Global NEST Journal, Vol 13, No 3, pp 229-236, 2011

MULTIVARIATE ANALYSIS OF HYDROCHEMICAL DATA OF THE GROUNDWATER IN PARTS OF KARWAN – SENGAR SUB - BASIN, CENTRAL GANGA BASIN, INDIA

Taqveem Ali KHAN

Department of Geology
A.M.U., Aligarh, India

Received: 19/09/09
Accepted: 05/02/10

*to whom all correspondence should be addressed:
e-mail: taqveemk@yahoo.co.in

ABSTRACT

Ganga basin is one of the world's biggest aquifer repositories. The thick alluvium of the basin hosts its three tier aquifer system. The aquifer of the basin is under high stress due to unethical human intervention in the natural system. This warrants the need to evolve the basic hydrochemistry of every bit of the basin to make a scientific planning followed by a pragmatic execution. A multivariate statistical analysis was carried in order to give the hydrochemistry of the shallow aquifer a new dimension which is easily understood at a glance. In the present paper an attempt has been made to study the hydro chemical analysis data of shallow groundwater in parts of Karwan – Sengar sub basin, Central Ganga basin. The study is made of shallow aquifer of the region in which the movement of groundwater is from northwest to southeast. The descriptive statistical analysis was done beside Pearson correlation, principle component and regression analysis. All these are synthesized here to decipher the dynamics involved in the hydrochemistry of the area. The principle component analysis identified five factors that are responsible for the data structure explaining 83.49 % of the total variance of the data set. Factor 1 to 5 explains variance of 31.23, 19.445, 13.131, 12.105 and 8.647% respectively. Regression analysis show that Electric Conductivity (EC) as an independent variable which can be used to measure Carbonate (CO_3^{2-}), Chloride(Cl^-), Sodium (Na^+), and Total Dissolve Solids (TDS). Further Magnesium (Mg^{2+}) can be used to calculate the Total Hardness (TH) directly in the area.

KEYWORDS: Groundwater quality, Statistical analysis.

1. INTRODUCTION

The rise in population coupled with changing lifestyle has led to higher consumption of water for domestic, industrial, and irrigation purposes. The situation arisen has made it imperative to prevent and control water pollution and have reliable information on water quality for its effective management. The observations made through physico - chemical analysis needs to be interpreted in a rational manner to decipher the spatial and temporal variation in the hydrochemical data. 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 requires multivariate projections methods for simultaneous and systematic interpretation (Ayoko *et al*., 2007).Taking this into consideration the multivariate statistical techniques are used to interpret the water quality of the study area and to give meaningful results that were not possible while assessing the data at a glance. In the present study statistical software SPSS software version 10 is used to carry out the statistical analysis. Besides, Pearson correlation coefficient and Principal Component Analysis (PCA), Regression analysis was also performed.

Correlation coefficient is used to measure the strength of association between two continuous variables. This tells if the relation between the variables is positive or negative that is one increase with the increase of the other or one decreases with increase of the other. Thus, the correlation measures the observed co-variation. The most commonly used measure of correlation is Pearson's r. It is also called the linear correlation coefficient because r measures the linear association between two variables (Helsel and Hirsch, 2002). The data were statistically computed using correlation coefficient in order to indicate the sufficiency of one variable to predict the other (Davis, 1986).

Principal component analysis (PCA), a multivariate statistical technique, was initially developed as a tool in the social sciences but has proven quite effective in groundwater quality studies (Love and Hallbauer, 1998; Olmez *et al.*, 1994; Reghunath *et al.*, 2002; Subbarao *et al.*, 1995). The technique is used for data reduction and for deciphering patterns within large sets of data (Wold *et al.*, 1987 and Farnham *et al.*, 2003). The multivariate analysis is used in making the relationship between variables (water quality data). This technique aims to transform the observed variables to a set of variables, which are uncorrelated and arranged in decreasing order of importance. The principal aim is to simplify the problem and to find new variables (principal components), which make the data easier to understand. (Mazlum *et al.*, 1999). The result of these techniques helps the interpretation of the data. The numbers of factors, called principal components (PC), were defined according to the criterion that only factors that account for variance greater than 1 (eigenvalue- one criterion) should be included. The rational for this criterion is that any component should account for more variance than any single variable in the standardized test score space (Andrade *et al*., 2005). Simple linear regression analysis was performed to evaluate the statistically significant variables of the system. The variables shown significance in the correlation analysis are subject to regression analysis and predictive model for the same is prepared.

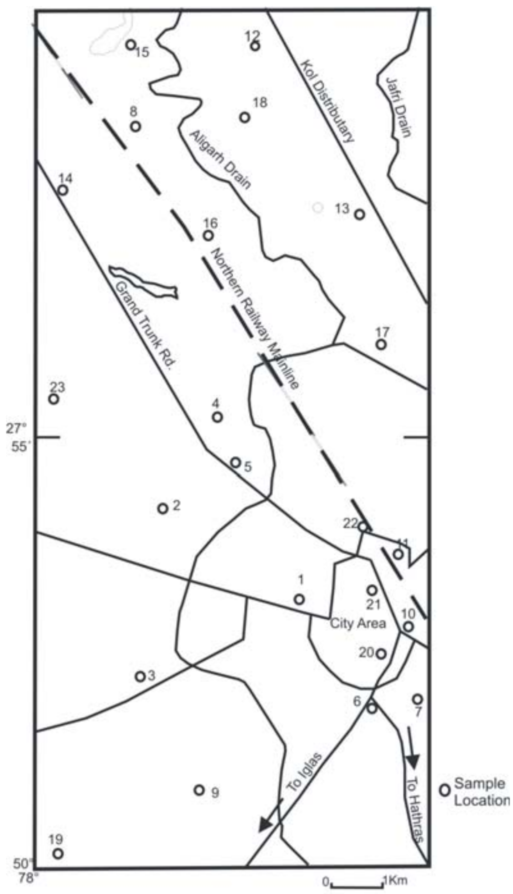

Figure 1. Showing the study area

2. MATERIAL AND METHODS.

2.1 Study Area

The study area is spread over 152 sq km. It lies between Karwan River in the west and Sengar River in the east and is a part of Central – Ganga basin. It lies between the latitude 27° 50' and 28° N and the longitude 78° and 78° 5' E (Figure 1). The central depression and western upland are two prominent physiographic units of the area. The NW-SE trending upland forms the eastern margin of the western upland and sub parallel to it lies the central depression due east. The level varies from

NW – SE with an average gradient of 0.26 m km^{-1}. Usually, the surface down to a depth of 20 to 25 cm is a well – drained soil and contains loose loam that can easily be cultivated. The pH of the soil ranges from 7 to 8. Iron and alumina remain constant, whereas, magnesia is less through out the area.

Table 1. Descriptive Statistics (Ions in mg L^{-1})

Variables	Minimum	Maximum	Mean	Std. Deviation
pH	7.1	8.6	7.82	0.3252
EC	379	1375	774.48	249.98
CO_3^{2-}	Nil	0.95	0.50	0.21
HCO_3^-	2.2	14.5	9.22	2.44
Cl^-	0.36	4.28	1.94	1.09
SO_4^{2-}	0.52	4.75	2.30	1.09
Na^+	0.973	10.661	6.46	2.71
K^+	0.273	3.687	0.90	0.90
Ca^{2+}	1.157	8.483	4.19	1.78
Mg^{2+}	1.282	2.379	1.73	0.30
TH	180.	504.0	268.26	74.33
TDS	243.	881.	496.43	160.22
SAR*	0.572	8.54	4.02	2.08

* Sodium Adsorption Ratio

The area falls under sub – tropical climatic zone and is characterized by hot summer and chilly winter. During summer the temperature shoots up to 47°C and in winter some time temperature falls to 2°C. The monsoon normally breaks in the second week of June and ends in September. Heavy precipitation takes place in the months of July and August. The area on an average receives 760mm of rainfall per year. The hydrochemical analysis data from 23 locations in the area was used in the present study. The summary of the results is given in the Table 1.

2.2 Synopsis of Geology and Hydrogeology

The Ganga basin is one of the largest groundwater basins of the world. It is located between the northern fringe of Indian Peninsula and Himalayas and extends from Delhi Hardiwar ridge in the west to Monghyr – Saharsa ridge in the east. The study area forms the part of this vast basin. In the study area the bed rock encountered at a depth of 340 meter below ground level (m.b.g.l) is upper Bhander red shale of upper Vindhyan group of Proterozoic age which is further overlain by quaternary alluvium. The river Ganga and its various tributaries derived from the newly risen Himalayas and also from the northern fringe of the peninsula deposited the quaternary sediments on the eroded surface of the upper Vindhyans.

Hydrogeologically speaking there occur three to four tier aquifer systems. Aquifers seem to merge with each other, thus, develop a single bodied aquifer. The granular zones comprise 40 – 50 percent of the total formations encountered at various depths. In the southeast the clay formation attains considerable thickness and predominance of the clay to the granular zones form 50% of the total litho units encountered. However, the clay beds pinch out laterally. The shallow aquifers in the area mainly comprise fine to medium sands and vary in thickness from 3 to 26 meters. The groundwater occurs in these aquifers under phreatic condition. Due to excessive withdrawal of water from these aquifers, they are highly strained. The discharge of these wells varies from 30 to 50 m h^{-1} at a nominal drawdown of 3 to 4.5 meters. The elevation of water table ranges between 179 meters (m) in the northwest to 171 m in the southeast above the mean sea level. The general flow of groundwater is northwest to southeast in consonance with the over all trend of groundwater flow in the Ganga basin save minor alteration that are governed by local lithologic and anthropogenic factors (Khan and Ahmad, 2002).

2.3 Methods of chemical parameters determination

The samples were collected so as to cover the entire area and from the hand pumps used for drinking purpose. All the physico chemical parameters were determined by the standard methods (APHA, 1975; Trivedi and Goel, 1984). The Cl^-, CO_3^{2-} and HCO_3^-, were analyzed by volumetric method and SO_4^{2-}, by turbedemetric method. The concentration of other major elements was done by atomic absorption spectrophotometer.

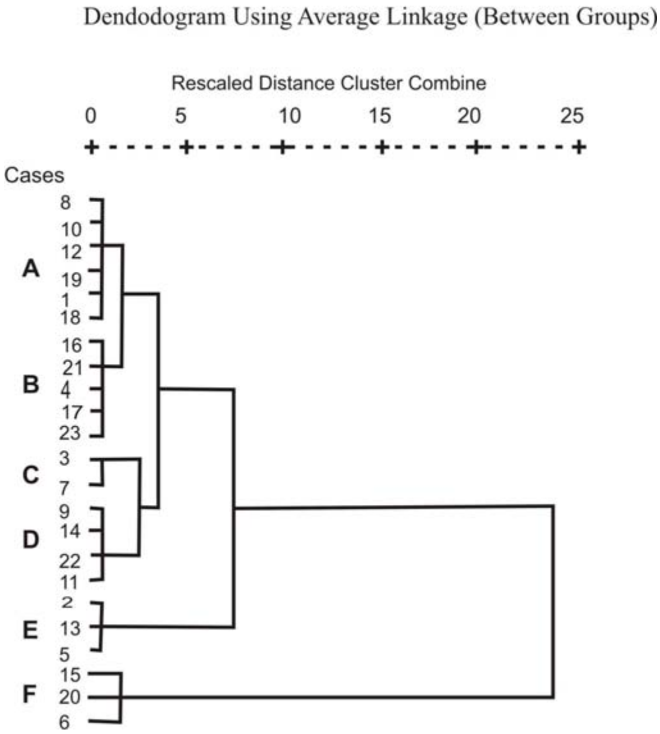

Figure 2. Dendogram using average linkage (between groups)

Table 2. Cluster groups and their members

Group	Members (Location/sample No.)
A	8, 10, 12, 19, 1, 18
B	16, 21, 4, 23, 17
C	3, 7
D	9, 14, 22, 11
E	2, 13, 5
F	15, 20, 6

3. Results and discussion

3.1 Q- mode cluster analysis

Hierarchical cluster analysis is a powerful tool for analyzing water chemistry data (Seyhan *et al.*, 1985; Reeve *et al.*, 1996; Ochsenkuhn *et al.*, 1997) and has been used to formulate geochemical models (Meng and Maynard, 2001). It is an exploratory data analysis tool used to sort out different objects into groups. In clustering the objects are grouped such that the similar objects fall into the same class (Danielsson *et al.*, 1999). The degree of association between two objects is maximal if they belong to the same group and minimal otherwise. Hierarchical clustering joins the most similar observations and then successfully the next most similar observation. The levels of similarity at which observations are merged are used to construct a dendogram (Chen, 2007). The Euclidean distance is represented on the horizontal axis of the dendogram. It gives the similarity between two clusters. The weighted pair group method was used and the Euclidean distance was selected as the measure of similarity (Khan, 2008). Khan (2008) carried out Hierarchical Cluster Analysis of hydrochemical variables of the study area. The results of the cluster analysis are presented in Figure 2. The dataset were classified in six groups named as A, B, C, D, E, and F. A contains six samples and F contains only three. Clusters of samples are listed in Table 2, which indicate that each cluster has a water quality of its own which is different from the other clusters. Group A consist of the samples from location no 8, 10, 12, 19, 1, and 18 (Figure 1). The values of EC, HCO_3^-, TH and TDS are in a narrow range. The water of the group is low in K^+. The group B has EC, TH, and TDS in close range. The water type of the area is dominated by pH, HCO_3^- and K^+, while the TDS is the lowest. The group C is dominated by TH and Mg^{2+} and has low HCO_3^-, SO_4^{2-} concentrations. However K^+ dominates the group D. The group E is deficient of almost all elements i.e., EC, CO_3^{2-}, Cl^-, Na^+, K^+, TDS, and Sodium Adsorption Ratio (SAR) all lowest among the area. Only SO_4^{2-} is high. This group F is quite distinct from the rest of groups. This is evident in the visual interpretation of the dendogram. This cluster traverses a large distance before joining the rest of the cluster group. The EC, CO_3^{2-}, Cl^-, SO_4^{2-}, Na^+, TH, and TDS dominate the water type of this group. The pH is low.

3.2 Pearson Correlation Coefficients

The close inspection of correlation matrix was useful because it can point out associations between variables that can show the overall coherence of the data set and indicate the participation of the individual chemical parameters in several influence factors, a fact which commonly occurred in hydrochemistry (Helena *et al.*, 2000). The Pearson correlation coefficient matrix is given in the Table 3. The variables having coefficient value (r) >0.5 are considered significant. Inspection of the table reveals that EC is positively related with Na^+, Cl^-, CO_3^{2-} and TDS.

The same matrix gives the maximum variance as shown in the principal component analysis-factor 1. This further substantiates the significance of the analysis. HCO_3^- shows no correlation either positive or negative with any variable. Cl^- is related to Na^+, Mg^{2+} and TDS. SO_4^{2-} in the groundwater of the area shows no affinity with any variable. The same is significant from the PCA in which it forms the fifth factor and has least variance. Na^+ associates itself with TDS and SAR. K^+ shows no significant correlation. Ca^{2+} has negative affinity with SAR. This is the sole case in the present Pearson matrix. Finally Mg^{2+} shows positive relation with TH. The variation in relationship indicates the complexity of the quality of groundwater. And further depicts the effect of rock – water interaction.

Table 3. Pearson correlation

	pH	EC	CO_3	HCO_3	CL	SO_4	NA	K	CA	MG	TH	TDS	SAR
pH	1.000												
EC	-.277	1.000											
CO_3^{2-}	-.025	.541	1.000										
HCO_3^-	-.189	-.093	-.041	1.000									
Cl^-	.052	.547	.062	-.017	1.000								
SO_4^{2-}	.012	.186	.155	.194	.158	1.000							
Na^+	.444	.528	.387	.014	.540	.151	1.000						
K^+	-.240	.004	.018	.378	.081	-.102	.042	1.000					
Ca^{2+}	-.300	-.096	-.161	.236	-.081	.010	-.481	-.204	1.000				
Mg^{2+}	.095	.293	.223	-.101	.613	-.178	.160	.189	.054	1.000			
TH	-.177	.154	.192	.095	.470	-.186	-.090	.271	.164	.842	1.000		
TDS	-.277	1.000	.541	-.093	.547	.186	.528	.003	-.097	.293	.154	1.000	
SAR	.423	.470	.376	-.082	.363	.156	.926	.084	-.686	.004	-.224	.470	1.000

3.3 Principal component analysis

In all the Principal component analysis generated five significant factors (Table 4). These factors explain 83.49 % of variance. Each factor consists of variable with eigen value more than 1. The factors are given in descending order depending upon the variance. The factor having highest variance is assigned number 1 position and with least variance is given the fifth place. Factor 1, accounting for about 31% of the total variance, provides information about EC CO_3^{2-}, Cl^-, Na^+, TDS, and SAR. TDS is associated with Na^+ and Cl^- in this factor, which suggest that both mixing and water-rock interaction are responsible for the salinity of groundwater. The case is true as the area consist of alluvium derived by the fluvial agencies from the newly risen Himalayas. Salt content of groundwater strongly affects the taste of drinking water. Concentrations of less than 600 mg L^{-1} are considered permissible, whereas are unpleasant if higher than 1200 mg L^{-1}. However, the salt content of the area is within the permissible limits. Factor 2 accounts for 19 % variance and clusters Ca^{2+}, Mg^{2+}, and TH with positive loading and pH and SAR with negative loading. Mg^{2+} is a significant variable in this factor, which happens to be one of the major ions in the hydrosphere and the most abundant divalent cation in the biosphere. It is an essential element for both plants and animals. This factor considered to be a TH factor and provides information about hardness variability in the groundwater of the area. It is well known that the TH is connected to Ca^{2+} and Mg^{2+} content of the water. The clustering of the present variables further explains the dissolution of soils and mineral in the sediments containing groundwater. Factor 3 accounts for 13 % of variance and pH and Mg^{2+} are the only variables grouped in this factor. Mg^{2+} once again makes impact in this factor suggesting the dominance of the variable in the area. Factor 4 contents variable HCO_3^- and K and shows the 12% variance. K is the main constituent of soluble fertilizers and could come from livestock excrement (Conrad *et al.*, 1999). Fifth factor explains only 8% of variance and is solely influenced by SO_4 suggest the signature of livestock excrement.

3.4 Simple Linear Regression Analysis

Linear regression analysis is an important tool for the statistical analysis of water resources data. It is used to describe the covariation between some variable of interest and one or more other variables. Regression analysis is performed to estimate or predict values of one variable based on knowledge of another variable, for which more data are available. Values of r^2 close to 1 are often incorrectly deemed an indicator of a good model. An r^2 near 1 can result from a poor regression model; lower r^2 models may often be preferable (Helsel and Hirsh, 2002).

Table 4. Showing result of principal component analysis

	Components				
	1	2	3	4	5
pH	.123	-.564	.616	-.12	.363
EC	.84	.199	-.414	-.157	-.118
CO_3^{2-}	.592	4.624E-02	-.242	-6.263E-02	-.252
HCO_3^-	-.101	.207	-.237	.761	.345
Cl^-	.705	.332	.217	-3.273E-02	.378
SO_4^{2-}	.198	-.178	-.484	.133	.605
Na^+	.834	-.396	.139	.144	.145
K^+	.105	.283	.154	.802	-.315
Ca^{2+}	-.413	.543	-.322	-.234	.375
Mg^{2+}	.45	.652	.524	-.129	9.781E-02
TH	.237	.821	.393	2.256E-02	3.530E-02
TDS	.84	.199	-.414	-.157	-.118
SAR	.775	-.560	.114	.183	-4.363E-02
% of variance	31.23	19.445	13.131	12.105	8.647
Eigen Values	4.060	2.528	1.707	1.435	1.124
Cumulative %	31.23	50.675	63.806	74.844	83.491

A positive correlation between EC and CO_3^{2-}, Na^+, Cl^-, TDS and between TH and magnesium is used to carry out the regression analysis (Table 5). The model for simple linear regression is

$Y = \beta_0 + \beta_1 X$

Where, Y is the dependent variable

X is the independent variable

β_0 is the intercept which is the coefficient of regression

β_1 is the slope

Table 5. Regression equations for various water quality parameters

CO_3^{2-}= 0.143 +.0005EC

Cl^- = 0.09+.0024EC

Na^+ = 2.021+.006EC

TDS= 0.025+.641EC

TH=-90.75+207.191MG

The results of the analysis show that by measuring the EC value the other variables that is CO_3^{2-}, Na^+, Cl^-, and TDS can be calculated. TDS and EC show a perfect relation that is clear from high r^2 value. TH can directly be calculated from magnesium with the given equation. Though the confidence level is not much high but the analysis helps draw conclusions on the correlation of the variables.

4. CONCLUSION

The present study suggests that the multivariate statistical techniques help in identifying the relationship between the variables that is difficult to get at first glance. The pearson correlation coefficient simplifies the complexity of hydrochemical data and show the extent of dependence of one variable on the other. The PCA of the hydrochemical data reduces the original data matrix into five components that explains 83.49 % of the total variance. The regression analysis confirms the positive relation of EC with CO_3^{2-}, Cl^-, Na^+ and Total Dissolve Solid and of Mg and Total Hardness in the study area. The results depict the rock water interaction in this part of the Central Ganga – basin. The results further substantiate the usefulness of the multivariate analysis in hydrochemical studies of the groundwater.

REFERENCES

Andrade E.M., Palácio H.A.Q., Crisóstomo L.A., Souza I.H. and Teixeira A.S. (2005), Índice de qualidade de água, uma proposta para o vale do rio Trussu, *Ceará. Rev. Ciênc. Agron.,* **36**(2), 135–142 (in Portuguese).

American Public Health Association (1975), Standard methods for examination of water and waste water, 14th edition, APHA, Washington, D.C.

Ayoko G.A., Singh K., Balerea S. and Kokot S. (2007), Exploratory multivariate modeling and prediction of the physico-chemical properties of surface water and groundwater, *Journal of Hydrology,* **336**, 115– 124.

Chen K., Jiao J.J., Huang J. and Huang R. (2007), Multivariate statistical evaluation of trace elements in groundwater in a coastal area in Shenzhen, China, *Environmental Pollution*, **147**(3), 771-780.

Conrad J.E., Colvin C., Sililo O., Gorgens A., Weaver J. and Reinhardt C. (1999), Assessment of the impact of agricultural practices on the quality of groundwater resources in South Africa. Water Research Commission Report 641/1/99.

Danielsson A., Cato I., Carman R. and Rahm L. (1999), Spatial clustering of metals in the sediments of the Skagerrak/Kattegat, *Applied Geochemistry,* **14**, 689–706.

Davis J.C. (1986) Statistics and data analysis in geology, John Willey and Sons, New York, 2nd ed.

Farnham I.M., Johannesson K.H., Singh A.K., Hodge V.F. and Stetzenbach K.J. (2003), Factor analytical approaches for evaluating groundwater trace element chemistry data, *Analytica Chimica Acta,* **490**, 123–138.

Helena B., Pardo R., Vega M., Barrado E., Fernandez J. and Fernandez L. (2000), Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga River, Spain) by principal component analysis, *Water Res,* **34**(3), 807–816.

Helsel D.R. and Hirsch R.M. (2002), Statistical Methods in Water Resources Chapter A3, USGS, (http://water.usgs.gov/pubs/twri/twri4a3/).

Khan T.A. (2008), Cluster analysis and quality assessment of groundwater in and around Aligarh city, U.P. India., Proceedings of All India Seminar on Advances in Environmental Science and Technology, Z. H. College of Engg. and Technology, A.M.U, Aligarh, India. 109-114.

Khan T.A. and Ahmad M.S. (2002), Micro – level Hydrogeological studies in parts of Central Ganga Basin – Aligarh City, U.P., India, *Asian Profile,* **30**(3), 221-232.

Love D. and Hallbauer D.K. (2000), Groundwater chemistry trends and possible interventions at a southern African iron ore mine, *Journal of African Earth Sciences,* **31**, 41–42.

Mazlum N., Ozer A. and Mazlum S. (1999), Interpretation of water quality data by principal component analysis, *J. of Engineering and Environmental Science,* **23** ,19-26.

Meng S.X. and Maynard J.B. (2001), Use of statistical analysis to formulate conceptual models of geochemical behavior: Water chemical data from the Botucata Aquifer in Sao Paulo state, Brazil, *Journal of Hydrology,* **250**, 78–97.

Ochsenkuehn K.M., Kontoyannakos J. and Ochsenkuehn P.M. (1997), A new approach to a hydrochemical study of ground water flow, *Journal of Hydrology,* **194**(1–4) 64–75.

Olmez I., Beal J.W. and Vilaaume J.F. (1994), A new approach to understanding multiple-source groundwater contamination: factor analysis and chemical mass balances, *Water Research*, **28**, 1095–1101.

Reeve A.S., Siegel D.I. and Glaser P.H. (1996), Geochemical controls on peat land pore water from the Hudson Bay Lowland: A multivariate statistical approach, *Journal of Hydrology,* **181**(1–4), 285–304.

Reghunath R., Murthy, T.R.S. and Raghavan B.R. (2002), The utility of multivariate statistical techniques in hydrogeochemical studies: an example from Karnataka, India, *Water Research,* **36**, 2437–2442.

Seyhan E., van-de-Griend A.A. and Engelen G.B. (1985), Multivariate analysis and interpretation of the hydrochemistry of a dolomitic reef aquifer, northern Italy, *Water Resources Research,* **21**(7), 1010–1024.

Subbarao C., Subbarao N.V. and Chandu S.N. (1995), Characterisation of groundwater contamination using factor analysis, Environmental Geology, **28**, 175–180.

Trivedi R.K. and Goel P.K. (1984), Chemical and biological methods for water pollution studies, Environment Publication, Karad, India, 215.

Wold S., Esbensen K. and Geladi P. (1987), Principal component analysis, *Chemometric and Intelligent Laboratory Systems*, **2**, 37–52.