TEST Engineering & Management

November-December 2019
ISSN: 0193-4120 Page No. 819 - 825

Influential Water Quality Parameter in Juru River Basin using PCA

Basharoh Abdul Karim[1], Shamshuritawati Sharif[2], Haslina Zakaria[3], Tun Mohd Firdaus Azis[4]

[1]Kolej Poly-Tech MARA Alor Setar, Alor Setar, Kedah, Malaysia.
[2]School of Quantitative Sciences, Universiti Utara Malaysia, Sintok, Kedah, Malaysia
[3]Institute of Engineering Mathematics, Universiti Malaysia Perlis, Arau, Perlis, Malaysia.
[4]Faculty of Applied Sciences, Universiti Teknologi Mara cawangan Perlis, Arau, Perlis, Malaysia.

Article Info
Volume 81
Page Number: 819 - 825
Publication Issue:
November-December 2019

Article History
***Article Received:** 3 January 2019*
***Revised:** 25 March 2019*
***Accepted:** 28 July 2019*
***Publication:** 25 November 2019*

Abstract

The river pollution or contamination is caused by several factor such as growth in the industrial development, urbanization transformation and increase in the population density. River pollution may negatively effect human health, well-being and the environmental condition. Juru River basin is among the contaminated rivers in Malaysia which located in Pulau Pinang, Malaysia. Even though many green activities have been implemented but the result still the same. In this paper, Principal Component Analysis is employed for pinpointing the most influential parameter in Juru River Basin. It is important to know what the root cause of the water quality is. There are 20 parameters in measuring water quality, therefore, PCA is useful statistical techniques in handling a larger number of parameter. Bimonthly data is gathered from 2008 to 2017 Department of Environment (DOE). From the analysis, there are eight influential parameter that have to control which are salinity, conductivity, dissolved solids, calcium, zinc, pH, arsenic and ammonia nitrogen. We expect to help the DOE to put the focus on these eight parameters instead of monitoring 20 parameters. Once we control all of the identified influential parameters, the rest will be in-control.

***Keywords:** Data reduction, Contaminated river, Principal component analysis, Water quality*

I. INTRODUCTION

Every one of us are put high expectation to have a clean river because the river is a key of surface water that can influence all the human daily life activities [1]. The number of contaminated rivers in Malaysia is projected to rise if the monitoring and recovery actions are not instantaneously and extensively made. From time to time there is a growth in the number of contaminated rivers. In 1998, there are 7 rivers, while after ten years it becomes 16 rivers in year 2008, and dramatically upturn to 46 contaminated rivers in year 2016 [2[, [3]. The change in the level of river pollution is caused by several factor such as growth in the industrial development, urbanization transformation and increase in the population density [4].

It give a negative impact on the environmental condition and its users. It can cause water supply crisis, threaten human health, reduce the amount of river species habitat which contribute to biodiversity and ecosystem imbalances. Besides, the pollution will indirectly affects the tourism industry in the form of rejuvenation and leisure. The river pollution is a universal problem that requires the cooperation of all parties to address it.

II. JURU RIVER, PULAU PINANG

Generally, Juru River Basin consists of four main rivers, specifically Juru River, Kilang Ubi River, Pasir River and Rambai River. The basin is placed in Pulau Pinang, Malaysia. Originally, all of four main rivers are located at Bukit Mertajam that having an area of 75km^2. All source from the river flow into the Straits of Malacca [1]. Crucially, Juru River Basin is among the contaminated rivers in Malaysia since 1976 [5]. Until now, it still known as the most contaminated river in Malaysia with Class IV in the year 2004 [6].

To recover the river pollution, the government as well as the non-governmental organizations initiate many green initiative activities, in example, Effective Microorganisms (E.M) Mud balls program. The EM Mud balls is the friendly and greenly solution to the environmental which can reduce water pollutants, and thus refining the quality of river water and fluids in the drains. The turmoil produced from the mud balls can lighten and abolish the NH_3-NL that usually contains in the human overflows and sewerage outflows into the water system. Beside that, an activity such as eco-camp park, river rehabilitation by Malaysian Resources Corporation Bhd, regular water health check, and many more activities have been working cooperatively in Juru River [7]. However, water quality of Sungai Juru is still contaminated even though the index was slightly improved from Class IV to Class III with a value of 55 [3]. Yet, the water is still not suitable to drink and require intensive treatment for drinking water.

There are several factors and activities that can be a root causes to the polluted river in Juru area such as manufacturing industry, timber activities, rubber industry, urbanization development for residential house and commercial building, untreated domestic wastes, and also an animal wastes [1], [6]. Additionally, the contaminated streams of water can disrupt the mangrove tree along the Juru river because there are a lot of contaminated that will end up or flow into the sea. Moreover, contaminated river will destroy the fish habitats, shrimps, and crabs whereby it might affect fishermen activities.

In this paper, to elaborate more on quality of water, we investigate the Juru River using a multivariate statistical method. At the end of the analysis, several influential parameters are statistically recognized for measuring the quality level of river water. Generally, there are 20 parameters is taking into consideration based on data that can be provided by DOE. Multivariate statistical technique known as principal component analysis (PCA) will play it role in reducing the number of parameter. With the identification, the pollutant or contaminant parameter can be recognized and therefore pollution root causes can be detected in mathematical way.

The rest of this paper is organized as follows. In the Section III, we will discuss the methodology, specifically we describe the secondary data that been used and the Principal Component Analysis that employed in investigating the most influential water quality parameters. Next, we deliver the results of the investigation in Section IV, followed by conclusions at the last section.

III. METHODOLOGY

In this section, we deliver the way to get ready the data for analyzing using principal component analysis (PCA).

A. Secondary Data

Firstly, the bimonthly data is gathered from the Department of Environment Malaysia (DOE). They provide us with 20 water quality parameters which are dissolved oxygen (X_1), biochemical oxygen demand (X_2), chemical oxygen demand (X_3), suspended solids (X_4), Ph (X_5), ammoniacal nitrogen (X_6), temperature (X_7), conductivity (X_8), salinity (X_9), turbidity (X_{10}), dissolved solids (X_{11}), total solids (X_{12}), nitrate (X_{13}), phosphate (X_{14}), arsenic (X_{15}), chromium (X_{16}), zinc (X_{17}), calcium (X_{18}), potassium (X_{19}), and magnesium (X_{20}). All of the parameter has been identified by DOE for measuring the quality of water. A collection of data for the past 10 years from 2008 to 2017 is used for PCA.

B. Data reduction technique

PCA is one of technique to reduce the data that introduced by Karl Pearson in 1901. It can reduce the number of parameters from a multivariate data to a new set of parameters by maintaining as much as possible important information [1]. It is the oldest multivariate technique [8].

The new set of parameters generated from this analysis are called Principle Components (PCs). The principal component (PC) can be express as follow.

$$Y_p = a_{p1}X_1 + a_{p2}X_2 + ...+ a_{pp}X_p \tag{1}$$

Where Y_p is the score of component, a_{pi} is the factor loading, X_p the parameter value. The Principal component will retain all the information that contains in all of the original parameters.

PCA are commonly used as the first phase in analyzing large data sets. The data set can have many parameters consist of similar or different scales of measurement. Besides that, this technique particularly useful when the parameters within the data set are highly interrelated. Interrelated indicates that there is redundancy in the data. Due to this redundancy, PCA can be used to reduce the original parameters into a PCs which explaining most of the variance in the original parameters.

In this study, PCA was accomplished to identify the most influential parameters to a water quality for a quick classification and easy monitoring. There are several steps and conditions before the PCA can be implemented.

Step 1. Data standardization.

For data that have many parameters with different scale of measurement, the data standardization needs to be done to get a reliable composite. Equation 2 is used to standardize a parameter.

$$Z_{ij} = \frac{X_{ij} - \bar{x}_j}{s_j}, \tag{2}$$

Where: X_{ij} is the raw data of parameter j, $\bar{x}_j$ is sample mean of parameter j, and s_j is sample standard deviation of parameter j. For a dataset that contains more than one parameter but using the same scale of measurement, it is not necessary to make sure of a data standardization.

Step 2: Measure of covariance

In what follows, covariance between x_1 and x_2 is calculated using Equation 3 to quantify the changes in two different parameters.

$$\text{cov}(x_1, x_2) = s_{12} = \frac{1}{n-1}\sum_{j=1}^{n}(x_{11} - \bar{x}_1)(x_{12} - \bar{x}_2). \tag{3}$$

We can compute the covariance matrix to look all the possible covariance in a dataset like this.

$$S_n = \begin{bmatrix} s_{11} & \cdots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{1p} & \cdots & s_{pp} \end{bmatrix}$$

In general, we have diagonal and non-diagonal elements of covariance matrix. The diagonal elements is represent the eigenvalues while non-diagonal elements is represent the variances. The eigenvalues is the amount of variation that retained by each of the principal component. There are two approach for deciding the number component to be retain.

Step 3. Deciding number of component retained

First approach is Guttmann-Kaiser criterion. Principle component with the eigenvalues more and equal to 1 were retained for the further analysis [9]. For example, if eigenvalues for the first, second, third and fourth principal components were 3.56, 2.97, 1.80 and 0.75, respectively. Therefore, we will retain Y_1, Y_2, and Y_3 for the further analysis since the three principal component consist of eigenvalues greater than 1.00. Large Eigenvalues are for the first PCs and small for the next PCs. Therefore, the PCs matches to the directions with the maximum amount of variation in the data set.

Besides that, the second approach is introduced by Cattell (1966). He proposed a graphical scree plot to determine the number of principal components to be retained. A scree plot is a plot of the eigenvalues in the y-axis while the component number in the x-axis. The plot always displays a downward curve. The component that appears before the slope of the curve is clearly flattened out are retained for further analyze while component that appear after it are not retained [9].

IV. RESULT AND DISCUSSION

In this analysis, we decide to implement the Guttmann-Kaiser criteria. From Table 1, the statistical results indicate that there are five principal components that must be maintained with an eigenvalue greater than 1. All of the first five principal components (Y_1 , Y_2 , Y_3 , Y_4 and Y_5) represent 78.08% of the variability in the data set with eigenvalues is around 15.62. Mathematically, the total variance is computed using eigenvalues. In details, the eigenvalues are divided by the number of parameters under study then multiplied the value by 100%. The highest total variance is derived from the highest eigenvalues described in the original parameters. There are three main conclusion of factor loading: Strong, Moderate, Poor: The parameters that produce a factor loading with value more than 0.7 are considered strong, while parameters that having a factor loading from 0.5 to 0.7 is moderate [10], and parameters that having a factor loading lower than 0.5 is poor. Therefore, in this analysis, we consider the factor loading that more than 0.5 [11].

As shown in the Table 1, the first principal component clarifies the total variance close to 42% with eight parameters show a strongly negative loading. The factor loading shows that the first principal component will increase when all the eight parameters which are X_5, X_6, X_8, X_9, X_{11}, X_{15}, X_{17}, and X_{18} decreases. It means that eight parameters significantly affect each other. Therefore, if the measurement value of one parameter increases, the reduction will occur to the remaining parameters and vice versa. Thus, the other parameters should be examined immediately if there are any changes to any of the above parameters. From the results, the two most strong factor loading in the first principal component are X_8, and X_9. In practice, the inorganic dissolved solids and organic compounds such as chloride, calcium, nitrate, aluminum cations, iron, magnesium, sulfate, sodium, alcohol, phenol, oil, and sugar causing X_8 to exist in water [12]. While intrusion of saltwater into the river mouth of the mangroves results in high X_9 content [13].

Additionally, the second principal component with a strong negative loading on X_3, and X_{12}, while adequate negative loading on X_{13}, and X_4 explains the total variance close to 13%. The second principal component will increase when there is a decrease in the parameter X_3, X_4 , X_{12}, and X_{13} . This means these four parameters interrelated each other. The parameters X_{12} should be monitored first compared to X_{13}, and X_4 if there is any change on the X_3 . The factor loading values on X_3, X_{12}, and X_4 parameters are due to the disposal of sewage from irregular settlements and the process of erosion of rock and river banks that occur naturally. In addition, the X_{13} content in this component also reflects surface runoff from agricultural activities [1].

The third principal component has positive loading on X_{19} and X_{20} , while moderate negative loading on X_2 which explains the total variance close to 10%. This component shows that the X_2 content in the river will decrease if the factor loading value on X_{19} and X_{20} increases. The X_2 is the key indicator of water contamination caused by the disposal of industrial waste, domestic waste, organic sources and a very high concentration of bacteria in sewage treatment [14].

TABLE 1. Factor loading for five principal components with detail of each parameters.

Pr	Y_1	Y_2	Y_3	Y_4	Y_5	Cm
X_1	0.077	0.153	-0.403	0.551	-0.325	0.602
X_2	-0.173	-0.017	-0.642	-0.118	-0.274	0.531
X_3	0.107	-0.834	0.058	0.128	0.152	0.750
X_4	-0.421	-0.532	0.087	-0.310	0.117	0.577
X_5	-0.978	0.010	0.065	0.087	-0.060	0.972
X_6	-0.969	0.032	0.052	0.083	-0.069	0.955
X_7	0.253	-0.339	-0.337	0.654	-0.239	0.778
X_8	-0.992	-0.011	0.040	0.042	-0.053	0.990
X_9	-0.992	-0.010	0.036	0.048	-0.056	0.990
X_{10}	0.338	0.296	-0.027	-0.231	-0.491	0.497
X_{11}	-0.983	0.017	0.068	0.066	-0.027	0.976
X_{12}	0.030	-0.788	-0.113	0.277	0.137	0.729
X_{13}	-0.225	-0.575	-0.338	-0.349	0.098	0.627
X_{14}	-0.083	0.434	0.353	0.353	0.519	0.714
X_{15}	-0.976	0.007	0.026	0.073	-0.055	0.962
X_{16}	0.214	-0.003	0.257	0.591	0.074	0.467
X_{17}	-0.979	-0.007	0.044	0.027	-0.069	0.965
X_{18}	-0.982	0.056	0.042	0.059	-0.091	0.981
X_{19}	0.203	-0.250	0.721	0.001	-0.392	0.777
X_{20}	0.263	-0.347	0.639	-0.012	-0.422	0.776
TVE	**8.320**	**2.532**	**1.966**	**1.617**	**1.181**	
% of TVE	**41.601**	**12.660**	**9.830**	**8.087**	**5.905**	

Key: Pr = Parameters, Cm=Communalities, Y_p =PC

The fourth PC show a moderate positive loading on X_1 , X_7, and X_{16} with explains approximately 8.087% of the total variance. Although these three parameters are at a low level, they still influence each other. This shows that changes in river water temperature will cause changes in X_{16} and X_1 . However, the differences in weather conditions, time and location of sampling can affect the temperature value which its effect the result of X_1 and other parameters [11].

The fifth principal component shows a moderate positive loading on X_{14} which explains approximately 5.905% of the total variance. This component is derived from the process of manufacturing agricultural products. Together these results found that these five principal components also explained communalities (Cm) value for 8 parameters which $X_5, X_6, X_8, X_9, X_{11}, X_{15}, X_{17}$, and X_{18} is more than 95% of the variability. The parameter X_8 and X_9 have the largest communalities value which is 0.990 or 99.0% of the variability. This means that in the process of monitoring the river water quality, these 8 parameters that influence each other should be given priority.

V. CONCLUSION

In this study, all the water quality parameters that measured from the Juru River have been reduced into five principal components by using PCA. PCA is shown to be an effective technique in investigating and identifying the main possible parameters that contribute to river water quality. These five principal components obtained from this analysis can interpret the relationship between all water quality parameters. The parameters that basically contribute to the water quality change was classified in a similar group. It makes the process of monitoring the management of water resources can be done quickly and efficiently. According to the five principal components X_5, X_6, X_8, X_9, X_{11}, X_{15}, X_{17}, and X_{18} which are pH, ammoniacal nitrogen, conductivity, salinity, dissolved solids, arsenic, zinc and calcium are the parameters that influence each other in the water quality at Juru River.

ACKNOWLEDGMENT

We are gratefully acknowledged to the Department of Environment of Malaysia (DOE) for providing and authorization their water quality parameters to be used for this study. Special thanks to Kolej Poly-Tech MARA Alor Setar (KPTMAS) for the financially supported and Universiti Utara Malaysia for research software provided. We are also very grateful and would like to thank the reviewers for their valuable ideas, which have led to a great improvement of the article.

REFERENCES

[1] H. Juahir, M. A. Zali, & A. Retnam. (2011). Spatial characterization of water quality using principal component analysis approach at Juru River Basin, Malaysia. *World Applied Sciences Journal*. *14*, 55-59.

[2] M. S. M. Zin, H. Juahir, M. E. Toriman, M. K. A. Kamarudin, N. A. Wahab, A. Azid, (2017). Assessment of water quality status using univariate analysis at Klang and Juru River, Malaysia. *Journal of Fundamental and Applied Science*, *9*(2S), 93–108.

[3] Department of Environment. (2018, October). *Water Quality Data Set*. Available: http://www.data.gov.my/data/ms_MY/dataset?q=water quality.

[4] S. A. Muyibi, A. R. Ambali, & G. S. Eissa. (2008). The impact of economic development on water pollution: Trends and policy actions in Malaysia. *Water Resources Management*. *22(4)*, 485–508.

[5] A. H. Rahman,” Suatu Tinjauan Terhadap Isu Pencemaran Sungai Di Malaysia,” unpublished.

[6] O.E. Toriman, N. Hashim, A. J. Hassan, M. Mokhtar, H. Juahir, M. B. Gasim & M. P. Abdullah. (2011). Study on the impact of tidal effects on water quality modelling of Juru River, *Malaysia Asian Journal of Scientific Research*. ***4****(2)*, 129–138.

[7] N. C. Chuan, (2014). Water pollution in Juru River, Penang Malaysia. Available:https://tunza.eco-generation.org/ambassadorReportView.jsp?viewID=9327.

[8] S. P. Mishra, U. Sarkar, S. Taraphder, S. Datta, D. Swain, & R. Saikhom, & M. Laishram. (2017). Multivariate Statistical Data Analysis- Principal Component Analysis (PCA). *International Journal of Livestock Research*, *7(5)*, 60–78.

[9] M. J. Masnan, A. Zakaria, A. Y. Md. Shakaff, N. I. Mahat, H. Hamid, N. Subari and M. J. Salleh. (2012). Principal Component Analysis–A Realization of Classification Success in Multi Sensor Data Fusion. *Principal Component Analysis Application*, 1-25.

[10] I. Mohd, A. M. Mansor, M. R. A. Awaluddin, M. F. M. Nasir, M. S. Samsudin, H. Juahir & N. Ramli. (2011). Pattern Recognition of Kedah River Water Quality Data by Implementation of Principal Component Analysis. *World Applied Science Journal*, ***14***, 66-72.

[11] A. R. Peterson. (2000). A Meta-Analysis of Variance Accounted for and Factor Loadings in Exploratory Factor Analysis. *Marketing Letters*. 11(3), 261-275.

[12] F. Al-Badaii, M. Shuhaimi-Othman and M. B. Ghasim. (2013). Water Quality Assessment of the Semenyih River. *Journal of Chemistry*. *2013* (5), 31–34.

[13] S. I. Khalit, M. S. Samsudin, A. Azid, K. Yunus, M. A Zaudi, S. S. Sharifuddin & T. M. Husin. (2017). River water quality assessment using APCS-MLR and statistical process control in Johor River Basin, Malaysia. *Journal of Fundamental and Applied Sciences*. 13(4), 764-768.

[14] S. Bhasin, A. N. Shukla, S. Shrivastava & U. Mishra, (2016). Control Chart Model for Assessment of Water Quality of a Tropical River-Kshipra Ujjain, India. *Haya:The Saudi Journal of Life Sciences*. 51–64].

AUTHORS PROFILE

Basharoh Abdul Karim completed her Bachelor of Science (Honors) in Mathematics at Universiti Putra Malaysia (UPM) on February 9, 2002. She has started her career as a lecturer in Kolej Poly-Tech MARA Alor Setar (KPTMAS) since June 2002, and her teaching experience had exceeded 17 years. She had obtained her Master of Science (Data Analysis) from Universiti Utara Malaysia in 2019. Her main expertise is in statistics and business mathematics.

Shamshuritawati Sharif currently a senior lecturer at Universiti Utara Malaysia. She obtained a PhD in Mathematics at Universiti Teknologi Malaysia. Her Master degrees in Decision Science were from Universiti Utara Malaysia in 2003. Her diploma and Bachelor of Science in Statistics were obtained from MARA Institute of Technology (ITM) in 2000. She is interested in multivariate hypothesis testing, industrial statistics, statistical quality control, network analysis and centrality measure.

Haslina Zakaria completed her Decision Science first degree in Universiti Utara Malaysia in 2004. She have started her career with Institute Engineering of Mathematics in 2006, and her teaching experience had exceed 13 years in Universiti Malaysia Perlis (UNIMAP). She had obtained her Master of Science (Data Analysis) from Universiti Utara Malaysia in 2019. Her main expertise is in engineering mathematics and statistics.

Tun Mohd Firdaus Bin Azis currently a lecturer at Universiti Teknologi MARA Cawangan Perlis. Her Master degrees in Entomology were from Universiti Kebangsaan Malaysia in 2013. He is interested in study of insect, and Biostatistics.