Journal of Environment and Earth Science
ISSN 2224-3216 (Paper) ISSN 2225-0948 (Online)
Vol 2, No.2, 2012

Application of Principal Component Analysis & Multiple Regression Models in Surface Water Quality Assessment

Adamu Mustapha[1*] Ado Abdu[2]

1. Department of Environmental Science, Faculty of Environmental Studies, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia.
2. Department of Resources Management & Consumer Studies, Faculty of Human Ecology, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia.

*E-mail of the corresponding author: amustapha494@gmail.com

Abstract

Principal component analysis (PCA) and multiple linear regressions were applied on the surface water quality data with the aim of identifying the pollution sources and their contribution toward water quality variation. Surface water samples were collected from four different sampling points along Jakara River. Fifteen physico-chemical water quality parameters were selected for analysis: dissolved oxygen (DO), biochemical oxygen demand (BOD_5), chemical oxygen demand (COD), suspended solids (SS), pH, conductivity, salinity, temperature, nitrogen in the form of ammonia (NH_3), turbidity, dissolved solids (DS), total solids (TS), nitrates (NO_3), chloride (Cl) and phosphates (PO_4^{3-}). PCA was used to investigate the origin of each water quality parameters and yielded five varimax factors with 83.1% total variance and in addition PCA identified five latent pollution sources namely: ionic, erosion, domestic, dilution effect and agricultural run-off. Multiple linear regressions identified the contribution of each variable with significant value (r 0.970, R^2 0.942, $p < 0.01$).

Keywords: River, Stepwise regression, Varimax factor, Varimax rotation, Water pollution

1. Introduction

With the growth of human populations, commercial and industrial activities, surface water has received large amount of pollutants from variety of sources (Satheeshkumar, and Anisa, 2011). The quality of surface water provides significant information about the available resources for supporting life in the ecosystem (Manikannan et al. 2011). The physical, chemical and biological compositions of surface water is controlled by many factors such as natural (precipitation, geology of the watershed, climate and topography) and anthropogenic (domestic, industrial activities and agricultural run-off). Increasing surface water pollution causes not only deterioration of water quality, but also threatens human health, balance of aquatic ecosystem, economic development and social prosperity (Milovanovic, 2007). It is imperative to prevent and control the surface water pollution and to have reliable information on its quality for effective management (Sing et al. 2005). Characterization of the spatial variation and source apportionment of water

quality parameters can provide an improved understanding of the environmental condition and help policy makers to establish priorities for sustainable water management (Huang et al. 2010). One of the major challenges in surface water quality assessment is identifying the sources of pollutants and the contribution of the parameters/variables in explaining water quality variation. An ever increasing literature on the use of principal component analysis (PCA) in identifying pollution sources and multiple linear regressions in estimating the contribution of parameters/variables suggest that the techniques are useful in revealing the latent pollution sources and it is practical in various types of data (Praveena et al. 2011). PCA provides information on the most meaningful variables that bring surface water quality variation and allowed the identification of a reduced number of latent factors/sources of pollution while multiple linear regressions examine the relationship between single depended variables and a set of independed variables to best represent relationship in a population.

Several researchers used PCA to identify water quality sources apportionment. For example: Shrestha and Kazama, (2007); Huang et al. (2010); and Juahir et al. (2011) studied spatial variability of surface water quality and sources apportionment and classified the studied water bodies into High pollution site (HP), Moderate pollution site (MP) and Low pollution site (LP). PCA revealed that the pollution levels in the three zones were mainly influenced by natural sources (temperature and river discharge) and anthropogenic sources (industrial, municipal and agricultural run-off). Onojake et al. (2011) in their studies, they discovered that Rivers in Delta State of Nigeria were heavily polluted as a result of industrial discharge and municipal waste (anthropogenic source of pollution). They used PCA to identify the latent factors that explain the chemistry of the surface water in which PCA yielded three PC's with more than 82% variance. Equally, Hai et al. (2009) studied Taihu lake region in China and discovered that, the surface water in the region is progressively susceptible to anthropogenic pollution, three PC's yielded correspond to urban residential subsistence, livestock farming and farmlands run-off. Similarly, recent study conducted by Koklu et al. (2010) revealed that, multiple regressions analysis identified important and effective parameters that contributed to the water quality variation in Melen River system, Turkey.

This study aims at evaluating the surface water pollution sources through PCA and estimating the contribution of the significant parameters towards water quality variation using multiple linear regressions model.

2. Materials and Methods

2.1 Study Area

Jakara Basin is located in the northwestern Nigeria and lies in the center of Kano city, the most populous city in the whole of Nigeria with over five million people. The region has rapid population growth and industrial development with increase the mass of sewage discharge (Mustapha and Aris, 2011; Mustapha and Nabegu, 2011). Jakara Basin is located on longitude 8° 31´ E to 8° 45´ and latitude 12°10´ N and 12° 13´ N. The basin is about 30km^2 with north-west, south-west orientation sprawling about 0.33°. Jakara

River receives many inputs both anthropogenic and natural in origin that may cause deterioration of the river water quality. The River runs through Kano city for a length of about 13.5 km. Approximately 27.2 km downstream of Jakara River, a dam has been constructed to supply the rural populace with portable water, and to aid irrigation agricultural activities. Jakara dam is the fifth largest of the 22 dams in Kano state, having a total storage capacity of 651,900,000 m^3 and a surface area of 16,590,000 m^2 (Mustapha and Aris, 2011)

2.2 Sampling and Analytical Procedures

The sampling network and strategy were designed to cover wide range of determinant at the key sites, which reasonably represent the surface water quality in the area. Sampling was carried out every day from 31st July to 30th September, 2011 at four different sampling locations along Jakara River. Grab samples were collected at 30cm below the water level using a water sampler and acid washed container to avoid unpredicted changes. The samples were immediately transported to the laboratory under low temperature conditions in ice-boxes and stored in the laboratory at 4° C until analysis.

All samples were analyzed for fifteen physiochemical parameters namely: dissolved oxygen (DO), five-day biochemical oxygen demand (BOD_5), chemical oxygen demand (COD), suspended solids (SS), pH, conductivity, salinity, temperature, nitrogen in the form of ammonia (NH_3), turbidity, dissolved solids (DS), total solids (TS), nitrates (NO_3), chloride (Cl) and phosphates (PO_4^{3-}). Water temperature, DO, pH, conductivity, turbidity, TS, DS, SS and NH_3 of the water samples were detected using multi-parameters monitoring instrument (YSI incorporated, Yellow Spring Ohio, USA). BOD_5 determination of the water samples was carried out using the standard method (APHA, 1998). The dissolved oxygen content was determined before and after the incubation. Sample incubation was for 5 days at 20°C in BOD bottle and BOD_5 was calculated after the incubation period. COD was determined after oxidation of organic matter in strong tetraoxosulphate VI acid medium by $K_2Cr_2O_7$ at 148° C with blank titrations. Cl was determined using 100 mg/l of the water sample which was measured into 250 mg/L conical flask and pH was adjusted with 1 M NaOH. 1 ml/g of $K_2Cr_2O_4$ indicator was then added and titrated with $AgNO_3$ solution. A blank titration was carried out using distilled water. Cl mg/L was then calculated. NO_3 and PO_4^{3-} were determined using calorimetric method (APHA, 1998).

2.3 Principal Component Analysis (PCA)

PCA is one of the best multivariate statistical techniques for extracting linear relationships among a set of variables (Simeonov et al. 2003). PCA is a pattern recognition tool that attempt to explain the variance of a large data set of inter-correlated variables with a smaller set of variables. PCA provides information on the significant parameters with minimum loss of original information (Singh et al. 2004). The PC's can be expressed using the equation below:

$$Zij = ai1x1j + ai2x2j + ... + aimxmj \quad (1)$$

Where Z is the component score, a is the component loading, x is the measured value of a variable, i is the component number, j is the sample number, and m is the total number of variables.

2.4 Multiple Linear Regressions

Multiple linear regressions is a statistical tool for understanding between an outcome variable and several predictors (independent variables) that best represent the relationship in a population (Koklu et al. 2010). The technique is used for both predictive and explanatory purposes within experimental and non-experimental designed. Multiple linear regressions can be expressed using the equation below:

$$Y = \beta o + \beta 1X1 + \beta 2X2 + \cdots\cdots + \beta mXm + \varepsilon \quad (2)$$

Where Y represent the dependent variable, $X1 \cdots Xm$ represent the several independent variables $\beta o \cdots \beta m$ represent the regression coefficient and ε represent the random error.

3. Results and Discussion

3.1 Descriptive Statistics

Table 1 present range, minimum, maximum, mean, standard deviation and variance of the parameters under study. It is clear that, SS, DS, TS, Cl and conductivity are the dominant parameters with high mean concentration of 115.90 mg/L, 104.66 mg L, 105.56 mg/L, 566.88 mg/L and 178.71 mg/L respectively. This showed that, these variables have common source of origin. The mean value of pH ranged from 6.27 to 7.99 mg/L which the average value of 7.99 mg/L which is slightly above neutral level. The concentration of DO, BOD_5, COD ranged from 0.53 to 6.98; 1.00 to 33.00; 20.00 to 154 mg/L which an average value of 2.99, 5.41, and 49.41 mg/L respectively. The order of abundance is COD > BOD_5 > DO, showing less anthropogenic pressure on the surface water.

3.2 Surface water pollution sources apportionment using principal component analysis

The main purpose of PCA is to reduce the contribution of less significant variables to simplify even more of the data structure coming from PCA. This can be achieved by rotating the axis defined by PCA, according to well established rules; the new PCs are now called varifactor (VFs). The measure of sampling adequacy obtained by the Kaiser-Meyer-Olkin method (KMO) was 0.687, indicating that the degree of inter-correlation among the variables and the appropriateness of PCA analysis was valid. Similarly, the Bartlett test of sphericity was significant (p < 0.01), confirming that, the variables are not orthogonal, but correlated. To reduce the overlap of original variables over each PC, a varimax rotation was conducted (Zhang et al. 2011).

Table 2 summarizes the PCA result after rotation, including the loadings, eigen values, the amount of variance explained by each VF and the cumulative variance. The results may be complemented by the examination of the loadings of the five retained components. VF1 explained 37.9% of total variance, had strong positive loading on salinity, TS, DS, conductivity and Cl and a moderate negative loading on NH_3. This factor group is highly and positively contributed by the variables related to natural factors (erosion) and refers to as ionic pollution factor group. The existence of lots of ions and their compound led to high loading of these variables (Zhang et al. 2011). VF2 had strong positive loading of SS and turbidity and explained 17.5% of total variance. High concentration of suspended solids will increase turbidity level, besides, the significant positive correlation between SS and turbidity which indicates common source between the parameters. The association of these variables may have occurred as a result of run-off around

the basin, which may increased the levels of SS and TS.

VF3 had strong loading on BOD_5 and COD and explained 12.1% of the total variance. The values of these parameters were generally higher, since they measure oxygen demand by both biodegradable and non-biodegradable pollutants. The high value obtained could suggest that a large amount of the product was lost to the waste stream. VF4 has strong loading on temperature and DO, explaining 8.5% of the total variance. This factor may be explained by the higher DO value as a result of increase in water volume in the river (Kamble and Vijay, 2011).

VF5 had strong loading on NO_3 and PO_4^{3-}, and explained 7.1% of the total variance. The higher value of nutrients in this factor could have been due to surface run-off from the surrounding farmlands that might have brought ionic substances such as NO_3 and PO_4^{3-} from fertilizer (Boyacioglu and Boyacioglu, 2008).

3.3 Surface water quality prediction using multiple linear regressions model

To find out the best predictor of water quality variation in the Jakara Basin, a stepwise multiple linear regressions model was used. Before interpreting the result, classical assumptions of linear regressions was checked: An inspection of normal p-p plot of regression standardized residuals revealed that all the observed values fall roughly along the straight line indicating that the residuals are from normally distributed population. Moreover, the scatter plot (standardized predicted values against observed values) indicated that, the relationship between the dependent variable and the predictors is linear and the residuals variances are equal or constant.

The water quality variation in the wet season was explained by five predictor variables namely: DO, BOD_5, SS, TS and Cl. The R-square of 0.94.2 revealed that 94.2% of the variation of water quality was explained by the mentioned five predictors. The estimate of coefficient of the model is presented in table 3. The Beta coefficient among the parameters calibrated by stepwise regressions analysis, TS makes the strongest unique contribution in water quality variation (0.668). The Beta value for DO (0.547) was the second highest, followed by Cl (0.545), BOD_5 (-0.491) and the least contributor was SS with -0.292.

The ANOVA table showed that the *F*-statistics (F = 112.697) was very large and the corresponding p value is highly significant (p = 0.0001) or lower than the alpha value (0.01). This indicated that, the slope of the estimated linear regression model is not equal to zero, confirming that, there is linear relationship between the predictors of the models.

4. Conclusion

In this study, principal component analysis (PCA) and multiple regression models were used to evaluate Jakara River water quality data sets. PCA yielded five PCs with 83.1% total variance correspond to five pollution sources namely: ionic, erosion run-off, domestic, dilution and agricultural run-off sources. Multiple linear regression supported PCA result and identified the contribution of each variable with significant values r =0.970, R^2 0.942. These statistical tools provide more objective interpretation of surface water quality variables. From the analysis, it is clear that DO, BOD_5, SS, TS, DS, Cl, salinity and conductivity were found to be the most abundance parameters responsible for water pollution in Jakara River. Therefore, there is need to properly manage wastes in the city and monitor human activities, in order to ensure minimal negative effects on the rivers.

References

Adamu, M. & Aris, A. Z. (2011) "Spatial aspect of surface water quality using chemometric analysis. *Journal of Applied Sciences in Environmental Sanitation*." 6(4), 411-426

Adamu, M., & Nabegu, A. B. (2011) "Surface water pollution source identification using principal component and factor analysis in Getsi River, Kano, Nigeria," *Australian Journal of Basic and Applied* Sciences, 5(12), 1507-1512

APHA (1998). Standard Methods for the Examination of Water and Wastewater, 19th edition. American Water Works Association, Washington, DC.

Boyacioglu, H. & Boyacioglu, H. (2008) "Water pollution sources assessment by multivariate statistical methods in the Tahtali Basin, Turkey," Environmental Geology, 54(2), 275-282

Hai, X. Zhang, Y. L. Mao, Z. G.. Guo, J. J. Xue, Y. S. & Pu, L. Z. (2009) "Anthropogenic impact on surface water quality in Taihu Lake Region," China, Pedosphere, 19(6), 765-778

Huang, F. Wang X. Liping, L. Zhiqing, Z. & Jiaping, W. (2010) "Spatial variation and source apportionment of water pollution in Qiantang River (China) using statistical techniques," Water Research, 44:1562-1572

Juahir, H. Zain, M. S. Yusoff, M. K. Hanidza, T. I. T. Mohd Armi, A. S. Toriman, M. E. Mokhtar, M. (2011) "Spatial water quality assessment of Langat River Basin (Malaysia) using environmetric techniques," Environmental Monitoring and Assessment, 173(1-4): 625-641

Kamble, R. S. Vijay, R. (2011) "Assessment of water quality using cluster analysis in coastal region of Mumbai,India," Environmental Monitoring and Assessment, 178(1-4): 321-332.

Koklu, R. Sengorur, B. Topal, B. (2010) "Water quality assessment using multivariate statistical methods, a case study: Melen River system," Water Resource management, (24)5: 959-978.

Manikannan, R. Asokan, S. Samsoor-Ali. A. M. (2011) "Seasonal variations of physic-chemical properties of the Great Vedaranyam Swamp, Point Calimere Wildlife Sanctuary, South-east coast of India," African Journal of Environmental Science & Technology, 5(9): 673-681.

Milovanovic, M. (2007) "Water quality assessment and determination of pollution sources along the Axios/Vardar River, Southeastern Europe," Desalinization, 213(1-3): 159-173.

Onojake, M. C. Ukerun, S. O. Iwuoha, G. (2011) "A statistical approach for evaluation of the effect of industrial and municipal wastes on Warri Rivers, Niger Delta, Nigeria," Water quality Exposure and Health, 3(2): 91-99.

Praveena, S. M., Kwan, O. I. Aris, A. Z. (2011) "Effects of data pre-treatment procedures on principal component analysis: a case study for mangrove surface sediment datasets," Environmental Monitoring and

Assessment, doi: 10.1007/s10661-011-2463-2.

Satheeshkumar, P. B. Anisa, K. (2011) "Identification of mangrove water quality by multivariate statistical analysis methods in Pondicherry coast, India," Environmental Monitoring Assessment, doi: 10.1007/s10661-011-2222-4.

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(4): 464-475

Simeonov, V. J. Stratis, C. J. Samara, G. J. Zachariadis, D. Voutsa, A. Anthemidis, M. Sofriniou, T. Koumtzis, T. (2003) " Assessment of the surface water quality in Northern Greece," Water Resources, 37(17): 4119–4124

Singh, K.P., Malik, A.D. Mohan, S. 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 resources,38(18): 3980-3992

Singh, K. P., Malik, A. Sinha, S. Vinod, K. Murthy, R. C. (2005) "Estimation of source of heavy metal contamination in sediments of Gomti River (India) using principal components analysis," Water Air and Soil Pollution, 166(1-4): 321-341.

Zhang, X. Q. Wang, Y. Liu, J. Wu, M. Yu, M. (2011) " Application of multivariate statistical techniques in the assessment of water quality in the Southwest, New Territories and Kowloon, Hong Kong," Environmental Monitoring and Assessment, 173(1-4): 17-27

Table 1. Descriptive statistics of the parameters under study

Parameters	Range	Min	Max	Mean	SD	Variance
DO	6.45	0.53	6.98	2.99	1.59	2.54
BOD	32.00	1.00	33.00	5.41	6.30	39.70
COD	134.00	20.00	154.00	49.41	24.09	580.35
SS	974.00	7.00	981.00	115.90	198.03	392.64
pH	1.72	6.27	7.99	7.14	0.34	0.12
NH_3	9.74	0.15	9.89	3.53	2.02	4.07
Temperature	5.17	27.21	32.38	29.53	1.07	1.14
Conductivity	408.00	309.00	412.00	178.71	115.66	134.00
Salinity	25.33	0.14	25.47	10.27	6.87	47.22
Tur.	842.20	7.80	850.00	108.65	192.11	369.72
DS	242.00	169.00	244.00	104.66	674.80	455.00
TS	241.00	348.00	244.00	105.56	672.86	452.00
NO_3	11.00	10.00	21.00	14.73	2.83	8.00

Cl	155.00	40.00	156.00	566.88	366.10	134.00
PO_4^{3-}	24.00	15.00	39.00	28.32	5.52	30.47

Table 2. Rotated component matrix

Parameters	VF1	VF2	VF3	VF4	VF5
Salinity	0.970	-0.036	-0.111	0.141	-0.058
TS	0.968	-0.055	-0.135	0.157	-0.031
DS	0.966	-0.083	-0.133	0.161	-0.029
Conductivity	0.962	-0.032	-0.111	0.176	-0.071
Cl	0.951	-0.114	-0.098	0.095	0.001
NH_3	-0.533	-0.368	0.351	0.291	-0.042
SS	-0.064	0.971	-0.028	-0.125	-0.086
Turbidity	-0.128	0.966	-0.029	-0.135	-0.004
COD	-0.189	0.015	0.909	-0.004	-0.028
BOD_5	-0.21	-0.086	0.871	0.056	-0.079
Temperature	0.219	-0.017	0.309	0.801	-0.016
DO	0.288	-0.136	0.075	0.637	-0.005
pH	0.001	-0.146	-0.249	0.630	-0.054
NO_3	-0.075	-0.060	-0.037	0.079	0.884
PO_4^{3-}	-0.020	-0.013	-0.058	-0.188	0.846
Eigen Value	5.7	2.6	1.8	1.3	1.1
Variance (%)	37.9	17.5	12.1	8.5	7.1
*CV (%)	37.9	55.4	67.5	76	83.1

*CV = Cumulative variance

Table 3. Estimates of coefficients of the multiple linear model

	Beta Unstandardized Coeffient	Std. Error	Beta standardized coefficient	*t*-value	*p*-value
(Constant)	40.689	7.211		5.642	0.000
DO	23.952	2.058	0.547	11.639	0.000
BOD_5	-17.866	1.654	-0.491	-10.805	0.000
SS	-5.825	0.953	-0.292	-6.11	0.000
TS	16.979	5.225	0.668	3.25	0.003
Cl	-10.995	4.078	-0.545	-2.696	0.011

Note: R = 0.970; R^2 = 0.942; Adj. R^2 = 0.933