Interpretation of Observed Surface Ponds Water Quality using Principal Components Analysis and Cluster Analysis

Ayeni A. O.

Department of Geography, University of Lagos, Lagos – Nigeria

E-mail: ayeniao2000@yahoo.com, aayeni@unilag.edu.ng; Phone: +234 (0) 8035894730

Abstract

Varieties of approaches are being used to interpret the concealed variables that determine the variance of observed water quality of various sources. A considerable proportion of these approaches are statistical methods, multivariate statistical techniques in particular. The use of multivariate statistical techniques is required when the number of variables is large and greater than two for easy and robust evaluation. By means of multivariate statistics of principal components analysis (PCA) and cluster analysis (CA), this study attempted to determine major factors responsible for the variations in the quality of 30 surface ponds used for domestic purposes in the 6 selected communities of Akoko Northeast LGA, Ondo State, Nigeria. It classifies the samples' location into mutually exclusive unknown groups that share similar characteristics/properties. The laboratory results of 20 parameters comprising 6 physicals, 8 chemicals, 4 heavy metals and 2 microbial from the sampled ponds were subjected to PCA and CA for further interpretation. The result shows that 5 components account for 97.52% of total variance of the surface pond quality while 2 cluster groups were identified for the locations. Based on the parameters concentrations and the land uses impacts, it was concluded that domestic and agricultural waste strongly influenced the variation and the quality of ponds in the area.

Key Words: Multivariate statistics, ponds, water quality, variance and interpretation

Introduction

The complexity of water quality as a subject is reflected in various types of measurements. These measurements include simple (in situ), basic and more complex parameters (Laboratory). For instance, pH, temperature and DO could be measured with a portable in-situ pH meter, a mercury thermometer and M90 Mettler Toledo AG DO meter, respectively (USGS 2006). BOD, TSS, Cu, Fe, Total bacterial counts, Total coliforms etc could be analyzed in the laboratory using standard methods for water samples examination (Ayoade, 1988; APHA, 1998, WHO 2006 and USGS 2006).

The surface water quality assessment is a matter of serious concern today due to its role in servicing domestic water needs of water stress areas (Yerel, 2010 and Ayeni *et al*, 2011). The

surface water (ponds) quality is principally influenced by the natural and the anthropogenic processes particularly in the urban areas and agricultural activities around the rural areas (Ayeni, 2010 and Ayeni *et al*, 2011). The level of water quality is relatively determined by the content of physical, chemical and biological parameters present in it. Relationship between two parameters may also lead increases or decrease in the concentration of others. This relationship or association is usually achieved using multivariate statistical techniques (Ifabiyi, 1997; Mazlum *et al*, 1999; Jaji *et al* 2007). This is because some analysis is primarily concerned with relationships between samples, while others trepidation are largely with relationships between variables. According to Mazlum *et al*, (1999) and Yerel, (2010), many multivariate statistical techniques have the capacity to summarize a large data by means of relatively few parameters. Nonetheless, the choice of using any of the multivariate statistical techniques lies on the nature of the data, problem, and objectives of the study. In view of the fact that the daily drinking and domestic water needs of the majority of residents in the area are met by unsafe surface water, in particular surface ponds (Ayeni, 2010), there is the need to understand the variables that control their quality variation. Principal Component Analysis (PCA) and Cluster Analysis (CA) of multivariate techniques are therefore adopted for the study. According Praus, (2005), PCA is used to search new abstract orthogonal eigenvalues which explain most of the data variations in a new harmonize structure. Each principal component (PC) is a linear combination of the original variables and describes different source of information by eigen-value based on the decomposition of the covariance/correlation matrix (Geladi and Kowalski, 1986). PCA is designed to modify the observe variables into uncorrelated variables of linear combinations of the original variables called "principal components" (Praus, 2005 and Yerel, 2010) as well as to investigate the factors which caused variations in the observed datasets (Mazlum *et al*, 1999). The principal component therefore provides information for interpretation and better understanding of the most meaningful parameters, which describes the whole data set through data reduction with a minimum loss of the original information. Cluster analysis (CA) is an exploratory analysis technique for classifying a set of observations into two or more mutually exclusive unknown groups based on combinations of interval variables (Stockburger, 1997; Trochim, (2006): Murali-Krishna *et al* 2008 and Yerel, 2010). According to Yerel (2010), CA organizes sampling entities into discrete clusters, such that within-group similarity is maximized and among-group similarity is minimized according to some objective criteria Its purpose is to discover a system of organizing observations and sort them into groups so that it is statistically easier to predict behavior of such observations based on group membership that share similar identities/properties. In this study observation, sampling location classification was done by the use of Hierarchical Cluster Analysis (HCA) procedure. HCA identify relatively homogeneous groups of variables (cases) through dendrogram based on selected characteristics. Dendrogram clearly distinguished locations bahaviours and interprets the description of the hierarchical clustering in a graphical format (Hastie et al, 2001 and Ryberg, 2006).

Study Area and Sampling Locations

The study area lies between longitude 5°38' and 6°04'E, and latitude 7°26' and 7°42'N in the northern senatorial part of Ondo State, SW – Nigeria (Fig. 1). It is bounded by Akoko North West LGA to the north, Edo State to the east, Akoko South East and West LGAs to the south, and Ekiti State to the west. It is primarily characterized with undulating relief ranges between 149 and 671m above sea level and located on basement complex rock formation (Ayeni, 2010). The complex composed mainly of granite, mischist, gneisses and metasediment (Barbour *et al*, 1982 and Adekunle *et al,* 2007). The area falls within sub-tropical climate with average rainfall over 1500mm per annum. The temperature ranging from about 30^0C to 38^0C while the vegetation cover is dominated by derived secondary rain forest. The soil is classified as Ferric Acrisols with relatively higher cation profiles (Fasona *et al,* 2007; Nwachokor and Uzu, 2008).

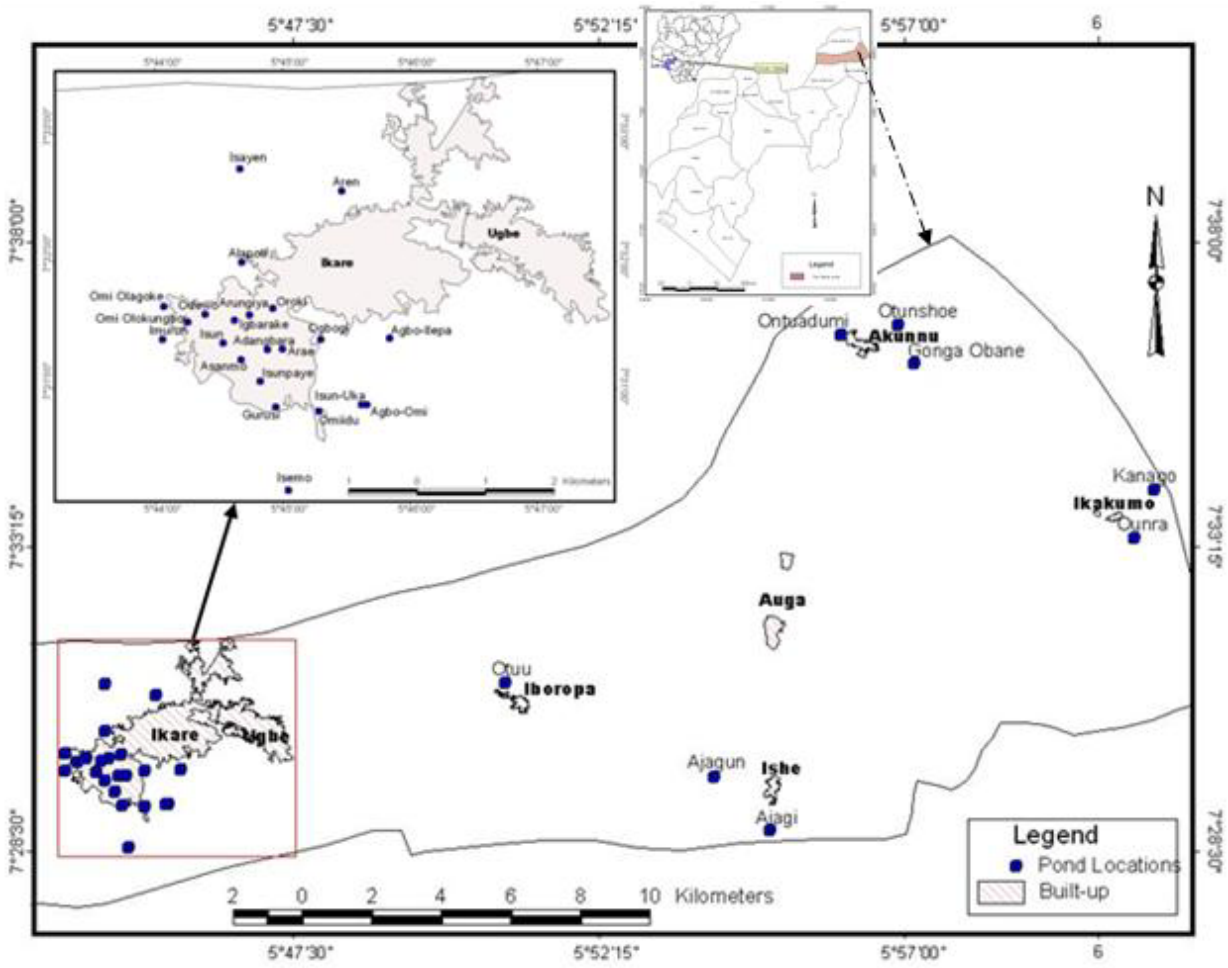

Fig. 1: Selected pond locations in Akoko Northeast LGA of Ondo State, SW - Nigeria

Methods

Twenty (20) water quality parameters from 30 surface ponds were monitored for 12 months. For each month, water sample from selected ponds are collected and analysed in the laboratory using APHA (2005) standard methods for the examination of water and wastewater. The coordinates of sampled ponds are interpolated on geo-rectified map of the study area (Fig.

1). The selected surface water quality parameters for the study are pH, temperature, Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD), Total Suspends Solid (TSS), Total Dissolved Solid (TDS), Turbidity, Total Hardness (TH), Calcium Hardness (Ca^{+}), Magnesium Hardness (Ma^{2+}), Chloride (Cl^{-}), Nitrate (NO_3^{-}), Phosphate (PO_4^{3-}), Oil & grease, Cupper (Cu), Iron (Fe), Manganese (Mn), Zink (Zn), Total bacterial counts (TBC) and Total coliforms (TC).

The laboratory results were evaluated using multivariate statistical techniques of PCA for selected parameters and CA for sample locations. The principal component is thus given by the formula:

$$Z_{ij} = a_{i1}X_{ij} + a_{i2}X_{2j} + a_{i3}X_{i3} + \ldots\ldots + a_{in}X_{nj} \quad (1)$$

where,

z = component score,
a = component loading,
x = measured value of variable,
i = component number,
j = sample number, and
n = the total number of variables.

In the case of cluster analysis, the formula is given thus:

$$d_{ij}^2 = \sum_{k=1}^{m} (z_{ik} - z_{jk})^2 \quad (2)$$

where

d^2ij = the Euclidean distance,
z_{ik} = the values of variable k for object i
z_{jk} = the values of variable k for object j
m = the number of variables.

Results and Discussion

The result of principal components analysis in Table 1 shows that of the 20 components, only 5 had extracted eignvalues over 1. This is based on Chatfield & Collin (1980) assumption which started that components with an eigenvalue of less than 1 should be eliminated. The extracted 5 components were subsequently rotated according to varimax rotation in order to make interpretation easier and fundamental significance of extracted components to the water quality status of the selected ponds. The result of rotation revealed that the percentages of the total variances of the 5 extracted components when added account for 97.52% (i.e. their cumulative variance) of the total variance of the observed variables. This indicates that the variance of the observed variables had been accounted for by these 5 extracted components. The calculated components loadings, eigenvalues, total variance and cumulative variance is shown in Table 2 while the scree plot of the eignvalues of observed components is depicted shown in Fig. 2.

Table 1: Principal Component Matrix of eignvalues less than 1 (5 components extracted)

	Component				
	1	2	3	4	5
pH	.838	.001	-.195	.379	.171
Temp	-.290	.273	-.589	.649	-.067
Do	.687	-.349	-.592	.150	.115
BOD	-.535	.334	.581	.450	-.243
TSS	.641	-.361	.378	.484	.200
TDS	.688	.715	.050	-.008	.092
Turbidity	.755	-.399	-.380	.324	-.042
TH	.734	.654	.145	-.004	-.097
Ca	.825	.476	.258	.116	.035
Mg^{2+}	.576	.776	.017	-.126	-.222
Cl^{-}	.449	.871	.018	-.152	.104
NO_3^{-}	-.361	.451	-.520	-.083	.578
PO_4^{3-}	.928	-.022	.045	.020	-.221
Oil & grease	-.273	.408	.312	.427	.663
Zn	.621	-.641	-.224	.379	-.084
Fe	-.513	.230	.428	.701	.069
Mn	-.353	.209	.096	.517	-.744
Cu	-.026	-.449	.822	-.309	.161
TBC	.462	-.439	.596	.304	.324
TC	-.770	.126	-.490	.346	.173

Table 2: Rotated Component loading matrix, eignvalues, total variance and cumulative variance

Variables	Component				
	1	2	3	4	5
TH	.983				
Mg^{2+}	.979				
TDS	.979				
Cl^{-}	.954				
Ca^{+}	.900				
Zn	.	.959			
Turbidity		.947			
Do		.840			
pH		.818			
TSS		.779			
PO_4^{3-}		.621			
Cu			-.943	.	
Temp			.872		
TC			.766		
TBC			-.699		
NO_3^{-}			.661		
Oil & grease				.953	
Fe				.745	
Mn					.952
BOD					.773
Eigenvalue	7.399	4.412	3.335	2.581	1.776
Total variance %	36.996	22.058	16.674	12.908	8.882
Cumulative variance %	36.996	59.054	75.728	88.634	97.517

Rotation Method: Varimax with Kaiser Normalization

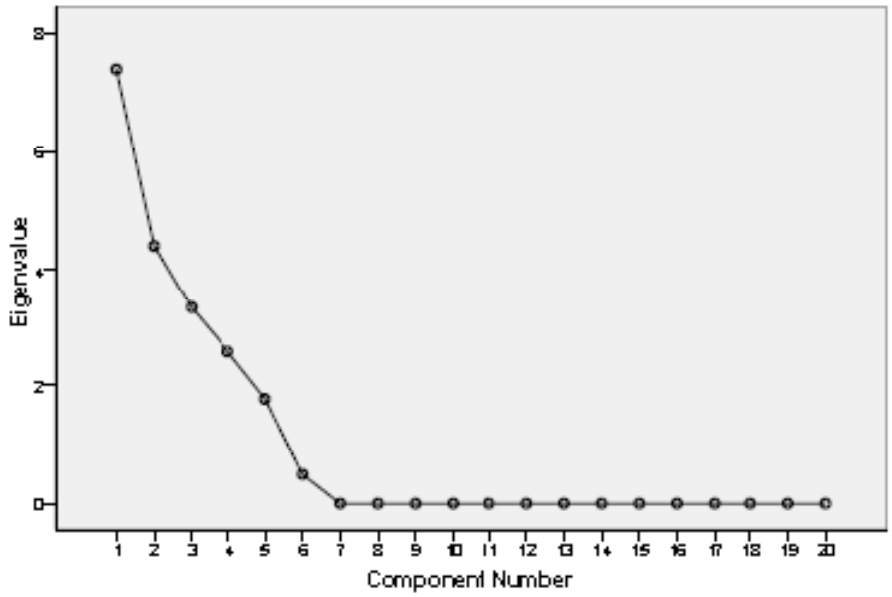

Fig. 2: The scree plot of the eignvalues

Based on the component loadings, the variables are grouped accordingly with their designated

components as follows:

Component 1 TH, Mg^{2+}, TDS, Cl^{-} and Ca^{+},
Component 2: Zn, Turbidity, DO, pH, TSS and PO_4^{3-}
Component 3: Cu, temperature, TC, TBC and NO_3^{-}
Component 4: Oil & grease and Fe
Component 5: Mn and BOD

Component 1, component 2, component 3, component 4 and component 5 explained 36.996%, 22.058%, 16.674%, 12.908% and 8.882% of the variance respectively. Classifying the component loading according to Liu *et al* (2003), the loading values greater 0.75 signifies "strong", the loading with absolute value between 0.75 and 0.50 indicate "moderate" while loading values between 0.50 and 0.30 denote as "weak"., respectively. Using this classification, all variable in component 1 and component 2 had strong positive loading except PO_4^{3-} with moderate positive. Amongst the 5 variables in component 3, two (2) had strong positive loading (Temperature and TC), NO_3^{-} had moderate positive loading while Cu and TBC were signified with strong and moderate loading respectively. All variables in components 4 and 5 explained strong positive loading.

An interpretation of the rotated 5 principal components is made by examining the component loadings noting the relationship to the original variables. Component 1 gives information about the variation in TH, Mg^{2+}, TDS, Cl^{-} and Ca^{+}. In this component, loading indicates that organic matter and organic acids which could be attributed to various anthropogenic activities and geological formation and/or composition of the area greatly influence the quality of selected ponds. The same also interpreted for component 2 but considers its eignvalue and total variance, it is quite lower compared with component 1.

Components 3 explained information about Cu, temperature, TC, TBC and NO_3^{-}. This component represents pollution from domestic and agricultural waste as well as geological composition of the area. However, the significance of NO_3^{-} in the component 3, indicates that nitrification takes place in the vicinity of the ponds. In the component 4, it can be understood that dissolved or emulsified oil and grease extracted from water especially unsaturated fats and fatty acids and Fe extracted from the parent rock of the area are of the significance in that component. In the component 5, Mn presence is an indication of the parent rock influence while BOD claimed that there is a high level of organic pollution, caused usually by poorly treated waste water.

The dendrogram of observed locations dataset was generated using Euclidean distance of HCA for CA result (Fig. 3 and Table 3). Based on Euclidean distance, two major clustering groups (cluster 1 and cluster 2) were observed. Cluster 1 characterized with low Euclidean distance correspond to locations 6, 9, 7, 3, 13, 19, 17, 22, 10, 30, 4, 5, 2, 12, 23, 27, 18, 26, 29, 21, 25, 28 and 24. Cluster 2 which has high Euclidean distance is coherent to locations 15, 16, 1, 20, 8, 11

and 14. Sub group clusters were also clarified within the major cluster 1 and vary with significance Euclidean distance.

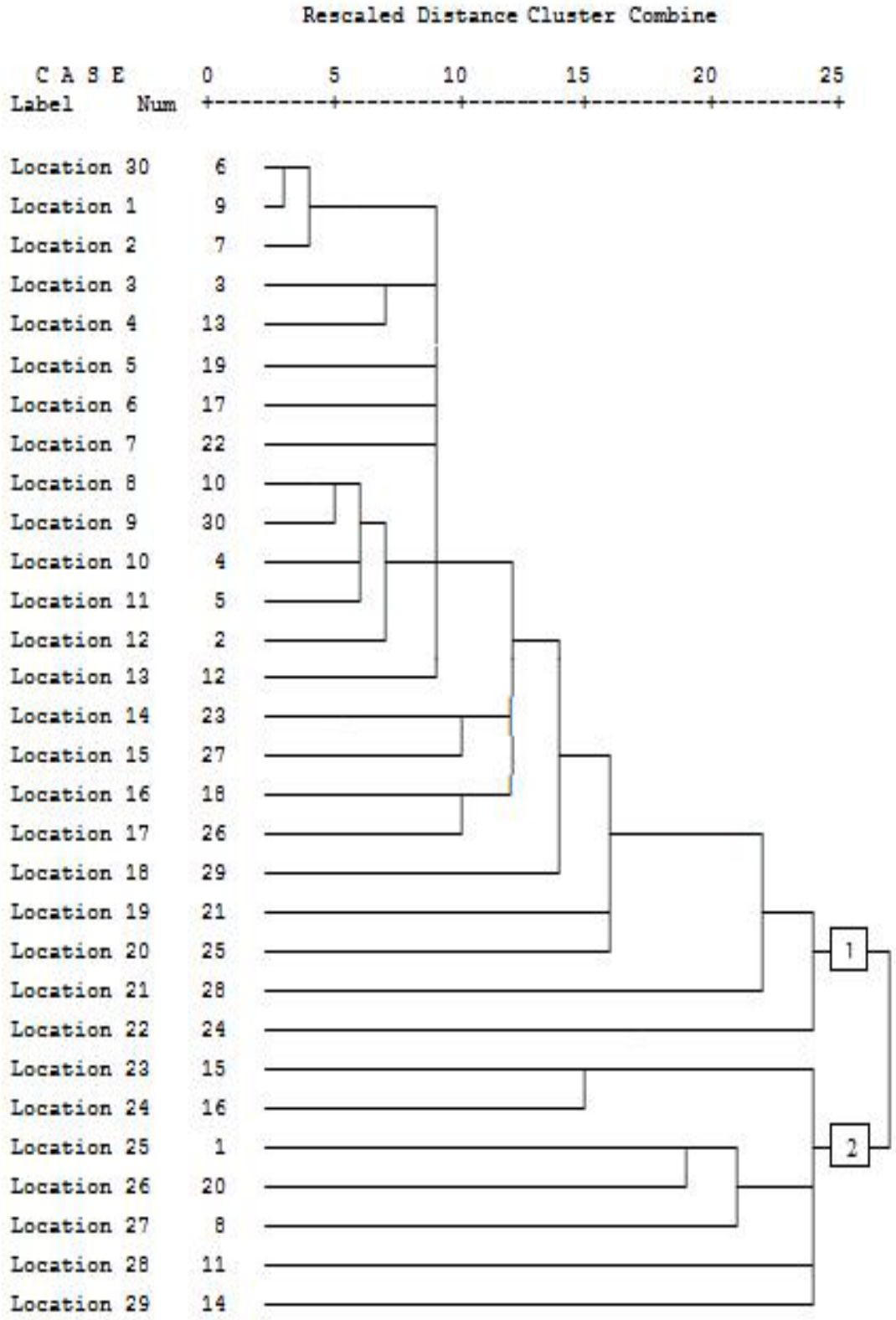

Fig. 3: Dendrogram of hierarchical cluster analysis

Table 3: Pond names and their cluster membership

Identification	Ponds' names	Clusters	Identification	Ponds' names	Clusters
Location 1	Omiidu	1	Location 16	Otunadumi	1
Location 2	Agboomi	2	Location 17	Gonga Obane	2
Location 3	Isun-uka	2	Location 18	Isun	2
Location 4	Isemo	2	Location 19	Imurun	2
Location 5	Gurusi	2	Location 20	Arungiya	1
Location 6	Alapoti	2	Location 21	Igbarake	2
Location 7	Isayen	2	Location 22	Omi-Alagoke	2
Location 8	Aren	1	Location 23	Omi-Olokungboye	2
Location 9	Agbo Ilepa	2	Location 24	Odewo	2
Location 10	Ajagi	2	Location 25	Oroki	2
Location 11	Ajagun	1	Location 26	Asanmo	2
Location 12	Ootu	2	Location 27	Ogbogi	2
Location 13	Ounra	2	Location 28	Arae	2
Location 14	Kanago	1	Location 29	Adangbara	2
Location 15	Otunshoe	1	Location 30	Isunpaye	2

The dendrogram clarifies cluster 1 as the abnormality observation which had high variation in the concentration of the surface water quality parameters compared to cluster 2 surface water samples concentration. The variation in cluster 1 might be due to low polluted effluents from non-point sources (agricultural and urban activities). Cluster 2 shows a high pollution from agricultural area which encompasses the ponds.

Conclusion

This study presents the usefulness of multivariate statistical techniques of large and complex dataset in order to obtain better information and interpretation concerning surface water quality. Principal component analyses helped in identify the factors responsible for surface water quality variations in 6 selected communities. The result revealed that the percentages of the total variances of the 5 extracted components when added account for 97.52% (i.e. their cumulative variance) of the total variance of the observed variables. The variation in components 1 and 2 loading indicate that organic matter and organic acids could greatly influence the quality of selected ponds. Components 3 ascribed mainly to domestic and agricultural waste of the ponds environment while component 4 and 5 respectively attributed to dissolved/emulsified poorly treated waste water. On the other hand, the result of cluster analysis revealed 2 major clustering groups resulting from influence of agricultural and urban activities around the samples' location. Cluster 1 characterized with low Euclidean distance corresponds to 23 locations and clarifies with sub groups that varies with significance Euclidean distance while cluster 2 coherent to 7 locations and observed high Euclidean distance with sub group of insignificance Euclidean distance. Therefore, it is worthwhile to conclude that PCA and CA are better tools for better understanding of the concealed information about parameters variance and datasets discrete information in water quality assessment studies.

References

APHA (2005): "Standard methods for examination of water and wastewater". American Public Health Association, 21th Edition, Washington DC, Pg 4 - 144

Ayeni, A. O. (2010) "Spatial Access to Domestic Water In Akoko Northeast LGA, Ondo State, Nigeria" Unpublished Ph.D. Thesis, University of Lagos, Lagos, Nigeria

Ayeni, A.O., I.I. Balogun and A.S.O. Soneye (2011) "Seasonal Assessment of Physico-chemical Concentration of Polluted Urban River: A Case of Ala River in South-western Nigeria", *Res. J. Environ. Sciences, 5(1): 21-35*

Ayoade, J. O. (1988): "Tropical Hydrology and Water Resources", London, Macmillan Publishers Ltd

Barbour, J. S. Oguntoyinbo, J. O. C. Onyemelukwe and J. C. Nwafor (1982): Nigeria in Maps, London, Hodder and Stoughton.

Chatfield, C. & Collin, A.J. (1980): Introduction to Multivariate Analysis". Published in the USA by Chapman and Hall in Association with Methuen, Inc. New York, USA. Cited in

Mazlum, N., A. Ozer and S. Mazlum (1999): Interpretation of Water Quality Data by Principal Components Analysis, *Tr. J. of Engineering and Environmental Science 23, 19 - 26.*

Fasona, A.S., F.O. Omolayo, A.A. Falodun and O.S. Ajayi (2007): Granite Derived Soils in Humid Forest of Southwestern Nigeria - Genesis, Classification and Sustainable Management, *America – Eurasian J. Agric & Environ. Sc. 2(2): 189 - 195*

Geladi P. and B. R. Kowalski (1986): Partial least square regression: A tutorial. *Anal. Chim. Acta 185: 1-17.*

Hastie, T., Tibshirani, R., and Friedman, J., (2001), The elements of statistical learning—data mining, inference, and prediction: New York, Springer Science+Business Media, Inc., 533 p.

Ifabiyi, I.P. (1997): "Variation in Water Gravity with Rainfall Incidences: A Case Study of Ogbe Stream Ile-Ife, Ife" *Research Publications in Geography*, *Vol. 6 No. 1 and 2 pp 139-144.*

Jaji, M.O, O. Bamgbose, O.O. Odukoya and T.A. Arowolo (2007): Water Quality Assessment of Ogun River, South West Nigeria, *Environmental Monitoring Assessment*, *133: 473-482* Springer Science and Business Media

Mazlum, N., A. Ozer and S. Mazlum (1999): Interpretation of Water Quality Data by Principal Components Analysis, *Tr. J. of Engineering and Environmental Science 23, 19 - 26.*

Murali Krishna P.S. M., S. G. Moses and V.S.G. K. Murali Krishna (2008): Application cluster analysis, discriminate analysis and principal component analysis for water quality evaluation for the river Godavari at Rajahmundry region, *International Journal of Applied Environmental Sciences 2(3): 195-209*

Nwachokor, M.A. and F.O. Uzu (2008): Updated Classification of Some Soil Series in Southwestern Nigeria *Journal of Agronomy 7(1): 76-81*

Praus, P (2005): Water quality assessment using SVD-based principal component analysis of hydrological data, *Water SA 31 (4): 417 - 422*

Ryberg, K. R (2006): Cluster Analysis of Water-Quality Data for Lake Sakakawea, Audubon Lake, and McClusky Canal, Central North Dakota, 1990-2003, Prepared in cooperation with the Bureau of Reclamation, U.S. Department of the Interior Scientific Investigations Report 2006–5202, U.S. Geological Survey, Reston, Virginia, 47p

Stockburger, D. W (1997): *Multivariate Statistics: Concepts, Models, and Applications,* www Version 1.0 http://www.psychstat.missouristate.edu/multibook/mlt04.htm accessed on the 29/06/2011

USGS (2006): Common water measurements Page Last Modified: Monday, 28-Aug-2006 14:56:21 EDT http://ga.water.usgs.gov/edu/characteristics.html Accessed on the 14/02/2009

WHO (2006): *Guidelines for Drinking-Water Quality,* Volume 1, Recommendations 1st Addendum to 3rd edition, World Health Organization, Geneva (Electronic version).http://www.who.int/water_sanitation_health/dwq/gdwq3rev/en/index.ht mlAccesse d on the 14/02/2009

Yerel, S. (2010): Water Quality Assessment of Porsuk River, *Turkey E-Journal of Chemistry, http://www.e-journals.net 2010, 7(2), 593-59.9*