International Journal of Engineering & Technology, 7 (4.43) (2018) 83-89

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Short communication

Pollution Sources Identification of Water Quality Using Chemometrics: a Case Study in Klang River Basin, Malaysia

Mohd Saiful Samsudin[1*], Azman Azid[1], Saiful Iskandar Khalit[1], Shazlyn Milleana Shaharudin[3], Fathurrahman Lananan[1,2], Hafizan Juahir[2]

[1]*Faculty Bioresources and Food Industry, Universiti Sultan Zainal Abidin (UniSZA), Besut Campus, 22200 Besut, Terengganu, Malaysia.*

[2] *East coast Environmental Research Institute (ESERI), Universiti Sultan Zainal Abidin, Gong Badak Campus, 21300 Kuala Nerus, Terengganu, Malaysia.*

[3] *Department of Mathematics, Faculty of Sciences, Universiti Pendidikan Sultan Idris, Malaysia*

**Corresponding author E-mail: saifulsamsudin294@gmail.com*

Abstract

The objectives of this paper are to review the status of Klang River Basin and to investigate the most significant pollutant using chemometric techniques. The extraction of secondary physico-chemical water quality data by Department of Environment, Malaysia from 2003 to 2007 were taken. Principal Component Analysis (PCA) used to identify the significant water quality parameters and Absolute Principal Component Scores-Multiple Linear Regression (APCS-MLR) was applied to evaluate which pollution source contributed the most significant from the other factors. Then, Hierarchical Agglomerative Cluster Analysis (HACA) was used to find which monitoring station of Klang River provided the most significant pollutant. The result shows 9 PCs (75% of the total variance) which extracted from PCA. APCS-MLR model revealed anthropogenic activities which NH_3-N as the main parameter is the most significant in Klang River. Based on HACA specified 1K16 has the highest amount of NH_3-N concentration in Klang River. Therefore, the utilization of envirometric techniques achieved the research objectives which will beneficial for present and future river water quality management.

Keywords: *Chemometric; Water Quality; Principal Component Analysis; Absolute Principle Component Scores- Multiple Linear Regressions; Hierarchical Agglomerative Cluster Analysis.*

1. Introduction

Malaysian rivers and their tributaries are the main medium of transportation before the establishment of tarmac roads, railways and air transportation in Malaysia. In the past, until the advent of the 20th century, the main method of transportation, river banks and estuaries were usually the main meeting point and place for marker and business trade. From the river upstream (highlands and mountains), the river flow continuously especially during the wet monsoon bringing with them alluvial soils that enrich the lowlands and gives benefits in agricultural activities. Fishing, agricultural activities and barter trade were the main business occupation of the village folks in the beginning. Fruits, paddy and other crops were planted in the interior (helped by irrigation from the river waters), the crop yields were transported by waterways, using rafts and boats to sold and trade at the markets. Settlements then started and slowly expanded and grew from the river mouths and along the river banks. Village developed into towns, towns into cities and cities into metropolitans. According to an article [1], using these parameters, a set of guidelines is established. These guidelines may have arranged in which the water must fall, a maximum value of a pollutant, or a minimum value of a necessary substance (such as dissolved oxygen). The quality of a water is then determined based on whether it meets all of the requirements. Table 1 provides an example of chemical water quality criteria:

Table 1: Typical fresh surface water quality criteria for general use.

Parameters	Value
pH	5.0 - 9.0
Total Ammonia Nitrogen	< 9.0 mg/L
Dissolved Oxygen	> 5.0 mg/L
Fecal Coliform	< 200/100 mL
Orthophosphate	< 0.01 mg/L

The water quality index (WQI) has been considered to give criteria for surface water classification based on the use of standard parameters for water characterization [2]. The purpose of an index is not to describe separately a pollutant's concentration or the changes in a certain parameters. It was reported that WQI development is the biggest challenge in to synthesize a complex reality in a single number since it is directly affected by a large number of environmental variables [3]. The water quality index (WQI) was use as a basis for assessment of watercourse in relation to pollution load categorization and designation of classes of beneficial uses as stipulated in the National Quality Standard of Malaysia (NWQS). Therefore, the NWQS has applied to the surface waters as a guideline to the classification of the different state of the river water quality. For references, the standards of related parameters are given in Table 2. Description of the classes, in terms of the utility, is also given in this section [4].

Table 2: Excerpt of National Water Quality Standards (NWQS) for Malaysia.

Parameter	Unit	Limit of Classes				
		I	II	III	IV	V

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Ammonical Nitrogen (NH_3-N)	(mg/l)	< 0.1	0.1 - 0.3	0.3 - 0.9	0.9 - 2.7	> 2.7
BOD	(mg/l)	< 1	1 – 3	3 – 6	6 – 12	> 12
COD	(mg/l)	< 10	10 – 25	25 – 50	50 – 100	> 100
Dissolved Oxygen (DO)	(mg/l)	> 7	6.0 – 7	5.0 – 3	3 – 1	< 1
pH	–	7.0	7.0	6.0	< 5.0	< 5.0
Total Suspended Solids	(mg/l)	< 2.5	25 – 50	50 – 150	50 – 30	> 300

Notes:
CLASS I: Represent water bodies of excellent quality. Standards are set for the conservation of natural environment in its undisturbed state.
CLASS IIA: Represent water bodies of good quality. Most existing raw water supply sources come under this category. Class IIA standards are set for the protection of human health and sensitive aquatic species.
CLASS IIB: The determination of Class IIB standard is based on criteria for recreational use and protection of sensitive aquatic species.
CLASS III: Is defined with the primary objective of protecting common and moderately tolerant aquatic species of economic value. Water under this classification may be used for water supply with extensive/advanced treatment.
CLASS IV: Defines water quality required for major agricultural irrigation activities which may not cover minor applications to sensitive crops.
CLASS V: Represents other water, which does not meet any of the above uses.

Malaysian rivers can be classified as Class II/III Rivers. In this study, secondary data were taken into consideration to extract the spatio-temporal information of Klang River Basin. The secondary data used in this study were collected from monitoring stations under the river water quality monitoring program by the Department of Environment (DOE) from 2003 to 2007. Practically, 30 physico-chemical parameters were involved in this study. Significant Parameters which contributed to Potential Pollution Source from PCA were determined by developed Multiple Linear Regression model which is to evaluate the most significant pollution source of those rivers.

2. Methodology

2.1. Site description

Klang River Basin (Fig.1) is a fourth and the largest basin between the Langat River Basin, Selangor River Basin and Bernam River Basin. Alignment of Klang River is approximately 120 km of 80 km in the Selangor and 40 km in the Kuala Lumpur City Hall (DBKL) area. Estimated total area of Klang River Basins covers 1.290 square kilometers [5]. The basin is located in the west coast of Peninsular Malaysia between latitudes 2°55'N-3°25'N and longitudes 101°I5'E - 101°55'E. The river starts on the western slopes of the Main Range of Peninsular Malaysia, at an altitude of about 1200 meters and flows southwestwards before being joined by the Gombak River at the heart of Kuala Lumpur. Two main estuaries of the Kelang arc the Gombak and Batu rivers with basin areas of 260 km^2 and 145 km^2 respectively. Other large estuaries are the Keroh, Jinjang, Ampang, Damansara, Kuyoh and Kerayong [6]. Klang River which flowed in the middle of Klang Town separate to two parts – North Klang and Klang South. Klang River is the pulse of the development of Klang City which river mouth become as the main Klang Valley doorway to foreign traders. Thus early on Klang River is a catalyst and the role on the development gives high impact for the State of Selangor in the future entirely [5].
The Klang River Basin is the most important, urbanized, and most populous and most muted river basin in Malaysia. The basin of the river, known for its murky waters and floating debris is part of the Klang Valley Metropolitan area, which is the most important region of the country in terms of economy, the institution and population [6]. The 2010's Census shows the population of Malaysia was 28.3 million, compared with 23.3 million in 2000. This gives an average annual population growth rate of 2.0 percent for the period 2000-2010. This rate is lower than the 2.6 per cent for the period 1991-2000 [7]. The Klang River facing enormous threats from various sources, more than ten years, due to the different types of industries such as food and beverage, chemical, semiconductor and electronics industries, etc. In fact, the river flows through a densely populated area are often with point and non-point sources make it difficult to track the pollutants' loading in the river [8]. Today, the public realized that serious effort has to be done to improve in several rivers in the Klang Valley. In Ulu Klang, leafy trees along Klang River have been spared for slope stabilization works. Where Kerayong River flows through Cheras, a water treatment plant is under construction. In Selayang, the banks of Sungai Gombak have prettified. The main problems which are polluting Klang River are: effluent from sewerage treatment plants (80%), commercial and residential centers (9%), industries and workshops (5%) and in the food industry, wet markets and restaurants (4.2%) and squats other (1.8%) [9].

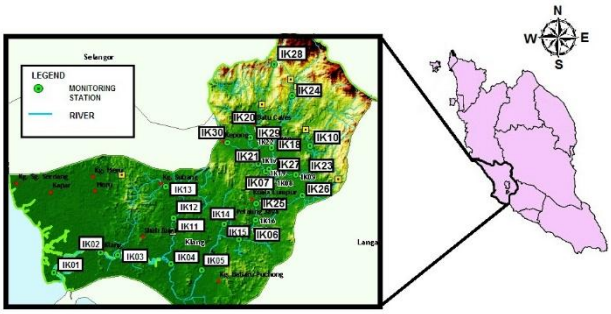


Fig. 1: Map of Klang River Basin. (Source: modified from Alam Sekitar Malaysia Sdn Bhd).

2.2. Chemometric Techniques

2.2.1. Pre-Processing Data

Preliminary work has been done in the data matrix that involved gathering and data transformation. The data that were lower the detection limit were unified with values equal to half the detection limit. Normal distribution tests were carried out with the support of the W (Shapiro-Wilk) test; the agreement of the distribution of the physico-chemical parameters of water with the normal distribution was tested [10, 11, 12]. Different variables with normal distribution were subjected to a transformation. The data pre-treatment is following the equation [13, 14] as cited in [15]:

$$Z_{ij} \, (X_{ij} - \mu) / \sigma \qquad (1)$$

Where Z_{ij} is the *j*th value of the standard score of the measured variable *I*; X_{ij} is the *j*th observation of variable i; μ is the variable's mean value; and σ is the standard deviation.
The raw water quality parameters were standardized through z-scale transformation to a mean of 0.0 and variance of 1.0 to certify that the dissimilar water quality parameters had equivalent weights in statistical analyses which can be found from Equation (1) [15]. The analysis results will be influenced most strongly by the variables having the highest magnitudes [14, 15, 16]. Furthermore, these transformations homogenize the variance of the distribution [13] and prevent any classification errors that may result from groups or classes described by variables of completely different sizes [15, 17].

2.2.2. Principal Component Analysis (PCA)

PCA has been designed to reduce the dimension of large data matrix to a lower dimension by retaining most of the original vari-

ability in the data [18, 19, 20, 21, 22]. This is achieved by converting a set of observational variables to a possible set of uncorrelated variables linearly called the principal components (PCs) [20, 21, 22]. The first principal component accounts for as much of the variation in the original data. Then, each succeeding component accounts for as much of the remaining variation subject to being uncorrelated with the previous component.

The covariance or correlation matrices obtained from data matrices play an important role in PCA to calculate the eigenvalues and eigenvectors to obtain relevant components that include most variations in data [23]. To achieve the objective in this study, correlation matrix was employed in the data set. At least 70% of cumulative percentage of total variation was recommended as a benchmark to cut off the eigenvalues in a large data set for extracting the number of components [22, 24, 25]. The reduced matrix is the component matrix of eigenvector loadings which defines the new variables consisting of linear transformation of the original variables that maximizes the variance in the new axes.

In this study, Pearson correlation was used in PCA for calculating eigenvectors and eigenvalues [26, 27]. Pearson correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations. In this case, Pearson correlation is used to measure the distance (or similarities) before implementing a clustering algorithm. The Pearson correlation coefficient between two vectors of observations is as follows:

$$r_{ij} = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_i)(X_j - \bar{X}_j)}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X}_i)^2 \sum_{i=1}^{n}(X_j - \bar{X}_j)^2}} \quad (2)$$

Where X_i and X_j refer to the vectors of observations in matrix data X with n observations, with $\bar{X}_i$ and $\bar{X}_j$ refer to the mean of the vectors.

The steps involved in PCA algorithm are as follows:

Step 1 : Obtain the input matrix.
Step 2 : Calculate the correlation matrix.
Step 3 : Calculate the eigenvectors and eigenvalues of the correlation matrix.
Step 4 : Select the most important principal components based on cumulative percentage of total variation.
Step 5 : Derive the new data set

This analysis is based on eigenvalue criteria by which a value >1 is deliberate significant, and a new group of variables was produced based on the similarity of the entire data set [28, 29]. Factor loading gives the correlation between the original variables and the varifactors (VFs), while the individual transformed observations are called factor scores [21, 22, 30]. The VF coefficients having a correlation 0.49–0.30 are considered 'weak' significant factor loadings, correlations in the range of 0.74–0.50 are considered 'moderate' and those in the range of >0.75 are considered 'strong' [29, 31, 32].

2.2.3. Absolute Principal Component Scores-Multiple Linear Regressions (APCS-MLR)

Quantitative contribution of the various identified sources was determined based on Multiple Linear Regression (MLR) of the Absolute Principal Component Scores (APCS) from PCA. In APCS-MLR, the predicted influence of each pollution source to the total concentration was determined using MLR with the denormalized APCS values produced by PCA as the independent variables and the measured concentrations of the particular pollutant [17, 21, 29, 33]. APCS–MLR model is based on the assumption that the total concentration of each contaminant is made up of the linear sum of the elemental contribution from each pollution component collected [21, 32]. Source contributions was calculated after grouping the water quality parameters for each basin in this study into number of factors and identify the possible sources by PCA. Therefore, in order to find the source of contribution, MLR is used to calculate sample mass concentration on the APCS [21, 34].

The contribution from each factor was estimated using MLR using APCS values as independent variables and measured water quality parameters as dependent variables. The quantitative contributions of each source of individual contaminants were compared with their measured values. A combination of PCA and MLR to form APCS-MLR is employed to find the source of apportionment using the equation as expressed in equation (3) [8]:

$$M_{n0} = d_{m0} + * \ D_{mn}(APCS)_{n0} \quad (3)$$

Where M_{n0} is the contaminant's concentration; d_{m0} is the average contribution of the nth contaminant from the sources not determined by PCA, D_{mn} is the linear regression contaminant and the bth factor and $(APCS)_{n0}$ is the absolute factor score for the bth factor with the nth measurement.

Coefficient of determination, R^2, Adjusted R^2 and Root Mean Square Error (RMSE) are the values that need to be considered in the best fitting regression linear equation [20]. R^2 is defined as the proportion of the variability in the dependent variable the fundamental measurement of the goodness of the fit of a linear model and is the fundamental measurement of the goodness of the fit of a linear model which is accounted for by the regression equation [20, 22, 35, 36]. The performance of the MLR model was assessed using correlation coefficient R^2, adjusted correlation coefficient R^2, Schwarz Bayesian Criteria (SBC) and Akaike's Information Criteria (AIC). The best model performance happens when the R^2, adjusted R^2, AIC and SBC values are close to unity. The minor difference in AIC and SBC indicate the MLR model fit for the prediction of possible pollution sources [1, 11, 32].

2.2.4. Hierarchical Agglomerative Cluster Analysis (HACA)

Hierarchical Agglomerative Cluster Analysis (HACA) is an independently pattern recognition technique that exposes constitutional structure or fundamental behavior of a data set deprived of creating an assumption about data, to categorize the items of the system into categories or clusters based on their closeness or similarity [21, 30]. HACA was accomplished on the normalized data set by means of the Ward's method, using Euclidean distances as a measure of similarity and usual have gathering items into groups so that items in a cluster are like each other while things located in other groups are dissimilarities with each other. Euclidean method is based on a single linkage (known as nearest neighbor) described as the ratio between the linkages distances divided by the maximal distance which can be seen in equation (4):

$$[(D_{link}/D_{max}) \times 100] \quad (4)$$

The linkage distance represented by the y-axis is standardized by multiply it with 100 [20, 37, 38, 39]. The result is showed by a dendrogram, representing the clusters and their proximity.

3. Results and Discussion

3.1. Principal Component Analysis (PCA)

PCA was primarily applied to the normalized datasets to compare between the compositional patterns of the analyzed water samples [8, 32, 40]. Besides, PCA was also utilized to identify the factor that influenced significantly to each of the variables [20]. Projections of the original variables on the subspace of the principal component (PCs) are called loadings and coincide with the alternation coefficients between the PCs and variable [22, 30, 40]. In the PCA, variables having loadings more than one are susceptible to the complex and difficult interpretation; therefore, it is advisable to conduct a varimax rotation that may ultimately group and

changes the factor loadings values spanning around 0 and 1 [41]. Furthermore, varimax rotation of the axis defined by PCA produced a new set of factors. Each one the factor involves primarily as a subset of the original variable and is divided into groups of variables that are independent towards each other [38]. In the PCA, the eigenvalue promotes as a measure of the factor loadings. Eigenvalues of one or greater are considered as significant [21, 22, 29]. In addition, the factor loading is classified as 'strong', 'moderate' and 'weak', corresponding to the absolute loading values of greater than 0.75, 0.75-0.50 and 0.50-0.30, respectively [8, 21, 22, 29, 30, 39, 42]. In this study, the concentrations of 30 water quality parameters for each of the river basin are used as variables for the PCA.

The results revealed that pollution pattern the resemblance in geographical structure and surrounding factors that contributed to the variation of variable concentrations. The PCA stipulated nine PCs with eigenvalues greater than one explaining 75 % of the total variance in the water-quality are described in Table 3. The results show that factor 1 (VF1) exhibit 29.9 % out of the total variance with strong positive loadings posed by count, Sal, DS, TS, Cl, Ca, K, Mg and Na. The direct relationship between these parameters coincides with the study that has been conducted by [30, 41]. Therefore, the VF1 are corresponded as a source of the mineral component that are overlying in the river water.

Table 3: Loadings of nine Varimax Factors (VFs) of Klang River Basin.

	VF1	VF2	VF3	VF4	VF5	VF6	VF7	VF8	VF9
DO	-0.091	-0.155	-0.113	-0.041	0.161	-0.652	-0.229	0.036	-0.162
BOD	-0.064	**0.830**	-0.016	-0.025	-0.034	0.171	0.180	0.056	-0.048
COD	-0.006	**0.828**	0.174	0.033	-0.049	0.233	0.092	0.054	-0.009
SS	0.000	0.139	**0.945**	0.009	0.002	-0.034	-0.022	-0.011	0.002
pH	-0.080	-0.023	-0.039	-0.001	-0.027	-0.113	-0.495	0.233	-0.341
NH_3-N	-0.058	0.127	-0.089	-0.026	-0.122	0.696	-0.145	0.209	-0.097
TEMP	0.084	0.133	0.011	0.112	0.007	0.709	-0.028	-0.016	0.002
COND	**0.992**	-0.018	-0.006	-0.013	0.005	0.000	0.002	-0.009	0.006
SAL	**0.993**	-0.018	-0.006	-0.013	0.004	-0.005	0.002	-0.009	0.006
TUR	-0.016	-0.012	**0.944**	-0.021	0.018	0.027	0.058	-0.012	-0.015
DS	**0.994**	-0.018	-0.010	-0.014	-0.001	0.001	0.004	-0.008	0.005
TS	**0.989**	-0.002	0.099	-0.013	-0.001	-0.002	0.002	-0.009	0.005
NO_3	0.054	-0.105	0.045	-0.032	**0.842**	-0.173	-0.025	-0.006	-0.022
Cl	**0.977**	-0.025	-0.007	-0.016	-0.025	-0.015	0.007	-0.017	0.018
PO_4	-0.023	0.242	-0.140	0.076	0.005	0.424	-0.027	0.619	-0.041
As	0.022	0.008	0.020	0.099	0.023	0.000	-0.006	0.716	-0.019
Cd	0.011	-0.019	-0.014	-0.064	0.014	0.165	-0.112	-0.144	0.549
Cr	0.096	0.358	-0.124	0.178	0.049	0.115	-0.199	-0.411	-0.284
Pb	0.003	0.010	-0.037	0.017	0.022	-0.058	0.035	0.053	0.706
Zn	0.032	0.188	0.060	-0.127	0.200	0.103	0.604	-0.174	-0.263
Ca	**0.921**	0.072	-0.049	0.008	0.023	0.109	-0.028	-0.012	-0.035
Fe	-0.090	0.096	0.021	-0.029	-0.100	-0.092	0.738	0.150	0.057
K	**0.970**	0.010	-0.026	-0.017	0.010	0.027	-0.012	0.018	-0.003
Mg	**0.978**	-0.019	-0.022	-0.018	0.016	-0.009	-0.017	0.005	-0.002
Na	**0.989**	-0.019	-0.006	-0.013	0.022	-0.010	-0.001	-0.005	0.003
OG	0.001	0.677	0.090	0.107	-0.053	-0.298	-0.205	-0.044	0.110
MBAS	-0.016	0.006	-0.016	-0.010	**0.866**	0.035	0.027	0.012	0.032
E-coli	-0.037	0.002	0.000	**0.911**	-0.010	0.051	0.037	-0.007	-0.010
Coliform	-0.051	0.050	-0.012	**0.922**	-0.025	0.007	-0.089	0.065	0.003
Eigenvalues	8.6852	2.873	2.1032	1.7673	1.52	1.385	1.27	1.088	1.039
Variability (%)	29.906	7.457	6.560	6.138	5.389	6.511	4.764	4.329	3.878
Cumulative %	29.906	37.363	43.923	50.061	55.450	61.960	66.724	71.053	74.931

Note: Values in bold indicate the variables having strong loading >0.75 and value underline indicate the moderate loading.

Meanwhile, factor 2 (VF2) signifies strong positive loadings of BOD and COD with 7.5 % of the total variance. This factor represents as the anthropogenic pollution sources. Theoretically, the BOD and COD have a strong significant correlation with one another and this is confirmed by the Pearson correlation that portrays $r= 0.745$ at $p < 0.001$. The strong significant correlation was due to the high levels of dissolved organic matter that will inevitably consume large amounts of oxygen that undergoes anaerobic fermentation processes, hence leading to the increase of COD values, BOD value and reducing the levels of dissolved oxygen required in the water column [30]. Whereas, VF3 indicates 6.56 % of the total variance have strong positive loadings of SS and turbidity ($r= 0.819$, $p = < 0.001$). The direct relationship between both parameters has been reported in [11]. Thus, this confirmed that suspended solids and turbidity does correspond to one another. Therefore, this factor is categorized as the surface runoff that comes from the field, transporting a high load of soil and waste disposal amenities into the river [11].

VF4 shows 6.14 % of the total variance with strong positive loadings of *E-coli* and coliform. Which is consistent with other study [44]. The result suggests that both parameters are suspected to be originated from the animal faecal, surface runoffs and sewage treatment plant effluent discharge. On the contrary, the VF5 accounted for 5.4 % of the total variance having strong positive loading of MBAS and NO_3. The alliances between these variables are suspected to be derived from the municipal waste discharge such as detergent [45]. The surfactants that are utilized in the detergents are also called as methylene-blue-active substance (MBAS) due to the colour changes that can be observed in aqueous solution of the methylene blue dye [46]. Generally, Klang River flows through the densely populated area in the urban city of Kuala Lumpur. Along the way the Klang River transported different kinds of particulates into the river, hence resulting in the variety of non-point pollution sources into the river. Therefore, this study confirmed that the association of these variables in VF5 is embodied into the river water due the municipal waste discharges.

VF6 exhibited to 6.5 % of the total variance with strong positive loadings of NH_3-N and temperature. The results suggest that it is potentially due to the high loading of dissolved organic matter or other liquid organic waste deposited into the river. This organic waste leads to the anaerobic conditions, hence resulted in the formation of ammonia and organic acids that cause the reduction of pH [40]. Moreover, the pH may decrease once the temperature increase. Therefore, the VF6 are corresponded as the liquid organic waste, a type of pollution sources available in the Klang River

Basin. VF7 accounted for 4.8 % of the total variance with strong positive loadings of Fe. This result is consistent with [47] and [8] which suggested that the presence of Fe in the river water are basically derived from the industrial effluent that held near the river. Moreover, Fe was also enlisted as one of the elements in the metal group. According to [48], soluble Fe can be found in groundwater, dead-ends in water distribution systems, oxygen-free reservoirs and scale (hard mineral coatings) within pipes. VF8 described 4.3 % of the total variance with the strong positive loading of As. This VF is referred as industrial application waste from industrial areas. Many chemicals including heavy metals, entered the river and accumulate in environmental compartments [48, 49]. Based on [48] research on Klang River, As is discharged into the environment by the several sources like the fusion process of Cu, Zn and Pb, likewise by the industrialisation of chemicals and glasses. Other sources are paints, rat poisoning, fungicides, and wood preservatives.

VF9 accounted for 3.9 % of the total variance with strong positive loadings of Cr. This factor is explained due to the chromes, dyes and leather industry waste that have contributed to the production and generation of Cr in the river. According to [48], the most relevant source of Cr in Klang River is from solid waste. Solid waste from processing plants chromate, when are improperly disposed of in landfills can be a source of contamination of the groundwater, the residence time of Cr may be several years. It is believed the leakage from topsoil and rocks is the most essential natural source of Cr access into water bodies [48]. The formation of Cr (VI) from Cr (III) happens, notably in the presence of common minerals containing Mn (IV) oxides. The summarization of all the possible pollutant sources is displayed in Table 4.

Table 4: Summary of Possible Pollution Sources for Klang River Basin.

Parameter	Possible Pollution Sources
VF1 : Cond, Sal, DS, TS, Cl, Ca, K, Mg, Na	Mineral Component
VF2 : BOD, COD	Industrial Waste/ Anthropogenic
VF3 : SS, TUR	Surface Runoff
VF4 : *E-coli*, Coliform	Sewage
VF5 : MBAS, NO_3	Municipal Waste (Detergent)
VF6 : NH_3-N, Temp	Liquid Organic Waste Product
VF7 : Fe	Abundant of Fe Element in Earth Crust
VF8 : As	Pesticide/Industrial Application Waste
VF9 : Cr	Chromes/Dyes/Leather Industry Waste

3.2. APCS-MLR for Source Apportionment

Source apportionment is a very crucial approach in environmetric. This approach assists in the identification of sources that contributed to each of the parameter concentrations [34]. Although, the PCA provides qualitative information concerning the source of pollution, it was still found inadequate to support the contribution of each type of the sources [43]. Therefore, to acquire significant sets of result, MLR was employed in an attempt to determine the percentage of contribution for each of the sources. This approach calculated the weight of sources in the total sum using multiple regressions. The outcomes were evaluated based on the R^2 values where if the R^2 is greater than 0.75, it indicated as a good fit between the measured and predicted concentration. Meanwhile, the goodness of the receptor modeling approach (PCA-MLR) to the source apportionment of the water variables is determined based on the excellent correlation computed [43].

This study revealed that the coefficient of determination (R^2) for the APCS-MLR model in Klang River indicates that it is relatively accurate with R^2 equivalent to 0.724 as this is displayed in Table 5. Although the R^2 values demonstrate a slightly different value in comparison to Wu et al. (2009) [43], the goodness of the receptor modeling approach (PCA-MLR) are still managed to illustrate a good correlation with the source apportionment of the water quality parameters in the Klang River.

Table 5: Goodness of fit statistic for regression of WQI.

Statistics	Value
R^2	0.724
Adjusted R^2	0.718
MSE	8.413
RMSE	2.900
AIC	829.812
SBC	869.344

Based on Table 6, VF6 has been identified as the major contributor to the NH_3-N and temperature with the percentage of 66.35%. The result suggests that the elevated amount of industrial activities near the Klang River will lead to the introduction of high pollution sources in the river. Therefore, the high loading of NH_3-N will literally result to the fluctuation in the water temperature. Hence, the contribution of these pollutants suggests that the river must undergo continuous monitoring of the levels of pollutants to control and protect the river water from being continuously polluted due to the human negligence.

Table 6: Goodness of fit statistic for regression of WQI.

Variable	R^2	Diff R^2	MSE	RMSE	% contribution
All Source	0.724		8.413	2.900	
L-VF1	0.663	0.061	10.254	3.202	8.354
L-VF2	0.717	0.007	8.606	2.934	0.966
L-VF3	0.688	0.037	9.503	3.083	4.986
L-VF4	0.698	0.027	9.201	3.033	3.637
L-VF5	0.721	0.003	8.494	2.914	0.464
L-VF6	**0.238**	**0.486**	**23.191**	**4.816**	**66.346**
L-VF7	0.663	0.061	10.248	3.201	8.326
L-VF8	0.724	0.000	8.403	2.899	0.057
L-VF9	0.674	0.050	9.922	3.150	6.866
Total for Diff R^2		0.733			100

3.3. Hierarchical Agglomerative Clustering Analysis for Significant Parameters

HACA was incorporated on the water quality data set to assess the spatial variations of the significant water quality parameter. HACA is used as a grouping and classifying tool based on the parameter similarity level. These prior studies have denoted the importance of HACA in the water quality assessment [11, 38, 40]. The clustering procedure first reflects the individual items as separate groups. Then, pairwise distances between groups are calculated and the pair with the minimum distance is linked to form new clusters [20, 50]. The dendrogram is used to illustrate the classification of the objects which display the similarity levels then quantified through Ward's method and Euclidean distance measurement. In fact, HACA has been widely used in the water quality assessment to identify one station in each cluster which can act as a reasonably accurate representation of the spatial assessment despite reducing the cost of monitoring and analysis [11]. The parameter that has been chosen for HACA analysis is based on the percentage of the source apportionment in the APCS-MLR.

The secondary data from the Klang river basin were subjected into HACA analysis to classify the monitoring stations based on the concentration of NH_3-N. The results revealed that the HACA managed to categorize the level of NH_3-N concentrations into two clusters. The concentrations of NH_3-N at each of the monitoring stations are categorized into two distinct groups in Fig. 2. The dendrogram classified IK16 as Cluster 1 shows to the highest level of NH_3-N concentration in the river. Meanwhile, the Cluster 2 grouped all of the other monitoring stations in Klang River as low level of NH_3-N concentration. These findings suggest that station IKI6 may have been induced with high level of NH_3-N concentration due to anthropogenic activities that occurred surrounding the area. According to [52], sedimentation is a major pollutant in Klang River, which high concentration of SS because of the development of substantially due to population growth. Besides, disposal of sewage and industrial waste also contribute to pollute

the Klang River. Furthermore, Klang River is regarded as main waterway that flows through the capital city of Kuala Lumpur which is known as a densely populated and urban industrialized area. Based on [53] research in Kerayong River, IK16 area is exceedingly urbanize with 74 % imperviousness. The residential area forms the largest fraction of the impervious surface, covering 50 % of the total catchment. The daily wash offs did not brush up the catchment due to the remaining pollutants washed off from an internal drain, which could be related with the sludge from sewerage treatment plant activities in the catchment [53]. The river was badly polluted even though the rainfall intensity was high. Theoretically, the improper sewerage treatment effluent and waste product that consist of sewage, liquid manure and other liquid organic waste product discharges into the running water course will lead to the increase of NH_3-N concentration in the river water. Moreover, Klang River basin is categorized as Class III and Class IV according to the Environmental Quality Report [54]. This indicates that the river is facing a current threat due to the one station (IK16) that emits high concentration of NH_3-N into the river. Therefore, the urge to monitor this station constantly needs to be implemented in attempt to reduce the amount of pollutants despite positioning IKI6 as a representative station in the Klang River Basin.

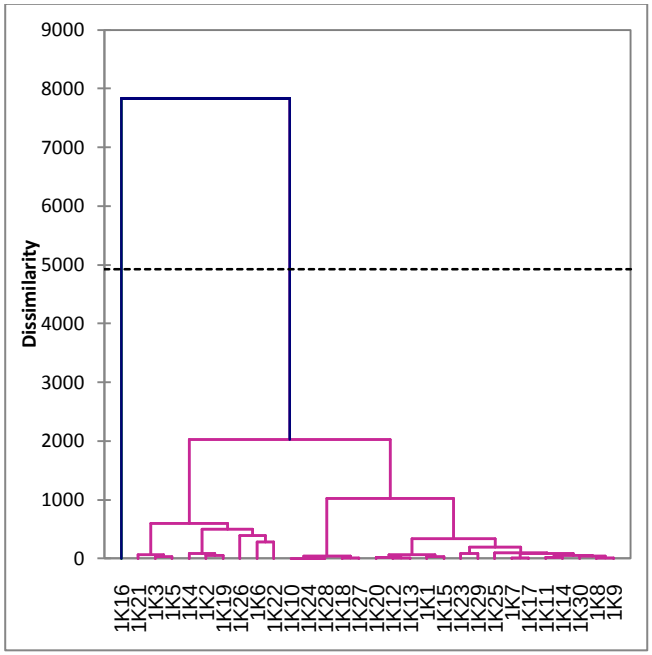


Fig. 2: NH_3-N dendrogram showing classified sampling sites located in the Klang River Basin.

4. Conclusion

PCA in Klang River Basin involved 9 PCs eigenvalues greater than one explaining about 75 % of the total variance in the data set of water quality. The result from PCA suggested mineral component, Industrial waste/ anthropogenic, surface runoff, sewage, municipal waste (detergent), liquid organic waste product, abundant Fe element in the earth's crust, pesticide/industrial application waste, chromes/dyes/leather industry waste as possible pollution sources in the Klang River Basin. NH_3-N has been revealed as the main parameters for polluting the Klang River Basin by using APCS-MLR model. Monitoring stations from Klang River Basin have been clustered by two groups which are high concentration of NH_3-N and low concentration of NH_3-N. HACA specifies 1K16 has the highest concentration of NH_3-N. SPC comes out with the result signifies that the mean concentration of NH_3-N in 1K16 were not in the control process. A potential clarification of this problem might be that the Klang River Basin is exposed to the anthropogenic activities in the highly industrialized area.

Acknowledgement

The authors acknowledged the assistance and support provided by the Department of Environment (DOE), Malaysia during the study conducted.

References

[1] Aertsen, W., Kint, V., Orshoven, J.V., Ozkan, K. & Muys, B. (2010) Comparison and ranking of different modelling techniques for prediction of site index in Mediterranean mountain forests. *Ecological Modelling* 221, 1119–1130.

[2] Hernández-Romero, A. H., Tovilla-Hernández, C., Malo, E.A. & Bello-Mendoza, R. (2004) Water quality and presence of pesticides in a tropical coastal wetland in southern Mexico. *Marine Pollution Bulletin* 48, 1130–1141.

[3] Lermontov, A., Yokoyama, Li'dia., Lermontov, M. & Machado, M. A. S. (2009) River quality analysis using fuzzy water quality index: Ribeira do Iguape river watershed, Brazil. *Ecological Indicators* 9, 1188–1197.

[4] Mamun, A. A. & Zainudin, Z. (2013) Sustainable River Water Quality Management in Malaysia. *IIUM Engineering Journal* 14(1), 29-42.

[5] Rahman, M. A. A. (2010) Laporan Awal Pemuliharaan dan Pembangunan Sungai Klang.Pengenalan Lembangan Sungai Klang; Selangor Town and Country Planning Development; http://www.jpbdselangor.gov.my; 27 Mac 2010.

[6] Balamurugan, G. (1991) Sediment Balance and Delivery In a Humid Tropical Urban River Basin: The Klang River, Malaysia. *Catena Verlag* 18, 271-287.

[7] DOS (Department of Statistic) (2010). Kompendium Perangkaan Alam Sekitar, Malaysia 2010; Portal Rasmi Jabatan Perangkaan Malaysia; www.dosm.gov.my; 23 December 2010.

[8] Nasir, M. F. M., M.S. Samsudin., I. Mohamad., M.R.A. Awaluddin., M.A. Mansor., H. Juahir. & N. Ramli. (2011) River Water Quality Modeling Using Combined Principle Component Analysis (PCA) and Multiple Linear Regressions (MLR): A Case Study at Klang River, Malaysia. *World Applied Sciences Journal* 14, 73-82.

[9] Balan, A. (2012) Saving Sungai Klang. Klang, Selangor: The Malaysian Times, The Malaysian Times Sdn Bhd.

[10] Sojka, M., M. Siepak., A. Zioła., M. Frankowski., S. Murat-Błażejewska. & J. Siepak (2008) Application of multivariate statistical techniques to evaluation of water quality in the Mała Wełna River (Western Poland). *Environ Monit Assess* 147, 159–170.

[11] Juahir, H., Retnam, A., Zali, M. A. & Hashim, M. F. (2011) *A comparison between multiple linear regression (MLR) and artificial neural network (ANN) for river class prediction at Klang river, Malaysia*. Contemporary Environmental Quality Management in Malaysia and Selected Countries. Universiti Putra Malaysia Press, Serdang.

[12] Samsudin, M. S., Juahir, H., Zain, S. M. & Adnan, N. H. (2011) Surface river water quality interpretation using environmetric techniques: Case study at Perlis River Basin, Malaysia. *International Journal of Environmental Protection* 1(5), 1-8.

[13] Güler, C., Thyne, G.D., McCray, J.E. & Turner, A.K. (2002). Evaluation of graphical and multivariate statistical methods for classification of water chemistry data. *Hydrogeology Journal* 10 (4), 455–474.

[14] Kowalkowski, T., Zbytniewski, R., Szpejna, J. & Buszewski, B., (2006) Application of chemometrics in river water classification. *Water Research* 40(4), 744–752.

[15] Gazzaz, N. M., Yusoff, M. K., Ramli, M. F., Aris, A. Z. & Juahir, H. (2012) Characterization of spatial patterns in river water quality using chemometric pattern recognition techniques. *Marine Pollution Bulletin* 64 (4), 688-698

[16] Reimann, C., Filzmoser, P. & Garrett, R.G. (2002) Factor Analysis Applied To Regional Geochemical Data: Problems And Possibilities. *Applied Geochemistry* 17(3), 185–206.

[17] Simeonov, V., Einax, J.W., Stanimirova, I. & Kraft, J., 2002. (2002) Environmetric modeling and interpretation of river water monitoring data. *Analytical and Bioanalytical Chemistry* 374 (5), 898–905.

[18] Shaharudin, S.M., Ahmad, N. & Zainuddin, N.H. (2013) Improved Cluster Partition in Principal Component Analysis Guided Clustering. *International Journal of Computer Applications* 75(11), 1162-1167.

[19] Everitt, B. S. & Dunn, G. (2001) *Applied Multivariate Data Analysis*. London: Arnold Publisher.

[20] Dominick, D., Juahir, H., Latif, M. T., Zain, S. M. & Aris, A. Z (2012) Spatial assessment of air quality patterns in Malaysia using multivariate analysis. *Atmospheric Environment* 60, 172-181.

[21] Samsudin, M. S., Azid, A., Khalit, S. I., Saudi, A. S. M. & Zaudi, M. A. (2017) River water quality assessment using APCS-MLR and statistical process control in Johor River Basin, Malaysia. *International Journal of Advanced and Applied Sciences* 4(8), 84-97.

[22] Azid, A., Juahir, H., Toriman, M. E., Kamarudin, M. K. A., Saudi, A. S. M., Hasnam, C. N. C., Aziz, N. A. A., Azaman, F., Latiff, M. T., Zainuddin, S. F. M., Osman, M. R. & Yamim, M. (2014) Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: a case study in Malaysia. *Water, Air, & Soil Pollution* 225(8), 2063.

[23] Neware, S., Mehta, K. & Zadgaonkar, A. S. (2013) Finger knuckle identification using principal component analysis and nearest mean classifier. *International Journal of Computer Applications 70*(9), 18-23.

[24] Jolliffe, I. (2011) Principal component analysis. In *International encyclopedia of statistical science* (pp. 1094-1096). Springer, Berlin, Heidelberg.

[25] Shaharudin, S.M., Ahmad, N. & Zainuddin, N.H. (2018) Identification of Rainfall Patterns on Hydrological Simulation using Robust Principal Component Analysis. *Indonesian Journal of Electrical Engineering and Computer Science* 11(3), 22-25.

[26] Peñarrocha, D., Estrela, M. J. & Millán, M. (2002) Classification of daily rainfall patterns in a Mediterranean area with extreme intensity levels: the Valencia region. *International Journal of Climatology 22*(6), 677-695.

[27] Wickramagamage, P. (2010) Seasonality and spatial pattern of rainfall of Sri Lanka: Exploratory factor analysis. *International Journal of Climatology* 30(8), 1235-1245.

[28] Osman, R., Saim, N., Juahir, H. & Abdullah, Md. (2009) Chemometric application in identifying sources of organic contaminants in Langat river basin. *Environmental Monitoring and Assessment*, 1-14.

[29] Samsudin, M. S., Khalit, S. I., Azid, A., Juahir, H., Saudi, A. S. M., Sharip, Z. & Zaudi, M. A. (2017) Control limit detection for source apportionment in Perlis River Basin, Malaysia. *Malaysian Journal of Fundamental and Applied Sciences* 13(3), 294-303.

[30] Vega, M., Pardo, R., Barrado, E. & Deban, L. (1998) Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. *Water Research* 32, 3581-3592.

[31] Liu, C. W., Lin, K. H. & Kuo, Y. M. (2003) Application of factor analysis in the assessment of groundwater quality in a Blackfoot disease area in Taiwan. *Science in the Total Environment* 313, 77-89.

[32] Retnam, A., Zakaria, M. P., Juahir, H., Aris, A. Z., Zali, M. A. & Kasim, M. F. (2013) Chemometric Techniques In Distribution, Characterisation And Source Apportionment Of Polycyclic Aromatic Hydrocarbons (PAHS) In Aquaculture Sediments In Malaysia. *Marine Pollution Bulletin* 69, 55–66.

[33] Feng Zhou, Gordon, H. H., Huaicheng, Guo, Wei Zhang & Zejia Hao. (2007) Spatio-temporal patterns and source apportionment of coastal water pollution in eastern Hong Kong. *Water Research* 41, 3429 – 3439.

[34] Simeonova, V., Stratisb, J. A., Samarac, C., Zachariadisb, G., Voutsac, D., Anthemidis, A., Sofonioub, M. & Kouimtzisc, T. H. (2003) Assessment of the surface water quality in Northern Greece. *Water Research* 37, 4119–4124.

[35] Allen, P. D. & Richard, H. M. (2005*) Stormwater Management for Smart Growth*.United States of America: Springer Science + Business Media, Inc.

[36] Ilten, N. & Selici, T. (2008) Investigation the impacts of some meteorological parameters on air pollution in Balikesir, Turkey. *Environment Monitoring Assessment* 140, 259-266.

[37] Singh, K. P., Malik, A., Mohan, D. & Sinha, S. (2004) Multivariate statistical techniques for the evaluation of *spatial and temporal variations in water quality of Gomti River (India)—a case study*. *Water Research* 38, 3980–3992.

[38] Singh, K. P., Malik, A. & Sinha, S. (2005) Water quality assessment and apportionment of pollution sources of Gomti river (India) using multivariate statistical techniques: a case study. *Analytica Chimica Acta* 538, 355-374.

[39] Juahir, H., Zain, M. S, Aris, A. Z., Yusoff, M. K. & Mokhtar, M. (2010) Spatial assessment of Langat River water quality using chemometrics. *Journal of Environmental Monitoring* 12, 287-295

[40] Shrestha, S. & Kazama.F. (2007) Assessment Of Surface Water Quality Using Multivariate Statistical Techniques: A Case Study Of The Fuji River Basin, Japan. *Environmental Modelling & Software* 22, 464-475.

[41] Helena, B., Pardo, R., Vega, M., Barrado, E., Ferna´ndez, J. M. & Ferna´ndez, L. (2000) Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga river, Spain) by principal component analysis. *Water Research* 34, 807-816.

[42] Juahir, H., Zain, M. S., Yusoff, M. K., Ismail, T. T. H., Samah, A. M. A., Toriman, M. E. & Mokhtar, M. (2010) Spatial water quality assessment of Langat River Basin (Malaysia) using environmetric techniques. *Environ Monit Assess* 173, 625–641.

[43] Wu, B., Zhao, D., Zhang, Y., Zhang, X. & Cheng, S. (2009) Multivariate statistical study of organic pollutants in Nanjing reaches of Yangtze River. *Journal of hazardous materials* 169(1), 1093-1098.

[44] Chigor, V.N., Umoh, J.V., Smith, I.S., Igbinosa, O.E. & Okoh, I.A. (2010) Multidrug Resistance and Plasmid Patterns of Escherichia coli O157 and Other E. coli Isolated from Diarrhoeal Stools and Surface Waters from Some Selected Sources in Zaria, Nigeria. *International Journal of Environmental Research and Public Health* 7, 3831-3841.

[45] Wyrwas, B. & Zgoła-Grześkowiak, A. (2013) Continuous Flow Methylene Blue Active Substances Method for the Determination of Anionic Surfactants in River Water and Biodegradation Test Samples. *Journal of Surfactants and Detergents* 17(1), 1-8.

[46] Ghose, N. C., Saha, D. & Gupta, A. (2009) Synthetic Detergents (Surfactants) and Organochlorine Pesticide Signatures in Surface Water and Groundwater of Greater Kolkata, India. *Journal of Water Resource and Protection*. 4, 290-298.

[47] Juahir, H., Ekhwan, T. M., Zain, S. M., Mokhtar, M., Zaihan, J. & Ijan Khushaida, M. J. (2008) The use of chemometrics analysis as a cost-effective tool in sustainable utilisation of water resources in the Langat River Catchment. *American-Eurasian Journal of Agricultural & Environmental Sciences* 4(1), 258-265.

[48] Ismail, Z., Salim, K., Othman, S. Z., Ramli, A. H., Shirazi, S. M., Karim, R. & Khoo, S. Y. (2013) Determining and comparing the levels of heavy metal concentrations in two selected urban river water. *Measurement* 46, 4135–4144.

[49] Karcher, S., Caceres, L., Jekel, M. & Contreras, R. J. (1999) Arsenic removal from water supplies in Northern Chile using ferric chloride coagulation. *Water and Environment Journal* 13, 164 –169.

[50] Hoang, T. H., Bang, S., Kim, K. W., Nguyen, M. H. & Dang, D. M. (2010) Arsenic in groundwater and sediment in the Mekong River delta, Vietnam. *Environmental Pollution* 158, 2648-2658.

[51] Pires, J. C. M., Sousa, S.I.V., Pereira, M.C., Alvim-Ferraz, M.C.M. & Martins, F.G. (2008) Management of air quality monitoring using principal component and cluster .Analysise part I: SO2 and PM10. *Atmospheric Environment* 42, 1249-1260.

[52] Mokhtar, M., Bahari, I., Hoh, Y. C. & Poon, A. (2001) Kajian kualiti air di sekitar Kawasan Perindustrian Subang Jaya dan Shah Alam, Lembah Kelang.*Malaysian Journal of Analytical Sciences* 7(1), 139-149.

[53] Dom, N. M., Abustan, I. & Abdullah, R. (2012) Dissolved Organic Carbon Production and Runoff Quality of Sungai Kerayong, Kuala Lumpur, Malaysia. *International Journal of Engineering & Technology* 12(4), 44-47.

[54] Department of Environment, M. (DOE) (2017) Malaysia Environmental Quality Report 2016. Putrajaya, Malaysia. p. 1-135.