Ayoko, Godwin and Singh, Kirpal and Balerea, Steven and Kokot, Serge (2007) Exploratory multivariate modeling and prediction of the physico-chemical properties of surface water and groundwater. *Journal of Hydrology* 336(1-2):pp. 115-124.

Exploratory multivariate modeling and prediction of the physico-chemical properties of surface water and groundwater

Godwin A. Ayoko*[a], Kirpal Singh [b], Steven Balerea [b], Serge Kokot[a]

[a]School of Physical and Chemical Sciences, Queensland University of Technology, GPO 2434, Brisbane 4001, Australia.

[b]School of Natural Sciences, University of Papua New Guinea, P.O. Box 320, University Post Office, NCD Papua New Guinea

.

Author to whom correspondence should be addressed. Email: g.ayoko@qut.edu.au; Fax: +61 7 31381804; Tel: +61 7 31381206

Abstract

Physico-chemical properties of surface water and groundwater samples from some developing countries have been subjected to multivariate analyses by the non-parametric multi-criteria decision-making methods, PROMETHEE and GAIA. Complete ranking information necessary to select one source of water in preference to all others was obtained, and this enabled relationships between the physico-chemical properties and water quality to be assessed. Thus, the ranking of the quality of the water bodies was found to be strongly dependent on the total dissolved solid, phosphate, sulfate, ammonia-nitrogen, calcium, iron, chloride, magnesium, zinc, nitrate and fluoride contents of the waters. However, potassium, manganese and zinc composition showed the least influence in differentiating the water bodies. To model and predict the water quality influencing parameters, partial least square analyses were carried out on a matrix made up of the results of water quality assessment studies carried out in Nigeria, Papua New Guinea, Egypt, Thailand and India/Pakistan. The results showed that the total dissolved solid, calcium, sulfate, sodium and chloride contents can be used to predict a wide range of physico-chemical characteristics of water. The potential implications of these observations on the financial and opportunity costs associated with elaborate water quality monitoring are discussed.

Keywords: multivariate modeling and prediction, water quality

Introduction

Water is vital to health, well-being, food security and socio-economic development of mankind. Therefore, the presence of contaminants in natural freshwater continues to be one of the most important environmental issues in many areas of the world, particularly in developing countries, where several communities are far away from potable water supply (WHO, 1993; WHO, 1996). Low-income communities, which rely on untreated surface water and groundwater supplies for domestic and agricultural uses are the most exposed to the impact of poor water quality. Unfortunately, they are also the ones that do not have adequate infrastructure to monitor water quality regularly and implement control strategies (Ongley and Booty, 1999). Many of such communities abound in developing countries, where irregular supply of electricity for continuous pumping of treated water, absence of piping systems in some areas, variability of rainfall and increased demand for water as the population grows seriously impede access to potable water supply. Drastic changes in climatic conditions make the situation worse. For example, as a result of the 1997 El Nino, it was estimated that over a million people in Papua New Guinea faced acute food shortage and at least 45, 000 people were without fresh water (UNOCHA, 1998). However, the problem of potable water supply is a global issue. Many of the earth's major rivers and groundwater supplies are either overexploited or polluted due to population growth, agricultural activities, urbanisation and industralisation. In Poland, three quarters of the rivers were thought to be too polluted even for industrial use; two-thirds of China's rivers were regarded as contaminated; forty rivers in Malaysia were reported as not being able to support aquatic life due to pollution; and in Manila, Philippines, over 60% of the main rivers supposedly contain untreated sewage (IRC, 1995). Thus, it is increasingly desirable to obtain reliable assessments of water quality, which can be used for water resource planning and assessment of policy options in order to sustain human well being, industrial growth and food security.

Differences in the pollutant loading of freshwaters from different sources may result from differences in the geological background, hydrological systems, anthropogenic activities and transformations of water components by microorganisms (Einax et al, 1997). Therefore, pollutant concentrations from different sampling sites and environmental phases are subject to high variability, which require careful evaluation and interpretation. In addition, water quality depends on a variety of physico-chemical parameters and meaningful prediction, ranking analysis or pattern recognition of the quality of water require multivariate projections methods for simultaneous and systematic interpretation. Thus a wide range of multivariate projection methods has been applied to hydrospheric samples (Einax et al, 1997). However, the multi-criteria decision making methods, PROMETHEE (Preference Ranking Organisation METHod for Enrichment Evaluation) and GAIA (Geometrical Analysis for Interactive Aid) (Brans, 1991 and 2002; Brans and Mareschal, 1989 and 2005; Kokot and Phuong, 1999; Epinasse et al, 1997; Ayoko, et al, 2004) have not been employed in the literature for multivariate ranking analysis of the parameters that influence water quality. Similarly, there are relatively few applications of multivariate predictive modeling in environmental problems (Eriksson, 2001).

To assess the quality of water for drinking and irrigation purposes, many variables are routinely monitored. This produces a large database but the process of data acquisition can be time-consuming, laborious and expensive while accurate interpretation of the multivariate data can be challenging.

This paper reports the physico-chemical properties of some surface - and ground - waters and the interpretation of the data with the aid of Chemometrics methods. To provide scientific data, which can be used for water resource planning and assessment of policy options in countries where availability of potable water supply is a problem, we used data from exploratory water quality studies that were conducted in Nigeria, Papua New Guinea, Egypt, Thailand and India/Pakistan to (i) model and predict the relationships between the factors (X) and the water quality influencing responses (Y), (ii) understand which X variables influence particular Y variable(s) and (iii) which X variables can be used as surrogates to provide information about water quality. The work enhances basic knowledge of the physico-chemical quality of ground - and surface - waters from developing countries, and provides a guide to the understanding of the relationships between the factors and responses, and the implementation of predictive models for water quality.

Materials and methods

Samples of surface water from rivers and groundwater from boreholes were collected in Papua New Guinea at a period that coincided with the end of dry season and the beginning of the raining season (November to December). For comparison, the quality indicators for water samples from Hudiara drain which extends over a distance of 44.2 km in India and 54.4 km in Pakistan (Afzal et al, 2000), as well as various groundwater and surface waters in Nigeria (Okoye,1991; Olajire et al, 2001; Agbu, 1984 and Alaribe, 1984, Ibe and Njemanze, 1999) Chao Praya and Mae Klong rivers in Thailand (Kruawal et al, 2005), groundwater samples from Egypt (El-dars, 2005) and Fly river in Papua New Guinea (Salomons and Eagle, 1990) were used as reported in the literature.

High-density polyethylene containers capable of taking up to 1L of water and equipped with screw caps were used for sample collection. Each container was washed with 1M HCl and rinsed several times with deionized water before sample collection. The samples were stored in an Esky containing ice and then transported immediately to the laboratory, where they were filtered through 0.45 μm millipore filters. The filtrates were acidified to pH <2 with 6M HNO_3 and stored at $4^{o}C$ in a refrigerator until analysed for the metals by standard methods (APHA, 1989). The pH, temperature, conductivity and Total Dissolved Solid (TDS) measurements were done on site using a Metrohm 620 pH meter for pH measurements and Hach conductivity/Total Dissolved Solid meter for electrical conductivity, temperature and total dissolved solid measurements. All other parameters were determined within two days of sample collection.

Ammonia-nitrogen was determined by the Nessler method. HACH 5 DR/2000 spectrophotometer, digital titrator was used for measurements of physical and inorganic ions (HACH, 1989) and Atomic Absorption Spectrometer (Perkin Elmer 310) was employed for the determination of the metals. Standard calibration curves were obtained by analyzing standards prepared by serial dilution of 1000 ppm stock solutions. Duplicate determinations were made on most samples and the blank and standard curves were checked after every 10 determinations. The replicated measurements did not differ by more than 5%.

Data processing

All analytical data were initially processed using Microsoft Excel 2003 software for Windows (Microsoft Corporation) and then subjected to PROMCALC software and Decision LAB 2000 (Brans, 1991; Decision Lab 2000, 1999) for multi-criteria decision making (MCDM) analysis by the PROMETHEE and GAIA procedures or to SIMCA–P 10.0 Umetrics AB for principal component analysis and partial least squares.

PROMETHEE and GAIA Procedures

PROMETHEE and GAIA procedures have been described by various researchers (Brans, 1991 and 2002; Brans and Mareschal, 1989 and 2005; Kokot and Phuong, 1999; Epinasse et al, 1997; Ayoko, et al, 2004). Essentially, they are non-parametric methods based on pairwise comparisons of the objects and variables. PROMETHEE facilitates the ranking or ordering of a number of objects (in this work, the water bodies) according to preference and weighting conditions, which have been pre-selected by the user and are applied to the variables (e.g. concentrations of nutrients, pH, temperature, conductivity, and total dissolved solid). Therefore, the first step was to choose a preference function, which provides the mathematical basis for selecting one object in preference to another. Of the six preference functions available in the procedures, the V-shaped function (P), which required a threshold value to be applied to each variable was used in this work. The threshold was set at the WHO Limit for each variable. But when there is no WHO Limit for a variable, the highest value of the variable in a given column was used. Additionally, since a lower value indicates a better water quality, it was specified that lower variable values are preferred by choosing the 'minimise' function when modeling each variable.

To refine the preference selection process, positive and negative outranking flows, ϕ^{+} and ϕ^{-} respectively were computed within PROMETHEE. The former expresses how each object outranks all others while the latter indicates how each object is outranked by all the other objects. By applying the set of rules described previously (Brans, 1991 and 2002; Brans and Mareschal, 1989 and 2005; Kokot and Phuong, 1999; Espinasse et al, 1997; Ayoko, et al, 2004), PROMETHEE II , which provides a full ranking of all objects from the best to the worst based on their net outranking flow was obtained.

GAIA, on the other hand, uses principal component analysis techniques to evaluate and display PROMETHEE results visually. It facilitates the interpretation of the global performance of each object with reference to a decision vector, π, which appears in the biplot. Thus, useful information about the underlying trends in the data matrix such as clustering of objects or variables and characterisation of outliers may be obtained from GAIA biplots. One of the marked advantages of GAIA procedures over other multivariate data analysis methods is that the reduction and standardization of data to unit variance is unnecessary (Massart et al, 1997) since PROMETHEE serves as a data pre-treatment procedure for GAIA. Furthermore, PROMETHEE and GAIA use only two principal components to produce results that are similar to those from principal component analysis. Nevertheless, their outcomes are comparable to those of their common alternatives (Brans and Mareschall, 1989; Geldermann, 2001) and they have been rated as the best among sixteen multi-criteria decision making methods applied to solve a multi-criteria water bodies problem (Al-Shemmeri, 1997).

Partial Least Squares (PLS)

PLS works with two matrices X (e.g. factors) and Y (e.g. responses). The main objectives of this technique are to: (i) well approximate X and Y and (ii) to model the relationship between them (SIMCA P 10.0,UmetricsAB; www.umetrics.com). The response block is represented by the Y scores, U, while the predictive block (X) is described by X scores, T. PLS maximises the covariance between U and T by decomposing Y and X as shown in the equations below.

$$Y = UC^T + F \quad (1)$$

$$X = TP^T + E \quad (2)$$

where P and C are loadings or loadings vectors, and E and F are the residuals or errors in X and Y matrices respectively.

Validation of PLS Models

When data sets that were not originally designed for calibrations are analysed (as in this work), there is a high probability that over-fitting would occur and that chance correlation rather than real correlations are observed. To avoid this, model validation is performed by cross-validation ("leave-one-out" method) or by using a calibration set. In this work, internal validation using cross-validation, external validation and response permutation were used.

Cross-validation: This involved keeping out parts of the data during model development, developing the model from the reduced data, predicting the parts kept out by different models and comparing the predicted values with actual values (Wold 1978). A predicted variation, Q^2 (the fraction of the total variation of the X's that can be predicted by a component) is calculated for the optimal number of PLS components and this can be compared with the R^2 (the fraction of the sum of squares explained by the component). In this work, cross-validation was carried out using the software, SIMCA P 10.0,Umetrics AB.

External validation: The entire data matrix was split into two nearly equals. One half was used as the calibration set and the other as the prediction set.

Response permutation: This was performed in order to confirm the significance of the R^2Y (fraction of the sum of squares of all the Y's explained by a component) and Q^2Y (fraction of the total variation of Y's that can be predicted by the component) values obtained from the internal validation process. Detailed results are presented under results and discussion.

Results and discussion

General description of the five of the ground - and surface - waters

The results of the analyses carried out on the surface waters and groundwater from Papua New Guinea are presented in Table 1 (which also describes the abbreviations subsequently used for these water bodies).These results are the means of triplicate measurements, which agreed within ±5% of each other. It is evident that a wide variation exists in the quality of water from the water bodies sampled.

The pH of all of the samples is generally within the optimum range of 6.5-9.5 (WHO, 1996) with most having pH values less than 7. The pH of surface - and ground-waters usually reflects their humic acid, CO_2, CO_3^{2-} and HCO_3^- contents (Olajire and Imeokparia, 2001; Jior et al, 1991) and the observed pH values may suggest the presence of acidic matter such as humic acids and free CO_2 in these water samples. The temperature (27.3-28.6^{0}C) is also within the international allowable standard. While its value has no direct effect on human health and well-being, it may impact on the rate of chemical and biochemical reactions, the solubility of gases in the water, and in turn the taste and odour. Consequently, strong unpleasant odour from water may reflect the release of dissolved gases at high temperatures.

The electrical conductivity of the samples varied widely from 60 to 530 μmho/cm and reflects the amount of charged substances in the water samples. Similarly, the Total Dissolved Solids (TDS), which gives a good indication of the salinity ranged from 30 – 250 ppm. WHO recommends that its value should be less than 500 ppm but pegs acceptable limits for water potability at 1500 ppm (WHO, 1996). Additionally, TDS in excess of 1000 ppm are objectionable to consumers and have adverse effects on crop production (Pescod, 1977).

The concentrations of the inorganic anions (F^-, NO_3^-, NO_2^-, PO_4^{3-} and SO_4^{2-}) in the waters were generally well within their respective WHO limits. Although there are considerable concerns about the level of fluoride in water (Akher, 1998), the highest observed fluoride level was 0.1 ppm for the Papua New Guinea groundwater 2 (PGG2). Even at this level, it is well below the threshold concentration (14 ppm) for the onset of crippling fluorosis but it is within the limit for the onset of mild dental fluorosis (WHO, 1996).

The sodium contents of the samples were within the desirable limit of 200 ppm for drinking water (WHO, 1996). WHO has no limit for potassium, which is usually present in water in lower proportions than sodium. The Ca and Mg contents are generally below 100 ppm in these samples. This suggests that they are soft waters, which have high tendencies to be corrosive to water pipes (WHO, 1996). Although these cations are not present at toxic levels in the water samples, the water samples may still be phytotoxic and this could limit their use for agricultural purposes (Pescod, 1992). In contrast to the concentrations of Na, Ca, and Mg, the ammonia concentrations showed little variation from one sample to another but were generally lower than the WHO guideline value limit of 1.5 ppm as were the zinc concentrations. At 0.4 ppm, the concentration of iron in PGG2 was above 0.3 ppm, which is the limit above which iron stains laundry (WHO, 1993;WHO, 1996). The manganese concentrations of two of the Papua New Guinea wells (PGG2 and Papua New Guinea groundwater 4 (PGG4) are above the WHO guideline value of 0.4 ppm and this is a reason for concern since manganese is known to cause adverse neurological effects following exposure from drinking water (WHO, 1996). The Cu, Pb, Cd, Zn and Ni concentrations of these waters were either below the detection limits of the Atomic Absorption Spectrometric method employed for the quantification of the metals or below the WHO guidelines values for the metals (WHO, 1993;WHO, 1996).

Only a few (fluoride, arsenic, nitrate and lead) of the chemical substances present in drinking water are known to cause widespread health effects in humans. Thus, the health risks posed by chemical substances are not as acute as those posed by microbial contaminants (WHO, 1993; WHO, 1996). Nevertheless, significant health effects may arise from exposure to the chemical constituents of water over a prolonged period. In this respect, it is noteworthy that substances, such as heavy metals, which have cumulative toxic properties, are not found at prohibitory levels in these water samples.

Ranking of water quality

To rank the water bodies and unearth patterns in the parameters that influence water quality, the data were subjected to PROMETHEE and GAIA analyses. The PROMETHEE II complete ranking results indicated that the net flow for the water bodies are 0.13, 0.07, 0.06, 0.03 and -0.02 respectively for PGS1, PGG2, PGS3, PGG5 and PGG4.. Thus, the most preferred source is PGS1 followed by PGG2, PGS3, PGG5 and PGG4 (in this order). The net outranking flow shows the spread of the objects (the water bodies) in such a way that the farther apart the outranking flows of any two water bodies, the larger the preference of the water body with the more positive outranking flow over that with the more negative flow.

Exploratory pattern recognition

In order to examine the variables that were most important in the ranking of the surface waters and groundwaters, Principal Component Analysis (PCA) of the 5 waters was performed with the aid of SIMCA P 10.0,Umetrics AB software. To minimise the skewness of the data (as a result of missing values) a constant number (100) was added to all variables and the data was log transformed and auto-scaled (mean-centered and scaled to unit variance) before PCA modeling. Approximately 72% of the variance is explained by the first two PCs. A close study of the scores plot displayed in Figure 1a reveals the following instructive details: The water samples were separated on the first principal component (denoted as t[1] in the SIMCA P-10 software). One cluster consisting exclusively of groundwater samples from a particular locality (PGG4, and PGG5) had negative t[1] scores while water samples from other locations (PGS1, PGG2, PGS3,) had positive t[1] scores (Figure 1a). The preliminary conclusion from these exploratory PCA was that the principal basis for the discrimination on t[1] is the geographic origin of the water supplies. It is well known that chemical constituents of water may arise from natural sources (e.g rocks and soils) as well as agricultural and industrial activities, which differ from one location to another (WHO, 1993; WHO, 1996).

The loadings plot (Figure 1b) showed that the first principal component loading vector (denoted as p[1] in the SIMCA P-10 software) has fairly large positive coefficients for NO_2^-, NO_3^- and NH_3-N, and relatively large negative coefficients for conductivity, Cl^-, Fe, SO_4^{2-} and PO_4^{3-}. The second principal component loadings vector (p[2]) has relatively large positive coefficients for Na and K, and large negative coefficients for conductivity. Thus, these are the dominant variables in ranking the water bodies. Five broad groups of such variables are apparent from the loadings plot in Figure 1b. Group A consisted of TDS, PO_4^{2-}, Cl^- and SO_4^{2-}, group B is made up of conductivity, Mg, F^- and Fe; group C consisted of Na and K ; group D is made up of NH_3-N, NO_2^- and NO_3^- and group E contained Ca, Zn, pH and Mn.

Exploratory comparison of international data

Overall PROMETHEE ranking: For the purpose of multivariate data analysis, the primary water quality data in Table 1 was treated as matrix 1, while additional data obtained from previous water quality studies carried out on water samples from Nigeria and Papua New Guinea (Okoye,1991; Olajire et al, 2001; Agbu, 1984 and Alaribe, 1984; Ibe and Njemanze, 1999; Salomons and Eagle, 1990) were treated as matrix 2**,** literature data on similar investigations conducted on Indian/Pakistani waters (Afzal et al, 2000**)** were treated as matrix 3 and data from Thai and Egyptian studies (Kruawal et al, 2005; El-dars, 2005) were treated as matrix 4. The water sources were compared because they represented examples of surface - and ground- waters qualities in developing countries where water resource planning and water quality assessment policy options are most urgently required. In order to compare the results of the multivariate analysis obtained in the current study with those from other similar water quality studies carried out in Nigeria and Papua New Guinea (matrix 2) and India/Pakistan (matrix 3) and Thailand and Egypt (matrix 4), the four matrices were combined into a single matrix, the variables were given equal weights and the matrix analysed by PROMETHEE. A complete PROMETHEE II outranking flow for the combined matrix is presented in Table 2. Sample NGS21 (surface water no.21 from Nigeria (Okoye, 1991) has the most positive outranking flow value. Therefore, its quality outranks those of all other water samples. Conversely, sample EGG52 has the lowest net outranking flow value and it is outranked by all other water samples. Generally, based on the physico-chemical properties of the water samples, most Nigerian water samples (Okoye, 1991; Ibe and Njemanze, 1999; Ekpo and Inyang, 2000) are among the best performing water bodies. The Indian/Pakistani water samples are generally amongst the medium and best performing waters, the Papua New Guinean water samples were generally medium performers and the Thai and Egyptian samples were among the least performing. Evidently, the qualities of the water samples are significantly influenced but not solely determined by their geographic origins. It is, however, noteworthy that variables such as PO_4^{3-}, SO_4^{2-}, NH_3-N, TDS, Mn, NO_2^-, Mg, Cl, Ca, Na, and K, which account for large data variances in the analysis of the combined matrix, contribute significantly to the ranking of the water bodies.

Selection of global key variables

Overall Principal Component Analysis : Next, using SIMCA P-10 software, we carried out PCA on matrices consisting of (i) matrix 1, (ie results from the current study) (ii) matrix 2 (results from published studies conducted in Nigeria (Okoye, 1991; Ibe and Njemanze, 1999;;Ekpo and Inyang, 2000) and Papua New Guinea (Salomons and Eagle, 1990) (iii) matrix 3 (data obtained from a study carried out by Afzal et al (Afzal et al, 2000) in

India/Pakistan, matrix 4 (data obtained from Kruawal et al, 2005; El-dars, 2005 studies) and (iv) a combination of matrices 1, 2, 3 and 4. When variables reported in these literature references were weighted equally, the most important and least important variables identified from each PCA are presented in Table 3. Of these variables TDS, Ca, SO_4^{2-} , Na and Cl appear among the most important variables in each of the matrices in Table 3. Therefore, they were selected as the most important variables influencing the water qualities. Interestingly, the scores plot for the combined matrix (Figure 2) showed that the water bodies from matrix 1 and most from matrix 2 (Agbu, 1984 and Alaribe, 1984 Okoye, 1991; Ibe and Njemanze, 1999; Ekpo and Inyang, 2000) had positive t[1] scores (Cluster C) while all of the India/Pakistan water bodies had negative t[1] and t[2] scores (Cluster A). However, some of the objects from matrix 2 (Salomons and Eagle, 1990) had negative t[1] and t[2] scores (Cluster A) while most of the objects from matrix 4 had positive t[2] but negative t[1] scores (Cluster B).

The key variables for prediction

Overview PLS model: The main conclusions from the above analyses are (i) the water bodies were separated to a large extent on t[1] and t[2] on the basis of their geographic origins, although this is not the only parameter influencing their qualities and (ii) TDS, Ca, SO_4^{2-} Na and Cl^- are the most important variables that influence the water qualities. Kettaneh et al (2005) have suggested that not all variables in a matrix are important. If there are N (26 in this case) variables in a matrix, they suggested that the salient feature of the matrix will be dominated by $\sqrt{N}$ variables (about 5 in this case). We therefore set out to test whether the five variables identified as the most important variables could be used as surrogates to predict physico-chemical properties of water in developing countries where, due to lack of the necessary infrastructure and expertise, it is not possible to carry out elaborate water quality studies. Thus, a PLS model in which TDS, Ca, SO_4^{2-} Na and Cl^- were used as X variables was developed for the combined data matrix (N = 57) yielding two significant components with the cumulative R^2X (sum of squares for the X-block) = 0. 84; cumulative R^2Y (sum of squares for the Y block) = 0.34 and cumulative Q^2 (fraction of the total variation of the X's that can be predicted by the components) = 0.30. According to Sun (Sun, 2004), Q^2 values equal or higher than 0.3 can be interpreted, Q^2 greater than 0.5 is associated with a good model while Q^2 is greater than 0.9 for an excellent model. Furthermore, the inner relationship of the Y-block PLS scores (denoted as u[1] in the SIMCA p-10 software) against the X-block scores (t[1]) was linear with the regression equation $y = x -7 \text{ x } 10^{-7,}$ $R^2 = 0.72$ and $N = 57$ at 95% confidence level. This indicates that the correlation between the Y block and X block is significant at 95% confidence level and that the X variables can be used to predict the Y variables encountered in this study.

To confirm the validity of the model, several parallel models in which the X (factors) data in the calibration set is kept constant and the Y (responses) data randomly reordered were developed [Eriksson et al, 2001]. New values of R^2Y and Q^2Y computed from the permuted Y data were then compared with the estimates of the R^2Y and Q^2Y from the parent PLS model in order to appraise the statistical significance of the latter values. In this work, the permutation procedure were repeated two hundred times and if every time lower R^2Y and Q^2Y values were obtained than those from of the original data, the significance of the "real" PLS model was confirmed (Eriksson, 2001). The result of response permutation obtained in the present work showed that the intercepts obtained for plot of R^2Y and Q^2Y (Y-axis) against the correlation coefficients between the permuted and original response variables (X-axis) for a model in which X = Total Dissolved Solid, Ca, SO_4^{2-}, Na and Cl^- ; Y= Conductivity and number of samples = 57 were $R^2 = 0.0$, 0.04 and $Q^2 = 0.0$, -0.06. Eriksson et al (2001) have shown that when $R^2 < 0.3$-0.4 and $Q^2 < 0.05$ the explanatory and predictive powers of the model are much higher than those obtained from randomly fitted Y data. Therefore the present model is valid.

When the water bodies with odd sample numbers (N =29) were used for model calibration and those with even sample numbers (N = 28) were used as validation sets, the correlation coefficients (R^2) of plots of the observed and predicted variable values together with the root mean square errors of prediction (RMSEP) are presented in Table 4. Given the facts that (i) the water bodies are from widely different origins, (ii) the studies were undertaken under different conditions and (iii) the limit of acceptable R^2 at 95% coefficient level is 0.28 for 50 samples (Minium et al, 1993), it is evident that variables like PO_4^{3-}, NH_3-N, Mg, NO_3^-, NO_2^-, Fe, F^-, Mn, As, Se and B can be confidently predicted. The root mean square of the errors of prediction (RMSEP), which might have arisen from (i) errors in the calibration set, (ii) errors in the prediction set, and (iii) errors in the prediction, are generally low (cf Eriksson, 2001). Similarly, the percentage absolute error (defined as 100 X (observed value-calculated value/observed value) is generally below 10% (Haus et al, 2003). Hortwitz (1982) suggested that at 1 ppm level, about 16% error is expected in the interlaboratory comparison measurements. Thus, the prediction efficiency of the model for these variables is comparable to the interlaboratory analysis experience.

Implications

The paper reported the use of PROMETHEE and GAIA procedures for the systematic interpretation of surface - and ground - waters quality in developing countries. PROMETHEE ranked the water bodies based on 26 water quality-influencing variables. Importantly, apart from the present study, the data used for the modeling and prediction was obtained from various investigations, indicating that globalization of information with the use of chemometrics is a very feasible approach for studying performance and prediction issues in water quality. Information from modest, cost-effective and affordable water quality studies may be brought together for meaningful exploratory (or large) investigations. In this exploratory study, the application of chemometrics techniques to a moderate size data set, consisting of globally diverse samples, has led to the extraction of information, which potentially could have significant benefits for lowering water analysis costs, especially in the developing countries. Patterns from PCA plots identified TDS, Ca, SO_4^{2-}, Na and Cl^- as the most important variables influencing the ranking of the water bodies. These variables were subsequently found (by PLS analysis) to be useful for modeling and predicting the levels of the other water pollutants. Available methods for monitoring the variables are comparatively cheap and it is palpable that the chemometrics procedures highlighted in this paper could (i) reduce the financial and opportunity cost associated with extensive monitoring of the chemical and physical qualities of surface water and groundwater and (ii) be used in developing countries to obtain reasonably good estimates of the levels of other pollutants in a water resource from the TDS, Ca, SO_4^{2-} Na and Cl^- contents. Although the modeling described did not produce satisfactory prediction for some water quality indicators ,considering the fact that data used for the modeling were obtained for investigations that were not carried out under identical conditions, the results provide a support for the viability of our concept – the possibility of using multivariate data analysis methods to predict water quality from a few easily measured variables. More work is required on this concept in order derive appropriate types of indicators from which water quality can be confidently predicted.

Acknowledgements

We acknowledge the assistance of Robin Totome during the experimental stage of this work.

References

Afzal, S., Ahmad, I., Younas, M., Din Zahid, M., Atique Khan, M. H., Ijaz, A. and Ali, K. 2000, Study of water quality of Hudiara drain, India-Pakistan, Environmental International, 26, 87-96

Agbu, A.A. 1984, Quality of Well Water in Samaru, MSc thesis, Ahmadu Bello University, Zaria, Nigeria

Akhter, M. S. 1998, Assessment of toxicity level of fluoride in underground waters used for irrigation in Bahrain, Environmental Toxicology and Water Quality, 13, 111-115.

Alaribe, H. C. 1984, Quality of Well Water in Old Zaria City, MSc thesis, Ahmadu Bello University, Zaria, Nigeria

Al-Shemmeri,T.; Al-Kloub, B. and Pearman, A. 1997, 'Model choice in multicriteria decision aid', *European Journal of Operational Research,* 97, 550-560.

APHA Standard methods for examination of water and wastewater; prepared and published jointly by: American Public Health Association (APHA)*;* American Water Works Association (AWWA), and Water Pollution Control Federation (WPCF), New York, 17 th Edition, 1989

Ayoko, G. A., Morawska, L., Kokot, S., and Gilbert, D. 2004, An application of multicriteria decision making methods to air quality in the microenvironments of residential houses in Brisbane, Australia, Environmental Science and Technology, 38, 2609-2616.

Brans, J. P.:1991 PROMCALC (Version 3:1), Centre for Statistics and Operations Research, Free University of Brussels, Brussels.

Brans, J.P.: 2002, Ethics and decision European Journal of Operational Research, 136, 340-352

Brans, J. P., Mareschal, B.: 1989 PROMETHEE-GAIA Visual Interactive Modelling for Multicriteria Location Problems, Internal Report University of Brussels, STOOTW/244

Brans, J. P., Mareschal, B.: 2005 PROMETHEE Methods in Figueira, J., Greco,S., and Ehrgott, M., Multiple Criteria Decision Analysis –State of the Art Surveys, 163-195, Springer, New York.

Decision Lab 2000 Executive edition, 1999, Getting started guide, Visual Decision Inc., Montreal, Canada,

Einax, J. W. Zwanziger, H. and Geiβ, S. 1997, Chemometrics in Environmental Analysis, VCH, Weinheim, pp 284-317 and references therein..

Ekpo, N. M. and Inyang, L. E. D. 2000, Radioactivity, Physical and Chemical Parameters of Underground and Surface Waters in Qua Iboe River Estuary, Nigeria, Environmental Monitoring and Assessment, 60, 47-55.

El-dars, F. M.S.2005, Evaluation of groundwater quality within a typical Egyptian village, North of Cairo, Egypt, Annali di Chimica, 95, 357-368.

Eriksson, L. Hagberg, P., Johansson, E., Rannar, S., Whelehan, O., Astrom, A., and Lindgren, T.2001, Multivariate process monitoring of newsprint mill. Application to modelling and predicting COD load resulting from de-inking of recycled paper, Journal of Chemometrics, 15, 337-352.

Espinasse, B., Picolet, G., and Chouraqui, E. 1997,Negotiation support systems: A multi-criteria and multi-agent approach', European Journal of Operational Research,103, 389-409.

Geldermann J. and Zhang K.: 2001, Software Review: Decision Lab 2000, Journal of multi-criteria Decision Analysis, 10, 317-323.

HACH Chemical Co. Conductivity/TDS Meter-Model 44600 Manual. Hach Chemical Company, Loveland, Colorado, USA, 1989 ed.

Haus, F; Boissel, O. and Junter G. A. 2003, Multiple regression modelling of mineral base oil biodegradability based on their physical properties and overall chemical composition, Chemosphere, 50, 939-948

Hortwitz, W. 1982, Evaluation of analytical methods used for regulation of foods and drungs, *A*nalytical Chemistry, 54, 67A-76A

Ibe, K. M.(Sr) and Njemanze, G. N., 1999, The impact of urbanization and protection of water in Owerri and environs SE, Nigeria. Environmental Monitoring and Assessment, 58, 337-348

IRC, Report: 1995, A developing Crisis, Water and Sanitation for all: A world Priority, Prepared by IRC International Water and Sanitation Centre, The Hague, Ministry, Spatial Planning and Environment.

Jior, R. S., Saxena, P. K., and Kondal, J. K., 1991, Impact of the Budha Nallah Brook on the river Satluj waters. I - Some physico-chemical parameters, International Journal of Environmental Studies, 39, 101-112.

Ketteneh, N., Berglund A.,and Wold S.: 2005, 'PCA and PLS with very large data sets', Computational Statistics & Data Analysis, 48, 69-85.

Kokot, S. and Phuong, T.D. 1999, Elemental content of Vietnamese rice Part 2. Multivariate data analysis, Analyst, 124, 561-569.

Kruawal, K. Sacher, F., Werner, A., Muller, J.and Knepper, T. P. 2005, Chemical water quality in Thailand and its impacts on the drinking water production in Thailand, Science of the Total Environment, 340, 57-70.

Massart, D. L., Vandeginste, B. G. M., Buydens, L. M. C., De Jong, S., Lewi, P. J.and Smeyers-verbeke, J.: 1997, 'Other Optimisation', Methods In Handbook of Chemometrics and Qualimetrics: *Part A,* Elsevier, Amsterdam, Chapter 26, pp 771-804.

Minium, E. W., Kingle, B. M. and Bear, G. 1993 Statistical Reasoning in Psychology and Education, New York, Wiley, 3rd Edition.

Okoye, B. C. O.1991, Nutrients and selected chemical analysis in the Lagos Lagoon surface waters, International Journal of Environmental Studies, 38, 131-136

Olajire, A. A.and Imeokparia, F. E. 2001, Water quality assessment of Osun river: Studies on inorganic nutrients, Environmental Monitoring and Assessment, 69, 17-28.

Ongley E. D. and Booty, W. G. 1999, Pollution remediation planning in developing countries: Conventional modeling versus knowledge based prediction Water International, 21, 31-38.

Pescod, M. B.1977, Surface water quality Criteria for tropical developing countries'-Editors Feacham, R., MacGarry, M. Mara, D. – Water, Wastes and health in hot climates, John Wiley & Sons, London 53.

Pescod, M. B.1992, Wastewater treatment and use in agriculture, Irrigation and drainage paper 47. Rome, Italy: Food and Agriculture Organisation of the United Nations.

Salomons, W., and Eagle, A.M. 1990, Hydrology, sedimentology and the fate and distribution of copper in mine related discharges in the fly river system, Papua New Guinea, *The* Science of the Total Environment, 97/98, 315-334.

Sun, H. 2004, A universal molecular descriptor system for prediction of Log P, Log S, LoB and absorption, Journal of Chemical Informatics and Computational Science, 44, 748-757

UNOCHA :1998, News in Brief, Water Research. Journal., June

Wold S.: 1978, Cross-validatory estimation of the number of components in factor and principal component models, Technometrics, 20, 397-405.

World Health Organisation. 1993, Guidelines for drinking water quality. 2nd Ed, Vol. 1. Recommendations. WHO, Geneva.

World Health Organisation 1996, Guidelines for drinking water quality, 2nd Ed, Vol. 2. Health criteria and other supporting information, Geneva, World Health Organization.

Table 1: Physico-chemical characteristics of the surface waters and groundwaters from Papua New Guinea*.

	Temp[a]	Cond[b]	pH[c]	TDS	SO_4^{2-}	PO_4^{3-}	Cl	F	NH_3-N	NO_3-	NO_2-	Na	K	Ca	Mg	Fe	Mn	Zn
PGS1	27.6	60	6.8	30	0.1	0.1	0.1	0.1	0.09	<0.1	<0.01	1.8	0.8	1.3	2	0.1	0.01	0.01
PGG2	27.3	120	6.8	60	3.0	0.1	0.1	0.2	<0.01	0.1	<0.01	3.2	1.8	0.6	4	0.4	0.2	0.02

PGS3	27.5	60	6.8	30	0.1	0.1	0.1	0.1	0.08	0.1	<0.01	8.0	13.5	1.2	2	0.1	0.04	0.01
PGG4	28.6	530	7.0	250	1.0	0.1	0.5	0.1	0.06	<0.1	<0.01	12.5	0.7	24	25	<0.01	1.2	0.01
PGG5	28.5	410	7.5	220	0.1	0.1	1.8	0.01	0.04	<0.01	<0.01	17.0	0.2	48	9	<0.01	0.04	0.09

* The Cu, Cd, Zn and Ni contents of the waters were generally below the detection limit of the atomic absorption spectrometric method used. [a] in ^{0}C; [b] in µs cm^{-1}; [c] in pH unit; all other measurements are in ppm; PGS1 = Papua New Guinea surface water 1; PGG2 = Papua New Guinea groundwater 2; PGS3 = Papua New Guinea surface water 3; PGG4 = Papua New Guinea groundwater 4; PGG5 = Papua New Guinea groundwater 5;

Table 2: PROMTHEE II complete ranking results for the water bodies from this study as well as other similar studies conducted in Nigeria, Egypt, Thailand, Papua New Guinea and India/Pakistan.

Object*	ϕ^+	ϕ^-	Net outranking flow (ϕ)	Rank	Origin*	Ref
NGS21	0.0385	0.0035	0.0350	1	NG	Ibe and Njemanze, 1999
NGS10	0.0384	0.0036	0.0348	2	NG	Ibe and Njemanze, 1999
NGS12	0.0382	0.0040	0.0342	3	NG	Ibe and Njemanze, 1999
PGS22	0.0382	0.0043	0.0340	4	NG	Salomons and Eagle, 1990
NGS20	0.0381	0.0044	0.0337	5	NG	Ibe and Njemanze, 1999
NGS11	0.0382	0.0048	0.0334	6	NG	Ibe and Njemanze, 1999
NGS9	0.0377	0.0045	0.0332	7	NG	Ibe and Njemanze, 1999
NGG6	0.0375	0.0043	0.0331	8	NG	Agbu, 1984
NGG7	0.0373	0.0045	0.0329	9	NG	Alaribe, 1984
NGS8	0.0374	0.0046	0.0327	10	NG	Ibe and Njemanze, 1999
NGS13	0.0373	0.0056	0.0316	11	NG	Ibe and Njemanze, 1999
PGS1	0.0358	0.0060	0.0298	12	PG	This work
NGS19	0.0335	0.0088	0.0247	13	NG	Ekpo and Inyang, 2000
IPG43	0.0238	0.0062	0.0177	14	IP	Afzal et al, 2000
IPG45	0.0238	0.0062	0.0177	15	IP	Afzal et al, 2000
EGG46	0.0238	0.0061	0.0177	16	EG	El-dars, 2005
NGS15	0.0245	0.0079	0.0166	17	NG	Ibe and Njemanze, 1999
NGG18	0.0228	0.0079	0.0149	18	NG	Ibe and Njemanze, 1999
NGS14	0.0222	0.0102	0.0120	19	NG	Ibe and Njemanze, 1999
PGG2	0.0254	0.0148	0.0106	20	PG	This work
NGG16	0.0203	0.0112	0.0091	21	NG	Ekpo and Inyang, 2000
PGS3	0.0258	0.0213	0.0045	22	PG	This work
IPG32	0.0164	0.0119	0.0045	23	IP	Afzal et al, 2000
IPG31	0.0163	0.0128	0.0035	24	IP	Afzal et al, 2000
PGG5	0.0234	0.0208	0.0026	25	PG	This work
IPG29	0.0142	0.0144	-0.0001	26	IP	Afzal et al, 2000
PGS23	0.0142	0.0145	-0.0003	27	PG	Salomons and Eagle, 1990
IPG40	0.0125	0.0159	-0.0034	28	IP	Afzal et al, 2000
IPG44	0.0126	0.0164	-0.0038	29	IP	Afzal et al, 2000
PGS24	0.0134	0.0179	-0.0045	30	NG	Salomons and Eagle, 2000
IPG39	0.0122	0.0169	-0.0047	31	IP	Afzal et al, 2000
EGG46	0.0121	0.0170	-0.0049	32	EG	El-dars, 2005
IPG42	0.0121	0.0171	-0.0050	33	IP	Afzal et al, 2000
PGG4	0.0157	0.0215	-0.0058	34	PG	This work
IPG37	0.0119	0.0179	-0.0060	35	IP	Afzal et al, 2000
IPG30	0.0110	0.0186	-0.0075	36	IP	Afzal et al, 2000
IPG36	0.0121	0.0199	-0.0078	37	IP	Afzal et al, 2000
IPG38	0.0107	0.0210	-0.0103	38	IP	Afzal et al, 2000
IPG28	0.0104	0.0213	-0.0108	39	IP	Afzal et al, 2000
IPG26	0.0102	0.0211	-0.0109	40	IP	Afzal et al, 2000
IPG27	0.0102	0.0211	-0.0109	41	IP	Afzal et al, 2000
IPG35	0.0098	0.0216	-0.0119	42	IP	Afzal et al, 2000
IPG34	0.0098	0.0220	-0.0123	43	IP	Afzal et al, 2000
IPG25	0.0099	0.0232	-0.0133	44	IP	Afzal et al, 2000
EGG49	0.0126	0.0263	-0.0137	45	EG	El-dars, 2005
NGG17	0.0158	0.0304	-0.0145	46	NG	Ekpo and Inynag, 2000
IPG33	0.0092	0.0239	-0.0147	47	IP	Afzal et al, 2000
EGG48	0.0121	0.0296	-0.0176	48	EG	El-dars, 2005

IPG41	0.0089	0.0281	-0.0192	49	IP	Afzal et al, 2000
EGG51	0.0112	0.0359	-0.0247	50	EG	El-dars, 2005
THS57	0.0079	0.0386	-0.0307	51	TH	Kruawal et al, 2005
EGG50	0.0055	0.0444	-0.0388	52	EG	El-dars, 2005
EGG54	0.0057	0.0450	-0.0392	53	EG	El-dars, 2005
THG56	0.0049	0.0488	-0.0439	54	TH	Kruawal et al, 2005
EGG53	0.0044	0.0503	-0.0458	55	EG	El-dars, 2005
EGG55	0.0036	0.0574	-0.0539	56	EG	El-dars, 2005
EGG52	0.0027	0.0660	-0.0633	57	EG	El-dars, 2005

*NG =Nigeria; PG = Papua New Guinea; IP= India/Pakistan; EG= Egypt; TH= Thailand; G =groundwater; S = surface water; and the suffix Arabic numeral the sample numbers.

Table 3: The PCA results for the water bodies from Nigeria, Egypt, Thailand, Papua New Guinea and India\Pakistan

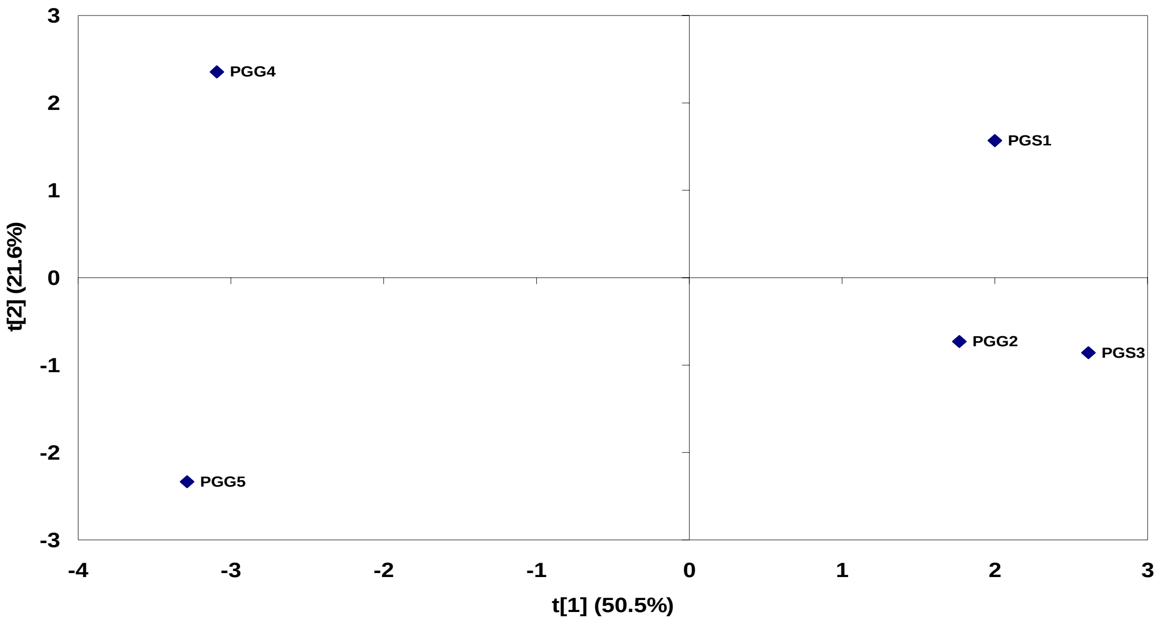
Matrix	Percent of variance accounted for by p[1] and p[2]	No of Objects	Most important variables (those with p[1] loadings >0.20 or <-0.20 or p[2] loadings >0.35 or < 0.35.)	Least Important variables (those with p[1] loadings >0.20 or <-0.20 or p[2] loadings >0.35 or < 0.35.)
This study	72	5	PO_4^{3-}, NH_3-N, SO_4^{2-}, pH, F^-, Na, Ca, Fe, K, Mg, TDS, Conductivity, NO_3^-, Cl^-, NO_2^-	Mn, NO_2^-, K, Zn
Other studies from Papua New Guinea and Nigeria	71	19	Temp, Cl^-, pH, TDS, SO_4^{2-}, NH_3-N, Zn, Na, K, Mg, NO_3^-, Ca,	F^-, NO_2^-, Fe, Mn, Li, B, PO_4^{3-}, NO_3^-, pH, conductivity.
A study from India/Pakistan	48	21	SO_4^{2-}, Ca, TDS,, pH, Na, Cl^-, K, NO_2^-, Li, NO_3-	Fe, B, Cr, Cd, As, P, Se, Hg
Egyptian and Thailand samples	87	12	Cl-, NO_3^-, Mg, PO_4^{3-}, NH_3-N, TDS, SO_4^{2-}, Ca, Na, K, Se, As	
Entire data	55	57	Temp, Cl-, B, NO_3^-, Mg, PO_4^{3-}, NH_3-N, TDS, Mn, F^-, SO_4^{2-}, pH, NO_2^-, Na, Ca, Li, Se,	Conductivity, Zn, Cd, Fe, As, Cr, K, P, Hg

Table 4: Correlation coefficients and errors of prediction for the validation set.

Y Variable	Correlation coefficient of observed Vs predicted value plot	Root mean square of root of prediction (RMSEP)
Temperature	0.045	0.032
Conductivity	0.63	0.29
pH	0.13	0.003
Phosphate	0.62	0.01
Fluoride	0.70	0.001
NH_3-N	0.60	0.08
NO_3^-	0.34	0.05
NO_2^-	0.69	0.001
K	0.19	0.028
Fe	0.67	0.04
Mn	0.60	0.003
Zn	0.54	0.002
Li	0.18	0.005
B	0.69	0.001
Mg	0.60	0.04
Cd	0.05	0.0000042
Cr	0.26	0.00005
P	0.22	0.0001
As	0.69	0.000034
Se	0.41	0.000035
Hg	0.21	0.000024

Figure 1: Scores (a) and loadings (b) plots for the water samples

(a)



(b)

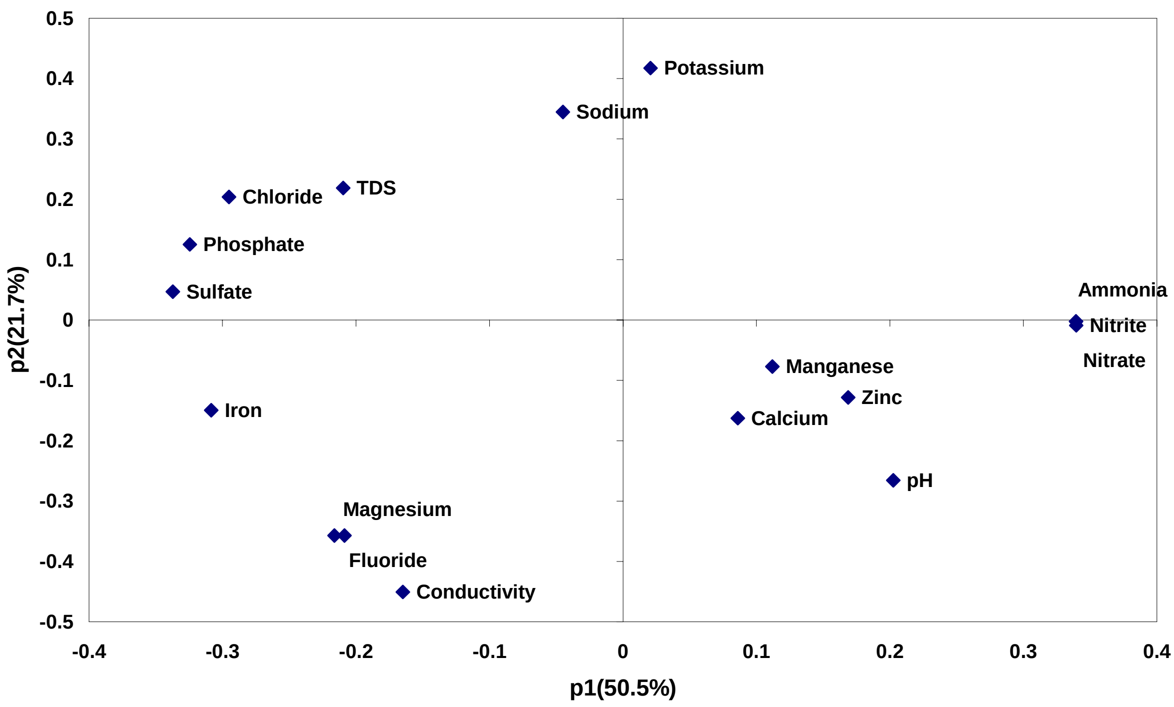


Figure 2: PCA Scores plot for the entire data matrix. Cluster A consists mainly of objects from India\Pakistan ; Cluster B contains most objects from the Egyptian and Thai studies and Cluster C contains many objects from the present and previous studies carried out in Papua New Guinea and Nigeria .