters as follows: 1000 (first phase) and 4000 (second phase) full data iterations, initial neighbourhood width, α = 0.05 (first phase) and 0.02 (second phase) and initial radius, R = 10 (first phase) and 3 (second phase). The relationships between the standardized variables [0, 1] can be indicated through the visualization of the component maps of selected variables (Fig. 1). The high neuron values relate to orange shades, whilst the lowest ones link to blue shades [3].

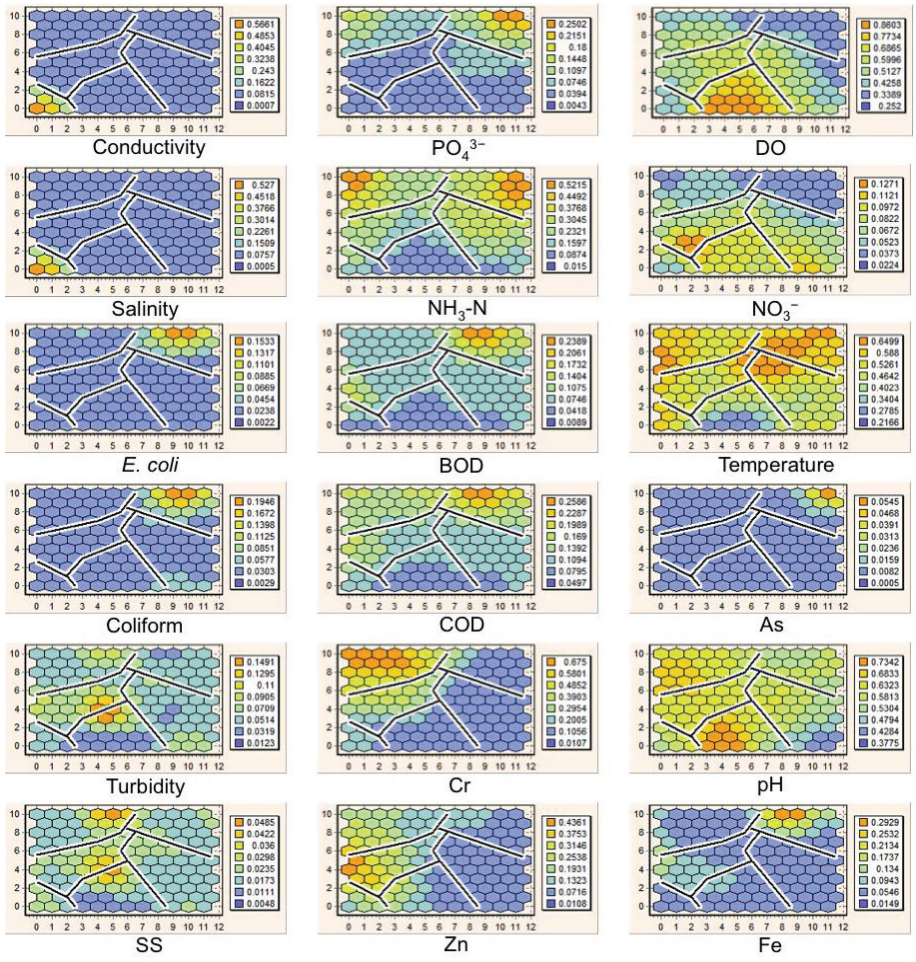

Fig.1. Component maps of selected water quality variables

Parallel colour gradients specify a positive correlation between variables, whereas anti-parallel gradients indicate an inverse correlation [4,6,7]. The strong positive correlated variables such as conductivity, salinity, DS, TS, Cl, Na, Mg, K and Ca have nearly identical map patterns. Other strong correlated variables include BOD – COD, *E. coli* – coliform and turbidity – SS. Since the increase in shade is at the opposite side of the maps, DO was found negatively correlated with other variables particularly NH_3-N and temperature. Complexity in every variable that has no clear correlation with other variables can be detected and included in the maps due to the capability of the SOM to express the non-linear relationships [4].

3.2 Cluster structure of the water quality data

The number on each hexagon of SOM hits map signifies the number of observations classified into each neuron (Fig. 2a). K-means map further defined the number of clusters into six (Fig. 2b) based on the reference vectors of the SOM. In order to recognize the spatiotemporal pattern, data that were classified into each cluster with respect of the stations on a bimonthly basis were considered. Over the periods of sampling, stations that were consistently classified into C1 are upstream stations: 1K24 (Gombak River), 1K28 (Batu River) and 1K10 (Klang River). Other stations that were frequently classified into this cluster especially in May (2006 – 2008), September (2008) and November (2007, 2008 and 2009) include the middlestream (1K18) and downstream (1K17) stations along the Gombak River. The middle stream of Batu River (1K20) was also alternately classified into this cluster. C1 is associated with the best water quality condition in this river basin with the highest levels of DO and pH. However, this cluster also contains high level of NO_3^-. The extensive use of fertilizers for agricultural activities and faulty septic systems may be the sources of NO_3^-. The high nutrient loading may be due to rainfall intensity when soil and nutrients are most vulnerable to erosion [8-10].

Station 1K01, which is located near the estuary was the only station classified into C6. C6 represents the extremely high levels of conductivity, salinity, DS, TS, Cl, Na, Mg, K and Ca, and significantly high temperature, SS and Zn. The high availability of mineral salt components may be linked to natural sources such as atmospheric salt cycling and river bed erosion, and anthropogenic sources such as industrial wastewater [8].

Indistinct spatial pattern to distinguish between C2 (less polluted), C3 (moderately polluted), C4 (polluted) and C5 (highly polluted) was observed. Most data measured from January 2006 to September 2007 were classified into C4 and followed by C3. The subsequent sampling periods (November 2007 to November 2009) contains data that were mainly classified into C2 and C5. C2 indicates the less polluted condition but with high level of NO_3^-. Data associated with C2 were mostly from stations along the Klang (1K02 – 1K09 and 1K25), Kuyoh (1K15), Damansara (1K11 – 1K13), Ampang (1K23) and Gombak (1K17 and 1K18) rivers. The data were mainly measured in November 2007 – September 2008 and all months in 2009 excluding July. C3 contains mostly data from station 1K27 at the Bunos River (constantly from January 2006 – March 2007). Data from other stations were also alternately classified into this cluster, especially stations along the Klang (1K08 and 1K25), Gombak (1K17 and 1K18), Damansara (1K11 and 1K13), Kuyoh (1K15) and Keroh (1K30) rivers. The most frequent data classified into this cluster were measured in January (2006 and 2007) and November (2006). This cluster comprises moderate levels of BOD, DO, Fe and NO_3^-.

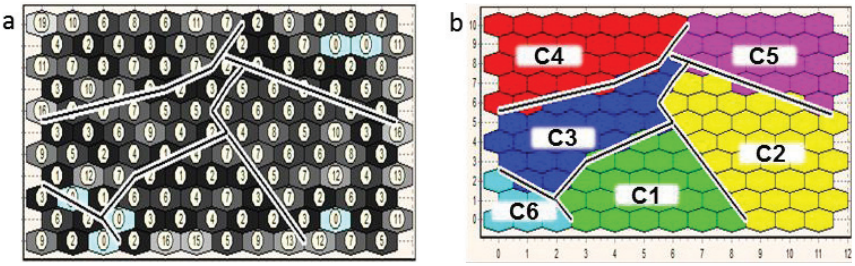

Fig. 2. Maps produced from the two-level clustering: (a) SOM hits map; (b) K-means map

C4 encompasses mostly data from stations along the Jinjang (1K22 and IK29) and Kerayong (1K16 and 1K26), Klang (1K02 – 1K09), Batu (1K19), Penchala (1K14), Keroh (1K21 and 1K30) and Damansara (1K11 and IK12), Ampang (1K23) and Kuyoh (1K15) rivers. Data measured in March, July and September (2008 and 2009) and May (2008) were frequently classified into this cluster. C4 represents the high levels of NH_3-N, PO_4^{3-} and temperature. C5 is highly linked to stations along the Kerayong (1K26), Penchala (1K14), Keroh (1K21 and 1K30), Jinjang (1K22 and 1K29) and Bunos (1K27) rivers. Most data measured in the driest month of 2009 (July) were classified into this worst water quality condition with the highest levels of PO_4, NH_3-N, BOD, coliform, Fe, COD, *E. coli*, temperature and As and the lowest level of DO.

Aside from urban runoff, domestic wastes, sewage treatment plants, septic tanks, the polluted rivers also received effluent discharge and wastes from industrial area. Districts with major industrial activities include Petaling, Klang, Gombak (Selayang and Seri Gombak) and Hulu Langat (Kajang and Ampang). Squatter settlements that exist along the riverbanks (particularly the Klang, Kerayong, Keroh and Jinjang rivers and followed by the Batu, Gombak and Ampang rivers that flow through Kuala Lumpur) may as well raise the water pollution level. The temporal pattern proves that the water quality conditions were highly influenced by land use modifications and seasonal variation. Surface runoff will be directed into waterways during heavy rainfall [11]. In some places, pollutants might be washed out with heavy rainfalls. However, most of the urban development in the Klang River Basin has taken place on the region that is prone to flooding. Construction and deforestation have led to the increase in impervious surface area. These developed environments do not allow the rain to infiltrate into the ground. Such alterations can affect the quality of water by the high transit speed and high volume of stormwater runoff [10]. In dry season, worse water quality conditions can be observed. This may be associated with high temperature that can reduce the DO concentrations and increase the amount of pollutants [7,12,13,14].

4. Conclusion

The complex relationships of 25 physicochemical and biological variables have been simplified using the SOM. The water quality dataset were then classified into six clusters in the next level by applying the K-means algorithm. The deteriorating river water quality in the hotspot areas was primarily due to rapid urbanization. Besides the more commonly recognized point sources that include direct discharge from industrial areas and sewage treatment plants, water pollution in this river basin was mainly initiated from the non-point sources such as urban runoff, construction sites and faulty septic systems. Sustainable development is compulsory when economic growth alone does not provide better lives but needs a healthy environment. In order to improve the weaknesses of river basin management, this research proposes employment of advance techniques using available data. Rather than applying the conventional methods of benchmarking schemes, machine learning algorithms suggest better solution for pattern recognition. This can provide reliable results for decisions to be made after identifying the sources of problem that deteriorating the quality of environment. Further work will aim at identifying the optimal prediction models for water quality conditions in this river basin based on different statistical models using linear and non-linear multivariate.

Acknowledgements

The authors would like to express the greatest acknowledgement to the Department of Environment (DOE), Malaysia for providing the water quality data. We are also grateful to the Ministry of Higher Education, Malaysia for the MyBrain15 scholarship and to those who have given guidance throughout the preparation of this paper.

References

1. Juntunen P, Liukkonen M, Lehtola M, Hiltunen Y. Cluster analysis by self-organizing maps: An application to the modelling of water quality in a treatment process. *Appl Soft Comput* 2013;**13**:3191-3196.
2. Othman F, Alaa EM, Mohamed I. Trend analysis of a tropical urban river water quality in Malaysia. *J Environ Monit* 2012; **14**:3164-3173.
3. Kanevski M, Pozdnoukhov A, Timonin V. *Machine learning for spatial environmental data: Theory, applications and software*. Switzerland: EPFL Press; 2009.
4. Gamble A, Babbar-Sebens M. On the use of multivariate statistical methods for combining in-stream monitoring data and spatial analysis to characterize water quality conditions in the White River basin, Indiana, USA. *Environ Monit Assess* 2012;**184**:845-875.
5. Sezer EA, Pradhan B, Gokceoglu C, Manifestation of an adaptive neuro-fuzzy model on landslide susceptibility mapping: Klang valley, Malaysia. *Expert Syst Appl* 2011;**38**:8208-8219.
6. Lamrini B, Lakhal EK, Lann MV, Wehenkel L. Data validation and missing data reconstruction using self-organizing map for water treatment. *Neural Comput Appl* 2011; **20**:575-588.
7. Jin YH, Kawamura A, Park SC, Nakagawa N, Amaguchi H, Olsson J. Spatiotemporal classification of environmental monitoring data in Yeongsan River Basin, Korea, using self-organizing maps. *J Environ Monit* 2011;**13**:2886-2894.
8. Gandaseca S, Rosli N, Ngayop J, Arianto CI. Status of water quality based on the physico-chemical assessment on river water at Wildlife Sanctuary Sibuti Mangrove Forest, Miri Sarawak. *Am J Environ Sci* 2011;**7**:269-275.
9. Bae HK. Changes of river's water quality responded to rainfall events. *Environ Ecol Res* 2013;**1**:21-25.
10. Coulliette AD, Noble RT. Impacts of rainfall on the water quality of the Newport River Estuary (Eastern North Carolina, USA). *J Water Health* 2008;**303**:473-482.
11. Nnane DE, Ebdon JE, Taylor JD. Integrated analysis of water quality parameters for cost-effective faecal pollution management in river catchments. *Water Res* 2011;**45**:2235-2246.
12. Naji A, Ismail A, Ismail AR. Chemical speciation and contamination assessment of Zn and Cd by sequential extraction in surface sediment of Klang River, Malaysia. *Microchem J* 2010;**95**:285-292.
13. Gevrey M, Comte L, De Zwart D, De Deckere E, Lek S. Modeling the chemical and toxic water status of the Scheldt basin (Belgium), using aquatic invertebrate assemblages and an advanced modeling method. *Environ Pollut* 2010;**158**:3209-3218.
14. Kusin FM, Jarvis AP, Gandy CJ. Hydraulic performance assessment of passive coal mine water treatment systems in the UK. *Ecological Engineering* 2012;**49**:233-243.