World Applied Sciences Journal
14 (Exploring Pathways to Sustainable Living in Malaysia: Solving the Current Environmental Issues): 73-82, 2011
ISSN 1818-4952; © IDOSI Publications, 2011

River Water Quality Modeling Using Combined Principle Component Analysis (PCA) and Multiple Linear Regressions (MLR): A Case Study at Klang River, Malaysia

[1]*Mohd Fahmi Mohd Nasir,* [1]*Mohd Saiful Samsudin,* [1]*Isahak Mohamad,*
[1]*Mohammad Roshide Amir Awaluddin,* [1]*Muhd Ariffin Mansor,* [1]*Hafizan Juahir,* [2]*Norlafifah Ramli*

[1]Department of Environmental Sciences, Faculty of Environmental Studies,
43400 UPM Serdang, Selangor, Malaysia
[2]Department of Environment Malaysia, Surface Water Monitoring Unit,
Water and Marine Division, Federal Government Administrative Centre, 62574 Putrajaya, Malaysia

Abstract: A collective set of data over five years (2003 to 2007) in Klang River, Selangor were studied in attempt to assess and determine the contributions of sources affecting the water quality. A precise technique of multiple linear regressions (MLR) were prepare as an advance tool for surface water modeling and forecasting. Likewise, principle component analysis (PCA) was used to simplify and understand the complex relationship among water quality parameters. Nine principle components were found responsible for the data structure provisionally named as soil erosion, anthropogenic input, surface runoff, fecal waste, detergent, urban domestic waste, industrial effluent, fertilizer waste and residential waste explains 72% of the total variance for all the data sets. Meanwhile, urban domestic pollution accounted as the highest pollution contributor to the Klang River. Thus, the advancement of receptor model was applied in order to identify the major sources of pollutant at Klang River. Result showed that the use of PCA as inputs improved the MLR model prediction by reducing their complexity and eliminating data collinearity where R^2 value in this study is 0.75 and the model indicates that 75% variability of WQI explained by the five independent variables used in the model. This assessment presents the importance and advantages poses by multivariate statistical analysis of large and complex databases in order to get improved information about the water quality and then helps to reduce the sampling time and cost for reagent used prior to analyses.

Key words: Water quality · Multiple linear regression · Principle component analysis

INTRODUCTION

The vast industrialization, uncontrolled urbanization and rapid economic development around many cities especially Kuala Lumpur have increase the levels of pollution to the environment. Klang River particularly suffers a lot since it flows through the state of Kuala Lumpur and Selangor. The pollution mainly comes from municipal infrastructure which raises numerous environmental concerns along the Klang River meanwhile intensive human activities have resulted in substantial hydrological deformation.

According to Tong and Chen (2002), land use types correlate with most water quality caharcteristic. Runoff from land surface carries the residues from the land into river system which known as non-point sources pollution [1]. Expansion of urban areas in any river basin effect the environment in terms of the increase of pollution load into river system and changes to surface water quality. Nevertheless, substantial modification on flood runoff and water quality found to be contributed by urban development [2]. Costa *et al*., (2003), found that the conversion of vegetation wil disrupt the hydrological cycle of a drainage basin by altering the balance between rainfall and evaporation of the area [3]. Urban rivers are also polluted with discharge from sewage treatment plants, overflowing sewage causes by rainfall [4] causing fecal contamination which is a major concern in the river near the town area [5] where the surface water are used by local residents. Nonetheless, industrial and household

Corresponding Author: Hafizan Juahir, Department of Environmental Sciences, Faculty of Environmental Studies,
43400 UPM Serdang, Selangor, Malaysia. Tel: +03-89467460,
Fax: +03-89467463.

waste which are discharge directly or through leakages in the sewage systems will flow into water sources thus causing excessive pollution of surface and underground water [6].

Monitoring programs with frequent water samplings and determination of physiochemical parameters may representatively provide the status of the surface water quality. Since 1978, Department of Environment (DOE) Malaysia has performed monitoring action resulting large data matrices and requires advance statistical tools such as multivariate and artificial intelligence for exceptional data illustration. Initially, the program covered all the river basin in Malaysia, involving mainly manual sampling and in-situ measurements of the river water samples. According to the DOE's 2007 Environmental Quality Report, 158 river basins in Malaysia involved in this program in order to monitor river water quality changes on a continuous basis [7]. Even though DOE have a regular monitoring program to provide the complex environmental data sets, however there is still lack of application in multivariate statistical methods in attempt to extract all possible information from the river water quality data sets thus unable to determine the major source influence the river class in Klang River. Therefore, the multivariate statistical technique and exploratory data analysis are the appropriate tools for an outstanding data reduction and interpretation of multi-constituent chemical and physical measurement.

Generally, water quality refers to the characteristics of water whether physical, chemical or biological characteristic. Based on the water quality data, the water quality index (WQI) was developed to evaluate the water quality status and river classification in Malaysia. WQI provides prediction in changes and trends in the water quality by considering multiple parameters. WQI is formed by six selected water quality variables namely as Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Suspended Solids (SS), Ammonical Nitrogen (AN) and pH [7]. WQI values ranges from 0 to100. The values in the range of 81to100 are considered as clean. Whereas, the value ranges from 60 to80 and 0-59 are classified as slightly polluted and polluted area respectively.

Hence, continuous monitoring of river water quality will reveals the chemical and physiochemical characteristic. Although, there are not often convincing for the interpretation of large data set with many variables; envirometric approach are still required to comprehend the variation on the data. In this paper, the large data matrix obtained from monitoring programme by DOE, Malaysia from year 2003 to 2007 were introduced to receptor models techniques involves varimax factor from principal components analysis (PCA) with numerical models multiple linear regressions (MLR). Furthermore, advance mathematical tools can be used as a medium to warned people as well as related environmental agencies to protect and sustain the river from being more polluted.

This study attempts to predict WQI values using MLR model from the varimax factors generated by PCA. On the hand, implement absolute understanding on how good the model is and delineate or narrow down the best model for WQI prediction in the Klang River.

MATERIAL AND MATHODS

Study Area: Klang River are one of the most important rivers in Malaysia where it flows directly through the major cities such as Kuala Lumpur, Shah Alam, Petaling Jaya and Klang. It has approximately about 120 km length which covers a total catchment area of 1288 km^2. The Klang river faced huge threats from various sources over ten years ago due to various types of industrial activities such as food and beverages, chemical manufacturing, semi conductor and electrical and etc. Indeed, a river flows through a heavily populated area are commonly associated with point and non point sources therefore it is difficult to trace the loading of pollutants in the river. Therefore, calls for research in accordance with advance mathematical tools are highly emboldened.

Data Set: In this study, thirty water quality parameters are observed along the Klang River monitoring stations. The selected stations were determined based on the data reported from 2003 to 2007, however in the raw data some stations are missing and some data are incomplete because not all readings are consistently taken due to technical failure by measuring instruments.

A total of 1105 observations were used for source apportionment and modeling techniques. The water quality index (WQI) was developed to evaluate the water quality status and river class classification based on water quality data. The thirty water quality parameters consists of dissolve oxygen (DO), biological oxygen demand (BOD), chemical oxygen demand (COD), suspended solid (SS), pH, ammoniacal nitrogen (NH_3NL), dissolved solid (DS), total solid (TS), nitrate (NO_3), chloride (Cl), phosphate (PO_4), *Escherichia coli* (*E. coli*), coliform and also several heavy metals parameters.. WQI provides a useful way to predict changes and trends in the water quality by considering multiple parameters. Hence, it is formed by six selected water quality variables namely as dissolved oxygen (DO), BOD, chemical oxygen demand (COD), SS, AN and pH [1].

Pretreatment Data Set: Data were initially arranged according to the stations and year of monitoring. Any particular variable that has not been detected (below detection limit), the value are normally set to half and no missing data was ensure in the overall data sets. Normality test were perform using XLSTAT2010 software based on Anderson-Darling test. Several data that are not normally distributed were pretreated which is a combination of centering, standardization and log-scaling method. Standardization opts was implemented to increase the influence of variables whose variance is small and vice versa [8]. Log scaling is very common in environmental data since some of the variables might exhibit too low or high values [9]. Then, statistical computation of PCA and MLR were conducted using XLSTAT2010 software.

Principle Component Analysis (PCA): Principal component analysis illustrates the most significant parameters, which describe whole data set rendering data reduction with minimum loss of original information [10-12]. PCA are sensitive to outliers, missing data and poor linear correlation between variables due to inadequate assigned variables [13]. Therefore, a detailed pretreatment data set needs to perform in order to get clearer image in complex data. It is a prominent technique for pattern recognition in attempts to explain the variance of a large set of inter-correlated variables and transforming into a smaller set of independent (uncorrelated) variables (principal components). The principal component (PC) is expressed as:

$$y_{mn} = z_{m1}x_{1n} + z_{m2}x_{2n} + z_{m3}x_{3n} + \ldots + z_{mi}x_{in} \qquad (1)$$

Where z is the component loading, y is the component score, x is the measured value of a variable, m is the component number, n is the sample number and m is the total number of variables. Meanwhile, factor analysis (FA) attempts to extract a lower dimensional linear structure from the data set. It further reduces the contribution of less significant variables obtained from PCA and the new group of variables known as varifactors (VFs) which is extracted through rotating the axis defined by PCA. In FA, the basic concept is expressed in Eq. (2),

$$y_{nm} = z_{p1}p_{1m} + z_{p2}p_{2m} + z_{p3}p_{3m} + \ldots + z_{pr}p_{rm} + c_{pm} \qquad (2)$$

Where y is the measured value of a variable, m refer to the factor loading, p is the factor score, c is the residual term accounting for errors or other sources of variation, m is the sample number, n is the variable number and r is the total number of factors. Principally, the two methods which is PCA and FA are expressed in a similar equations however, the difference are in PC where it is expressed as a linear combination of measured variables. Whereas FA, measured variable is expressed as a combination of factors and the equation contains the residual term and thus, a VF can include unobservable, hypothetical, latent variables [10-12, 14]. Principal component analysis or factor analysis was performed on correlation matrix of rearranged data (all observations for each group of sites), thus explains the structure of the underlying data set. The correlation coefficient matrix measures how well the variance of each constituent can be explained based on the relationship with each others. PCA of the normalized variables (water quality data set) was performed to extract significant PCs and to further reduce the contribution of variables with minor significance; these PCs were subjected to varimax rotation (raw) generating VFs.

The PCs resulted by PCA are sometimes not readily interpreted and varimax rotation need to perform to f reduce the dimensionality of the data and identify most significant new variables. Varimax factor (VF) coefficient with a correlation of >0.75 are explained as strong significant factor loading [15]. While correlation ranges from 0.75-0.50 and 0.50-0.30 are considered as moderate and weak factor loading respectively. In a nutshell, principle component analysis aims to uncover a more underlying set of factors that accounts for the major pattern across all the original variables [16]. Therefore, principal component often present information on the most meaningful reliable parameters, which define the whole data set affording data reduction with minimum loss of original information [8].

Absolute Principle Component Scores-multiple Linear Regression (APCS-MLR): Receptor modeling application based on APCS-MLR is a commonly apply in statistical technique for source apportionment of environmental contaminants in air pollution studies [17, 18]. It has freshly been employed to water pollution source apportionment worldwide. It is based on the assumption that the total concentration of each contaminant is made up of the linear sum of elemental contributions from each of the pollution source components collected at the receptor site,

$$Z_{no} = \Sigma\, Q_{mn} R_{no} \qquad (3)$$

Where Z_{no} is the normalized concentration of contaminant (variable), $_j$ is the number of pollution sources, Q_{mn} is the factor loadings, the coefficient matrix of the components relating the pollution sources to their elemental concentrations; and R_{no} is the scores. Since, Z_{no} in Eq. (3) is normalized value of variables, it cannot be used directly for computation of quantitative source contributions, the normalized factor scores determined in Eq. (3) were converted to unnormalized APCS [19]. The contribution from each factor was then estimated by MLR, then using the APCS values as the independent variables and the measured concentration of the particular contaminant as the dependent variable, as shows in

$$M_{no} = d_{m0} + \Sigma\, D_{mn}\,(APCS)_{no} \qquad (4)$$

Where M_{no} is the contaminant's concentration; d_{m0} is the average contribution of the nth contaminant from sources not determined by PCA/FA, D_{mn} is the linear regression coefficient for the m^{th} contaminant and the b^{th} factor and $(APCS)_{no}$ is the absolute factor score for the b^{th} factor with the n^{th} measurement. The values for M_{no}, d_{m0} and D_{mn} have the dimensions of the original concentration measurements. After determination of the number and identity of possible sources infiuenced the river water quality by PCA/FA, source contributions were computed through APCS-MLR technique. Quantitative contributions from each source for individual parameter or contaminant were compared with their measured values.

RESULTS AND DISCUSSION

Principle Component Scores (PCA): PCA was applied to the normalized data to compare the compositional patterns between the analyzed water samples and to identify the factors that influenced each samples. Rotation of the axis defined by PCA produced a new set of factors, involving primarily a subset of the original variables are divided into groups [20]. An eigenvalue illustrates the most significant factors the highest eigenvalues are the most significant. Eigenvalues of 1.0 or greater are considered significant [21].

PCA of the entire data set allowed forming nine PCs with nine eigenvalues greater than 1 explaining that 72% of the total cumulative in the water-quality data set as shown in Table 1. Projections of the original variables on the subspace of the PCs are called loadings and coincide with the alternative coefficients between PCs and variable. The principal component analysis showed that the eigenvalues of the two main principal components up to 36.03% of the total variance (PC1 28.78%; PC2 7.24%) for total observations. Considering the larger variability graph for factor loading 1 and factor loading 2 were plotted to explain the variance.

PCA was applied to the data set to compare the compositional pattern between the analyzed water samples and to identify the factor that reflecting each other [22]. PCA was performed on the raw data set comprising all the 30 water quality parameters with 1105 observation to identify the factors that contribute to pollution sources in Klang River Basin.

The nine Varimax Factors (VFs) were achieved after varimax rotation based on eigenvalues greater than 0.1 [21] as shown in the Table 2. Eigenvalues and the corresponding factors were sorted by descending order and the initial variability was representing in percentage. VF1 (eigenvalue 8.646) represent 28.78% of the total variability in one axis (VF1) which has strong positive loadings on Conductivity (COND), Salinity (SAL), Dissolved Solid (DS), Total Solid (TS), Chlorine (Cl), Calcium (Ca), Potassium (K) and Sodium (Na). VF1 can be interpreted as a mineral component of the river water. Vega *et al*. (1998) stated that this clustering variables points to a common origin for these minerals, likely from dissolution of limestone and gypsum soils which can be simplified as soil erosion [23].

Table 1: Eigenvalues from principal component analysis shows variability and cumulative.

Principal	Eigenvalues:								
Component Analysis:	F1	F2	F3	F4	F5	F6	F7	F8	F9
Eigenvalue	8.646	2.854	2.095	1.761	1.527	1.385	1.283	1.092	1.075
Variability (%)	28.819	9.513	6.982	5.870	5.091	4.617	4.277	3.641	3.583
Cumulative%	28.819	38.332	45.314	51.185	56.276	60.894	65.170	68.811	72.394

Table 2: Factor loadings after Varimax rotation

Parameter	Soil erossion	Anthropogenic input	Surface runoff	Fecal waste	Detergent	Urban domestic pollution	Industrial effluent	Fertilizer waste	Residential waste
DO	-0.095	-0.180	-0.104	-0.042	0.159	-0.617	0.338	0.019	-0.008
BOD	-0.065	0.838	-0.013	-0.031	-0.027	0.168	-0.088	0.061	0.093
COD	-0.005	0.830	0.169	0.034	-0.048	0.228	-0.050	0.046	0.041
SS	0.000	0.135	0.943	0.014	-0.002	-0.039	0.016	-0.014	-0.022
pH	-0.080	-0.052	-0.027	0.003	-0.039	-0.074	0.644	0.226	-0.025
NH3-NL	-0.057	0.130	-0.080	-0.020	-0.129	0.700	0.196	0.192	-0.020
TEMP	0.085	0.138	0.012	0.117	-0.002	0.705	0.027	-0.023	0.014
COND	0.992	-0.018	-0.006	-0.013	0.004	-0.001	-0.008	-0.008	0.003
SAL	0.993	-0.018	-0.006	-0.012	0.004	-0.006	-0.007	-0.008	0.003
TUR	-0.017	-0.011	0.946	-0.022	0.017	0.026	-0.042	-0.010	0.029
DS	0.994	-0.018	-0.010	-0.014	-0.001	0.000	-0.009	-0.007	0.003
TS	0.989	-0.002	0.099	-0.013	-0.002	-0.004	-0.007	-0.009	0.000
NO_3	0.053	-0.107	0.043	-0.030	0.841	-0.170	0.047	-0.008	0.011
Cl	0.962	-0.024	-0.009	-0.017	-0.029	-0.020	-0.035	-0.015	-0.007
PO_4	-0.024	0.244	-0.142	0.076	0.002	0.429	0.042	0.607	-0.060
As	0.020	-0.013	0.020	0.107	0.011	0.013	-0.011	0.713	0.005
Hg	0.013	-0.056	-0.051	0.056	-0.089	-0.086	-0.064	-0.001	0.643
Cd	0.013	-0.061	-0.024	-0.031	0.002	0.211	-0.129	-0.285	-0.307
Cr	0.095	0.331	-0.108	0.191	0.012	0.101	0.355	-0.355	0.275
Pb	0.003	-0.022	-0.051	0.042	0.033	0.002	-0.335	-0.119	-0.403
Zn	0.016	0.204	0.094	-0.159	0.210	0.127	-0.221	-0.136	0.579
Ca	0.921	0.070	-0.048	0.010	0.018	0.104	0.040	-0.002	0.029
Fe	-0.072	0.099	0.018	-0.066	-0.058	-0.104	-0.680	0.171	0.140
K	0.970	0.010	-0.026	-0.017	0.011	0.028	0.012	0.016	-0.004
Mg	0.978	-0.020	-0.022	-0.017	0.017	-0.008	0.017	0.003	-0.006
Na	0.987	-0.019	-0.006	-0.012	0.023	-0.009	0.002	-0.006	0.003
OG	0.004	0.666	0.076	0.127	-0.058	-0.318	0.106	-0.061	-0.186
MBAS	-0.014	0.009	-0.021	-0.005	0.863	0.033	-0.025	0.011	0.008
E. coli	-0.039	0.007	0.004	0.903	-0.002	0.056	-0.026	-0.004	0.020
Coliform	-0.051	0.045	-0.012	0.923	-0.026	0.005	0.060	0.068	-0.038

Table 3: Summary of regression of variable WQI

Regression of variable WQI: Goodness of fit statistics:	
Observations	1103.000
Sum of weights	1103.000
DF	1093.000
R^2	0.749
Adjusted R^2	0.747
MSE	48.875
RMSE	6.991
AIC	4299.812
SBC	4349.870

VF2 explain BOD, COD and Oil and Grease in new variable with strong factor loadings represent the anthropogenic input typically organic pollution. Thus, VF clearly explained that high levels of dissolved organic matter and biological organic matter comes from runoff of solids or waste disposal activities [24]. Meanwhile for VF3 (6.982% of variance) was firmly correlated with SS and Turbidity representing the surface runoff sources. This factor loaded with solids indicates that the surface runoff originated from the fields contains high load of solids and waste disposal sources [24].

Whereas, VF4 (5.870% of variance) shows strong positive loading of *E-coli* and coliform. The VF represents fecal waste according to its factor. On contrary, VF5 (5.091% of variance) shows strong positive loading of NO_3 and MBAS which can represent the abundance of detergent deposited into the Klang River. According to Berna *et al*. (1991) detergent contains nutrients such as nitrogen can provoke algae gloom. Meanwhile in aquatic environment, surfactants may form a surface film and reduce transfer at the water surface. Moreover, some surfactants may be toxic to aquatic organisms[25].

VF6 (5.091% of variance) shows strong positive loading of NH_3-NL and Temperature and moderately negative loading of DO which may be consider as urban domestic source. This is due to the fact that urban domestic waste might be discharges into the river such as waste from sewage treatment plant, poultry farms and surface run off from urban area [26]. Similarly to VF7 (4.277% of variance), it shows strong positive loading of pH and moderately negative loading of iron (Fe) which may represent the input from industrial effluent. According to Juahir *et al*. (2008), Fe are basically one of the element in the metal group, therefore Fe in Klang River might be originated from industrial effluents[27]. Since Klang river basin is a very vast industrialization area therefore the introduction of iron in the river seems unsurprising.

VF8 (3.641% of variance) has strong positive loading of arsenic (As) and moderately positive loading of PO_3. The VF may indicate the waste from fertilizer which is commonly used in urban landscaping. Therefore, PO_3 may come from the fertilizer used in agricultural and horticultural activities moreover a wide range of fertilizer might also come from urban landscaping in the state of Kuala Lumpur itself. Meanwhile, the presence of arsenic in Klang River might be due to the fact that some fertilizer may contains arsenic [27]. VF9 (3.583% of variance) has moderate positive loading of mercury (Hg) and zinc (Zn) which is represent residential waste [28].

Source Apportioning by Absolute Principal Component Scores (APCS): PCA aims to exclude redundant information from the original raw data set by obtaining a small number of variables that make it more comprehensible and for furthering the analysis such as modeling of the data set. Generally, source apportioning was well studied in many environmental areas such as air pollution and water quality study. Nevertheless, source apportioning study integrated with water quality index was less documented in tropical regions. In air pollution studies, PCA and any environmetric techniques have been used extensively to determine possible natural and anthropogenic contributions in the formation of the determinants total mass and concentration [29]. For receptor modeling or source apportioning the computation of APCS for each observation of interest is required. In this study, factor scores from PCA after varimax rotation were used in receptor models development using MLR and ANN. Both models were further compared to evaluate the performance on the data set. The use of PC based models was considered more dynamic, due to elimination of collinearity problems and prediction improvement [30]. Moreover the utility of APCS that contain minimum input for both model compared to the raw data set was beneficial since it will increase the computational efficiency and interpretability and reduce the noise and redundancy for the model.

APCS-MLR model: It is possible to develop a MLR model to describe the behavior of the old variables in terms of the new variables. Basically MLR is based on a linear least-squares fitting process and required a trace element or property to be determined for each source or source category [31]. PCA and MLR were combined to identify potential pollution sources to the Klang River. Two basic types of receptor models that generally applied for source apportionment are chemical mass balance (CMB) and multivariate techniques [32]. It is also noted that that factor analysis (FA) identifies tracers that represent specific sources and the sources are selected as input (independent variables) to predict dependent variables [33]. MLR was practiced as well in this study to explain the relationship between the source apportionment generated by PC and their correlation to WQI values. In order to prove it, MLR was applied to search the relationship of each source to the dependant variable (WQI) with 5VFs as independent variables for the MLR model.

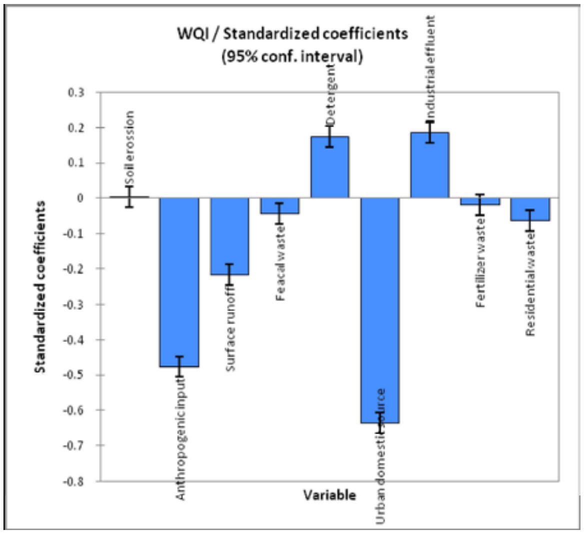

Fig. 1: Standardized coefficients for each variable

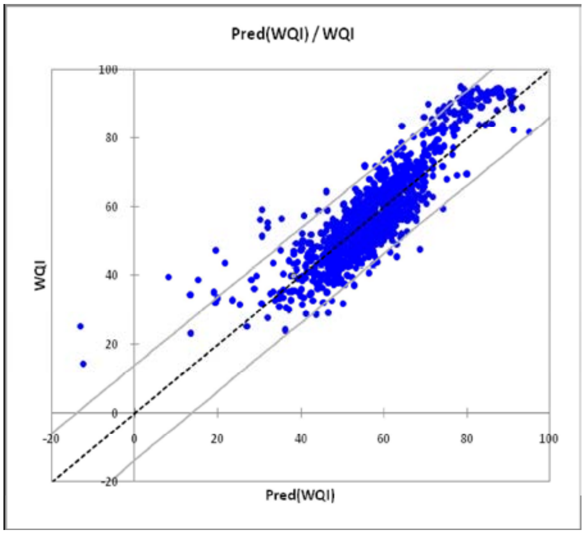

Fig. 2: Water Quality Index (predicted) versus Water Quality Index (actual)

The source apportionment is a vital environmetric techniques aiming to the estimation of contribution of identified sources to the concentrations of each parameter [34]. After determining the number and characteristics of possible sources by PCA with varimax rotation, source contributions were then calculated with APCS-MLR to identify main pollution origin in Klang River. The most commonly used criterion to evaluate model performance is coefficient of determination (R^2) [35]; however R^2 is not a good comparison measurement of different model since R^2 values only provide how good the model fits the data used to build the models and not how well it performs on external data [36]. The R^2 value for APCS-MLR model in this study is 0.75 (Table 2) and the model indicates that 75% variability of WQI explained by the five in dependent variables used in the model. While for adjusted R^2 that also adjust for the number of explanatory terms used in the model and always be less than R^2 and increases only if the new term improve the model [36]. Mean Square Error (MSE) and Root Mean Square Error (RMSE) measure residual error which give estimation of the mean difference between observed and modeled values of WQI. The minimum value of MSE for APCS-MLR result (Table 2) correspond to best network topology [37].

Best model performance are with Akaike's Information Criteria (AIC) and Schwarz Bayesian Criteria (SBC) values and R^2 and adjusted R^2 values closes to unity [36]. In general AIC, Bayesian Information Criteria (BIC) and SBC estimate the loss of accuracy caused by accounting a number of parameters and the number of data points used in its calibration. The small difference for AIC and SBC values signify that MLR was fit method for WQI prediction. The high and great difference between values of AIC and SBC from the APCS-MLR model in this study (Table 2) indicate the model has inadequacy in term of fitness and robustness.

Figure 1 shows the standardized coefficients of independent variable of the WQI linear regression model. Urban domestic pollution account as the highest pollution contributor to Klang River while for the next main contributor was anthropogenic input that may come from the vicinity area of Klang River. The negative standardized coefficient of independent variables (Anthropogenic input, surface runoff, fecal waste, urban domestic source, fertilizer waste and residential waste) owing to negatively correlation to WQI values (as all the four independent variable decrease, WQI value increase). Insert Figure 1.

Figure 2 shows the graph plotting for calculated WQI and predicted WQI. It is known that 26 observations from overall observations were out of the upper and lower boundary range (95% mean of the confidence interval). This proved that this model able to predict WQI values from the varimax factor of PCA with negligible precision. This is due to great difference between calculated WQI and predicted WQI for some of the observations from the training and testing set.

Figure 3 illustrated the residual analysis of the observed and predicted WQI using APCS-MLR model. The results show the deficiency of the APCS-MLR

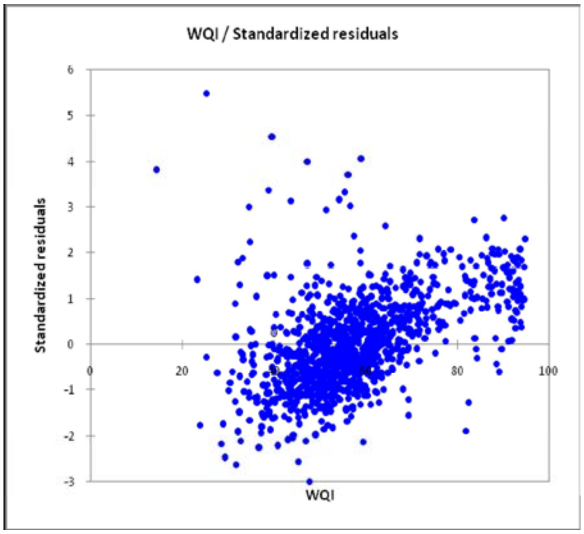

Fig. 3: Standardized residuals between Actual WQI and predicted WQI

model as the data set show great difference in the range of -3 to 6. The verification and applicability of the model was influence by the existence of the outlier observations as shown also in Figure 2.

CONCLUSIONS

Urban domestic waste was the majority of pollutant sources that contributed to changes of the water quality in Klang River, Malaysia and followed by anthropogenic activities. Two water quality parameters are identified as an input that influence the DO concentration NH3-NL and temperature which is accounted as strong pollution loading. This might be due to the discharges from sewage treatment plants, poultry farms and surface run off into the river. Moreover, this also leads to natural geological changes to the studied areas. The results above shows that APCS-MLR model gives good accuracy for WQI forecasting where R^2 value is 0.75 (Table 2) and the model indicates that 75% variability of WQI explained by the five independent variables used in the model. MLR employed to water pollution source apportionment worldwide because able to stimulate the complex relationship between the data set and consequently able to verify the complicated water quality parameters. Moreover without eliminating any data and parameters, main pollutant contributors to the basin were justified by using PCA. The applications of PCA in this model are better rather than using the original data, because PCA reduced the number of inputs and decreased the model complexity. Considering MLR and PCA can provides good performances index and more efficient due to elimination of collinearity problems and reduction of the number of predictor variables. In fact, the use of PCA based on MLR advanced the forecasting of WQI. Thus, proved that APCS-MLR model are a useful tool for DOE or others public agency in order to conduct a more efficient environmental monitoring in Malaysia. Hence, this advance model also will help to reduce the sampling campaign and cost of reagent used in the analyses.

ACKNOWLEDGEMENTS

The authors are grateful to Department of Environment, Malaysia for kindly providing the data used in this study and several colleagues who help in reading and sharing ideas for this journal.

REFERENCES

1. Tong, S.T.Y. and W. Chen, 2002. Modeling the relationship between land use and surface water quality. J. Environ. Manage, 66: 377-393.
2. Tong, S.T.Y., 1990. The hydrologic effects of urban land use: a case study of the Little Miami River Basin. Landscape Urban Plan, 19: 99-105.
3. Costa, M.H., A. Botta and J.A. Cardille, 2003. Effects of large-scale changes in land cover on the discharge of the Tocantins River, South eastern Amazonia. J. Hydrol., 283: 206-217.
4. Nix, P.G. and C.H. Merry, 1990. Use of sediment bags as a monitor of fecal pollution in streams. Bulletine Environmental Contamination Toxicol., 45: 864-869.
5. Miyabara, Y., J. Suzuki and S. Suzuki, 1994. Classification of urban rivers on the basis of water pollution indicators in river sediment. Bulletine Environmental Contamination Toxicol., 52: 1-8.
6. Akcay, H., A. Oguz and C. Karapire, 2003. Study of heavy metal pollution and speciation in Buyak Menderez andGediz river sediments. Water Research, 37: 813-822.
7. DOE, 1997. Environmental Annual Report 1997, Kuantan. Department of Environment Pahang, Ministry of Natural Resources and Environment.
8. Krishna, A.K., M. Satyanarayanan and P.K. Govil, 2009. Assessment of heavy metal pollution in water using multivariate statistical techniques in an industrial area: A case study from Patancheru, Medak District andhra Pradesh,India. J. Hazardous Materials, 167: 366-373.

9. Felipe-Sotelo, M., J.M. Andrade, A. Carlosena and R. Tauler, 2007. Temporal characterisation of river waters in urban and semi-urban areas using physico-chemical parameters and chemometric methods. Analytica Chimica Acta, 583: 128-137.
10. Helena, B., R. Pardo, M. Vega, E. Barrado, J.M. Fernandez and J. Fernandez, 2000. Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga River, Spain) by principle component analysis Wat. Res., 34(3): 807-816.
11. Vega, M., R. Pardo, E. Barrado and D. Luis, 1988. Assessment of Seasonal and Polluting Effects On The Quality of River Water by Explanatory Data Analysis. Water Res., 32(12): 3581-3592.
12. Wunderlin, D.A., M.P. Diaz, M.V. Ame, S.F. Pesce, A.C. Hued and M.A. Bistoni, 2001. Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality. A case study: Suquia river basin (Cordoba, Argentina). Water Res., 35: 2881-2894.
13. Sarbu, C. and H.F. Pop, 2005. Principal component analysis versus fuzzy principal component analysis a case study: the quality of danube water (1985–1996) Talanta, 65: 1215-1220.
14. Singh, K.P., A. Malik and S. Sinha, 2005. Water quality assessment and apportionment of pollution sources of Gomti River (India) using multivariate statistical techniques: a case study. Analytica Chimica Acta, 35: 3581-3592.
15. Liu, C.W., K.H. Lin and Y.M. Kuo, 2003. Application of factor analysis in the assessment of groundwater quality in a black foot disease area in Taiwan. The Science of the Total Environ., 313: 77-89.
16. Saima, S., R. Osman, D.R.S.A. Spian, M.Z. Jaafar, H. Juahir, M.P. Abdullah and F.A. Ghani, 2009. Chemometric approach to validating faecal sterols as source tracer for faecal contamination in water. Water Res., 43: 5023-5030.
17. Simeonova, V., J.A. Stratis, C. Samara, G. Zachariadis, D. Voutsac, A. Anthemidis, M. Sofoniou and T. Kouimtzis, 2003. Assessment of the surface water quality in Northern Greece. Water Res., 37: 4119-4124.
18. Swlethcki, E. and R. Krejci, 1996. Source characterisation of the Central European atmospheric aerosol using multivariate statistical methods Nuclear Instruments and Methods in Physics Res., 109/110: 519-525.
19. Thurston, G.D. and G.D. Spengler, 1985. A quantitave assessment of sourcecontributins to inhalable particulate matter pollution in metropolitan Boston. Atmospheric Environ., 19(1): 9-15.
20. Singh, K.P., A. Malika, D. Mohana and S. Sinha, 2004. Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India)—a case study. Water Res., 38: 3980-3992.
21. Kim, J.O. and C.W. Mueller, 1987. Introduction to factor analysis: what it is and how to do it. Quantitative Applications in the Social Sciences Series. Sage University Press, Newbury Park.
22. Singh, K.P., A. Malik, D. Mohan, S. Sinha and V.K. Singh, 2004. Chemometric data analysis of pollutants in wastewater – a case study. Analytica Chimica Acta, 32: 15-25.
23. Vega, M., R. Pardo, E. Barrado and D. Luis, 1998. Assessment of Seasonal and Polluting Effects On The Quality of River Water by Explanatory Data Analysis. Water Res., 32(12): 3581-3592.
24. Yeung, I.M.H., 1999. Multivariate analysis of the Hong Kong Victoria Harbour water quality data. Environmental Monitoring and Assessment, 60(3-4): 365-380.
25. Berna, J.L., A. Moreno and J. Ferrer, 1991. The behaviour of LAS in the environment. J. Chem. Technol. Biotechnol., 50: 387-398.
26. Geiser, L.H., A.R. Ingersoll, A. Bytnerowicz and S.A. Copeland, 2008. Evidence of Enhanced Atmospheric Ammoniacal Nitrogenin Hells Canyon National Recreation Area: Implications for Natural and Cultural Resources. Air & Waste Management Association, 58: 1223-1234.
27. Juahir, H., S.M. Zain, M.K. Yusoff, T.I.T. Hanidza, A.S.M. Armi, M.E. Toriman and M. Mokhtar, 2008. Spatial water quality assessment of Langat River Basin (Malaysia) using environmetric techniques. Environ Monit Assess, DOI 10.1007/s10661-010-1411-x.
28. Zhou, F., Y. Liu and H. Guo, 2007. Application of Multivariate Statistical Methods to Water Quality Assessment of the Water courses in North western New Territories, Hong Kong. Environ Monit Assess, 132: 1-13.
29. Randolph, K., I. Larsen and J.E. Baker, 2003. Source Apportionment of Polycyclic Aromatic Hydrocarbons in the Urban Atmosphere: A comparison between three methods. Environ. Sci. Technol., 37: 1873-1881.
30. Sousa, S.I.V., F.G. Martins, M.C.M. Alvim-Ferraz and M.C. Pereira, 2007. Assembly of the Rodinia Supercontinent: Evidence from the Sakoli and Sausar Belts in Central India Environmental Modelling& Software, 22: 97-103.

31. Henry, R.C., C.W. Lewis, P.K. Hopke and H.J. Williamson, 1984. Review of receptor model fundamentals. Atmospheric Environ., 18(8): 1507-1515.
32. Gordon, G.E., 1988. Receptor models. Environ. Sci. Technol., 22: 1132-1142.
33. Morandi, M.T., J.M. Daisey and P.J. Lioy, 1987. Development of a modified factor analysis/multiple regression model to apportion suspended particulate matter a complex urban airshedin. Atmospheric Environment, 21(8): 1821-1831.
34. Simeonov, V., J.A. Stratis, C. Samara, G. Zachariadis, D. Voutsa, A. Anthemidis, M. Sofoniou and T. Kouimtzis, 2003. Assessment of the surface water quality in Northern Greece. Water Rese., 37: 4119-4224.
35. Pearson, K., 1986. Regression, heredity, panmixia. In Mathematical contributions to the theory of evolution. Philos. Trans. R. Soc.Lond, 187: 253-318.
36. Aertsen, W., V. Kinta, J. Orshovena, K. Özkan and B. Muysa, 2010. Comparison and ranking of differentmodelling techniques for prediction of site index in Mediterranean mountain forests EcologicalModelling, 221: 1119-1130.
37. Sousa, S.I.V., F.G. Martins, M.C.M. Alvim-Ferraz and M.C. Pereira, 2007. Multiple linear regression and artificial neural networks based on principal components to predict ozone concentration. Environmental Modelling and Software, 22: 97-103.