International Journal of ChemTech Research
CODEN (USA): IJCRGG, ISSN: 0974-4290, ISSN(Online):2455-9555
Vol.10 No.5, pp 688-698, 2017

Application of Multivariate Statistical Techniques in the Assessment of Groundwater Quality in Chrompet Industrial Area, Chennai

V. Karpagam

B.S. Abdur rahman university, Vandalurd-48, India

Abstract : Multivariate statistical techniques such as factor analysis (FA) and cluster analysis (CA) were applied for the evaluation of spatial variations and the interpretation of a large complex water quality data set of Chrompet in Chennai, Tamil Nadu. 11 parameters of water samples collected from 16 different sampling stations of city was determined. Factor analysis indicates three factors in pre monsoon and four factors in post monsoon, which explained 76.106% and 84.064% respectively of the total variance in water quality data set. Hierarchical cluster analysis grouped 16 sampling stations of city into two clusters, i.e., moderately polluted (MP) and highly polluted (HP) sites, based on the similarity of water quality characteristics. This study illustrates the benefit of multivariate statistical techniques for interpreting complex data sets in the analysis of spatial variations in water quality, and to plan for future studies.
Keywords : Factor analysis; cluster analysis; ground water; Chrompet.

Introduction and Experimental

Water is called matrix of life because it is an essential part of all living systems and is the medium from which life evolved and in which life exists [1]. The quality as well as the quantity of clean water supply is of vital significance for the welfare of humanity [2]. Polluted water is a source of many diseases for human beings [3]. Groundwater is the major source of drinking water in both urban and rural areas. Ground water is also frequently used as the alternative source for agricultural and industrial sector [4]. Distribution of groundwater quality parameters is controlled by complex processes. Ground water typically has a large range of chemical composition [5]. The ground water quality depends not only on natural factors such as the lithology of the aquifer, the quality of recharge water and the type of interaction between water and aquifer, but also on human activities, which can alter these ground water systems either by polluting them or by changing the hydrological cycle [6]. Water quality monitoring has one of the highest priorities in environmental protection policy. A monitoring program that provides a representative and reliable estimation of the quality of ground waters has become an importantnecessity. Consequently, comprehensive monitoring programs that include frequent water sampling at numerous sites and that consists a full analysis of a large number of physicochemical parameters are to be designed for proper management of water quality in ground waters. Real hydrological data are mostly noisy, it means that they are not normally distributed, often collinear or auto correlated, containing outliers or errors etc. In of order to avoid this problem multivariate methods such as the factor analysis and cluster analysis were used. The application multivariable statistical methods offer a better understanding of water quality for interpreting the complicated data sets.

Salient features of the study area

The southern suburbs of Chennai that are involved in the tanning processes involve areas of Pallavaram, Pammal, Nagalkeny and Chromepet. These areas include large scale, small scale and also cottage industries that promote tanning process. Both vegetable and mineral tanning processes are practiced. This study is to be carried out in Chromepet area, located in the southern part of Chennai city is shown in Fig1, which serves as a home town to a large number of tanning industries. The study area is 13 km away from Bay of Bengal and 20 kilometers from Chennai city. It is located in12°55'16"N to 12°57'56"N longitude and 80°8'19"E to 80°9'59"E latitude. In general, the climate of the area is with low humidity and high temperature. During winter the temperature is around 20°C. During the summer season the temperature increases up to a maximum of 44°C. The southwest monsoon from June to September contributes nearly 40% of the annual rainfall, which is about 1200 mm. The northeast monsoon is more important as it contributes to more than 60% of the annual rainfall from October to December. Topographically this region gently slopes towards west and east. The charnockite rocks of Archaean age occur as a basement rock in this area. Weathered charnockite rock occurs at the depth from 2 m to 4 m from the ground surface. The weathered rock is overlaid by soil of thickness ranging from 2 m to 4 m.

Physicochemical analysis of drinking water

The collected samples were analyzed for different physicochemical parameters such as pH, electrical conductivity (EC), Total dissolved solids (TDS), total hardness (TH), Bicarbonate (HCO_3), sulfate, fluoride, chloride, sodium, Potassium, calcium, magnesium, Fluoride (F-) and zinc according to the standard methods (Table 1)

Table 1 Analytical methods adopted for physicochemical analysis

Sl. No	Analysis	Method/instrument
1	pH and Electrical conductivity (EC)	Field testing kit
2	Calcium hardness (Ca)	EDTA Titrimetry
3	Magnesium hardness (Mg)	EDTA Titrimetry
4	Sodium (Na) & Potassium (K)	Flame photometer
5	Bicarbonates + carbonates	Titrimetry
6	Chlorides (Cl)	Mohr'sTitrimetry
7	Sulphates (SO_4)	Spectrophotometry
8	Zinc (Zn), Chromium(Cr), Copper (Cu) and Fluoride (F-)	Atomic Absorption Spectrometer

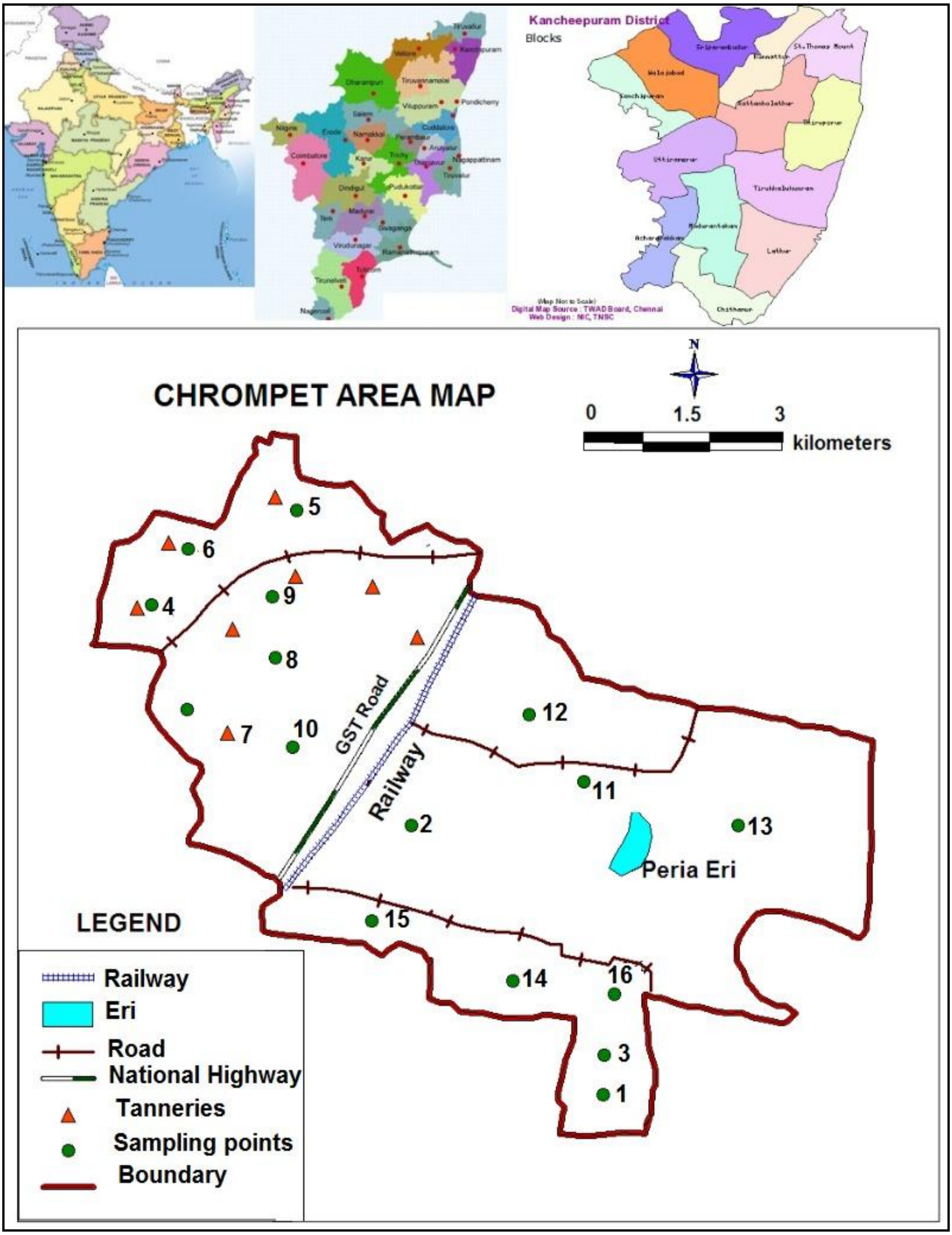

Fig 1 Study area location

Brief review of two multivariate statistical techniques used in this study

Multivariate statistical techniques can help to simplify and organize large data sets to provide meaningful insight. In the present study, two multivariate statistical techniques were used to Evaluate physicochemical parameters of groundwater samples. The statistical software package SPSS 17 was used for the multivariate statistical analysis

Principal Component Analysis

Principal Component Analysis reduces the multi-dimensionality of a complex data set to two or three dimensions by computing principal components or factors. This computation is achieved by transforming the observations from each sample (e.g. concentrations of parameters) into a "linear combination" of parameter concentrations. Principal Component Analysis produces several important outputs of which two namely eigenvalues*:* the variances accounted for by the component; and eigenvectors: that specify the directions of the PCA axes were considered in the analysis.

Factor analysis

Factor analysis (FA) is a very powerful technique applied to reduce the dimensionality of a dataset consisting of a large number of interrelated variables, while retaining as much as possible the variability presented in dataset and with a minimum loss of information [1]. This reduction is achieved by transforming the dataset into a new set of variables. Factors, which are orthogonal (non-correlated) and are arranged in decreasing order of importance. FA can also be used to generate hypotheses regarding causal mechanisms or to screen variables for subsequent analysis.

FA can be expressed as:

$$\mathbf{F_i = a_1\, x_1\, j + a_2\, x_2\, j + \ldots + a_m\, x_m}$$

Where F_i = factor
a = loading
x = measured value of variable
i = factor number
j = sample number
m = total number of variables

There are three basic steps to factor analysis:

1. Computation of the correlation matrix for all variables.
2. Extraction of initial factors.
3. Rotation of the extracted factors to a terminal solution.

Cluster analysis

It is a class of techniques used to classify cases into groups that are relatively homogeneous within themselves and heterogeneous between each other, on the basis of a defined set of variables. These groups are called clusters. Cluster analysis is a major technique for classifying a 'mountain' of information into manageable meaningful files. It is a data reduction tool that creates subgroups that are more manageable than individual datum. In cluster analysis there is no prior knowledge about which elements belong to which clusters. The grouping or clusters are defined through an analysis of the data (1). Cluster analysis was used to determine if the samples can be grouped into statistically distinct hydro chemical groups that may be significant in the geologic context. A number of studies used this technique to successfully classify water samples (7). Comparisons based on multiple parameters from different samples were made and the samples were grouped according to their 'similarity' to each other. Classifications of samples according to their parameters are known as Q-mode classifications. In the present study Q-mode HCA was used to classify the samples into distinct hydro chemical groups. The Ward's linkage method (Ward, 1963) was used in this analysis. A classification scheme using Euclidean distance (straight line distance between two points in *c*-dimensional space defined by *c* variables) for similarity measurement, together with Ward's method for linkage, produces the most distinctive groups where each member within the group is more similar to its fellow members than to any member outside the group. All 11 hydro chemical variables measured (consisting of EC, pH, Ca, Mg, Na, K, Cl, SO4, HCO3, TDS and TH) were utilized in this analysis. For statistical analysis, all the variables were log-transformed and more closely correspond to normally distributed data. Hydro chemical results of all samples were statistically analyzed

Results and Discussion

The physiochemical parameters of both pre monsoon and post monsoon seasons were calculated and compared with Bureau of Indian Standard as shown in table 1.

Table 1 Physiochemical parameters of Pre monsoon and post monsoon

Ion	Minimum concentration		Maximum concentration		Mean concentration		BIS (IS 10500: 2012)
	Pre monsoon	Post monsoon	Pre monsoon	Post monsoon	Pre monsoon	Post monsoon	
pH	7.1	7.2	7.9	8.8	7.481	8	6.5-8.5
EC (μScm^{-1})	1240	1060	2990	2140	1909	1600	1500-3000
Ca (mg/l)	50	100	154	444	91.63	272	75-200
Mg (mg/l)	40	96	96	310	73.50	203	30-100
Cl (mg/l)	184	137	511	347	344	242	250-1000
SO_4 (mg/l)	14.7	16	39.55	49	27.544	32.5	200-400
Na (mg/l)	17	112	3144	2396	1051	1254	200
K (mg/l)	0.5	0	103.4	24.2	15	12.1	-
HCO_3 (mg/l)	311	324	580	651	458	487.5	30-600
TH (mg/l)	372	600	878	2140	531	1370	200-600
TDS (mg/l)	672	678	1913.6	1325	1222	1001.5	500-2000
Cr (mg/l)	0.02	0.04	0.21	0.3	0.115	0.17	0.05
Cu (mg/l)	0.05	0.07	0.2	0.25	0.125	0.16	0.05- 1.5
Zn (mg/l)	0.02	0.03	0.36	0.42	0.19	0.32	5-15
F (mg/l)	0.13	0.1	1.13	1.23	0.63	0.68	1-1.5

Statistical analysis

Factor analysis for pre-monsoon

Totally 16 water samples were collected from Chrompet and 11 physicochemical parameters were determined. This water quality data was analyzed by using factor analysis (FA). Before conducting FA, the Kaiser-Meyer-Olkin (KMO) and Bartlett's sphericity test were performed on the parameter correlation matrix to examine the validity of FA. FA was conducted for all samples and it may be useful in providing significant reductions in dimensionality.

From data, 3 factors, explaining 76.106% of the total variance, was estimated on the basis of Kaiser Criterion of the eigenvalues greater or equal 1 and from a Cattle scree plot [1]. A scree plot shows the eigenvalues sorted from large to small as a function of the factor number. After the third factor (Fig. 1), starting the elbow in the downward curve, other components can be omitted. Factor was extracted by principal component method and rotated by Varimax. The factor loading, their eigenvalues, and variances are summarized in (Table 4).

Table 4 Components of pre monsoon water quality data

Parameters	Component		
	1	2	3
TDS	**.954**		
EC	**.900**		
Cl	**.881**		-.260
Ca	**.809**	.230	-.197
TH	**.782**	.597	
pH	-.702	**.207**	-.453
SO_4	-.527	.199	**.404**
HCO_3	**.487**	-.400	.209
Mg	.524	**.736**	.124
Na	.456	-.735	-.117
K	.450		**.708**

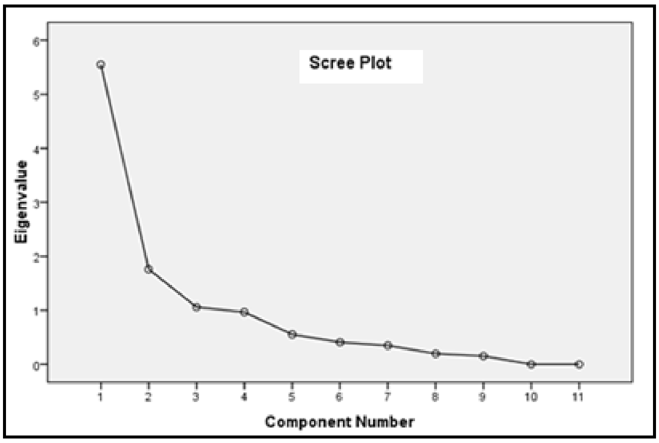

Fig 1 Scree plot of the Eigen values in pre monsoon water quality parameter

Table 5 Rotated Factor Loading Matrix, eigenvalues, % variance and cumulative variance values.

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.552	50.475	50.475	5.552	50.475	50.475	3.466	31.506	31.506
2	1.760	16.001	66.476	1.760	16.001	66.476	2.782	25.287	56.794
3	1.059	9.630	76.106	1.059	9.630	76.106	2.124	19.312	76.106
4	.966	8.778	84.884						
5	.554	5.036	89.921						
6	.409	3.714	93.635						
7	.349	3.170	96.805						
8	.197	1.795	98.600						
9	.154	1.400	100.00						
10	-1.690E-16	-1.536E-15	100.00						
11	-7.608E-16	-6.916E-15	100.00						

Parameters were grouped based on the factor loading and following factors were indicated:
Factor 1: TH, Ca, EC, Cl, TDS, Na, HCO_3
Factor 2: pH, Mg
Factor 3: K, SO_4

EC, TDS, Cl, Ca, TH, HCO_3 and Na marked factor 1, which explained 50.475% of the variance. Factor 1 had a high positive loading in EC, TDS, Cl, Ca, TH, HCO_3 and Na which were 0.954, 0.900, 0.881, 0.809, 0.782, 0.487, and 0.456 respectively. High positive loadings indicated strong linear correlation between the factor and parameters.

Thus, factor 1 can be termed as salinization factor. The electrical conductivity (EC) is positively correlated with the concentration of ions, which can thus be indirectly calculated from EC. Therefore, EC can be regarded as a water salinization index. Simultaneous drought and over pumping have led to deterioration of the groundwater. EC can be readily measured and used as a surrogate for the presence of the remaining parameters. Reducing the number of these parameters allows resources to be freed up for additional measurements elsewhere. The additional analyses could still be sampled, especially during periods when elevated EC is observed.

Factor 2, with higher loading of Mg and pH explained 16.001% of variance. pH of most of the water samples was greater than 7. Alkalinity of water may be due carbonate and bicarbonate of Na. Second factor can be called as alkalinity factor.

Factor 3, with higher loading of K, SO_4 explained 9.630% of variance and Potassium gave most contribution with a loading of 0.708. Tanning process is the main thing for those factors. This factor can be denoted as industrialization or anthropogenic factor.

Post monsoon factors

From data, 3 factors, explaining 84.064% of the total variance, was estimated on the basis of Kaiser Criterion of the eigenvalues greater or equal 1 and from a Cattle scree plot [1]. A scree plot shows the eigenvalues sorted from large to small as a function of the factor number. After the fourth factor (Fig. 2), starting the elbow in the downward curve, other components can be omitted. Factor was extracted by principal component method and rotated by Varimax. The factor loading, their eigenvalues, and variances are summarized in (Table 7).

Table 6 Components of pre monsoon water quality data

Parameters	Component			
	1	2	3	4
EC	**.881**	-.204	-.205	-.323
TDS	**.850**	-.204	-.205	-.323
Cl	**.839**	.399	-.275	-.148
Na	**.838**	.400	-.276	-.148
TH	**.781**		.534	.282
Mg	**.716**	-.313	.463	.259
HCO_3	**.558**	-.281	-.236	.508
SO_4	-.479		-.200	**.521**
pH	-.325	**.694**	-.328	.346
Ca	.567	**.649**	.273	.252
K	-.239	.400	**.627**	-.497

Table 7 Rotated Factor Loading Matrix, eigenvalues, % variance and cumulative variance values.

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.106	46.423	46.423	5.106	46.423	46.423	3.457	31.429	31.429
2	1.644	14.944	61.367	1.644	14.944	61.367	2.600	23.638	55.066
	1.407	12.793	74.160	1.407	12.793	74.160	1.616	14.695	69.762
4	1.089	9.903	84.063	1.089	9.903	84.063	1.573	14.301	84.063
5	.944	8.584	92.647						
6	.439	3.987	96.634						
7	.204	1.855	98.490						
8	.151	1.375	99.865						
9	.015	.135	100.00						
10	8.886E-6	8.078E-5	100.00						
11	2.624E-16	2.385E-15	100.00						

Parameters were grouped based on the factor loading and following factors were indicated:

Factor 1: EC, TDS, Cl, Na, HCO_3, Mg, TH
Factor 2: pH, Ca
Factor 3: K
Factor 4: SO_4

EC, TDS, Cl, Mg, Na, TH, HCO_3 and Na marked factor 1, which explained 46.4% of the variance. Factor 1 had a high positive loading in EC, TDS, Cl, Na, Mg, TH, and HCO_3 which were 0.881, 0.850, 0.839, 0.838, 0.781, 0.716, and 0.558 respectively. High positive loadings indicated strong linear correlation between the factor and parameters.

Thus, factor 1 can be termed as salinization factor. The electrical conductivity (EC) is positively correlated with the concentration of ions, which can thus be indirectly calculated from EC. Therefore, EC can be regarded as a water salinization index. Simultaneous drought and over pumping have led to deterioration of the groundwater. EC can be readily measured and used as a surrogate for the presence of the remaining parameters. Reducing the number of these parameters allows resources to be freed up for additional measurements elsewhere. The additional analyses could still be sampled, especially during periods when elevated EC is observed.

Factor 2, with higher loading of pH and Ca explained 14.9% of variance. pH of most of the water samples was greater than 7. Alkalinity of water may be due carbonate and bicarbonate of Na. Second factor can be called as alkalinity factor.

Factor 3 and 4 with higher loading of K and So_4 respectively explained 12.79% and 9.9% of variance and Potassium gave most contribution with a loading of 0.627. Tanning process is the main thing for those factors. This factor can be denoted as industrialization or anthropogenic factor.

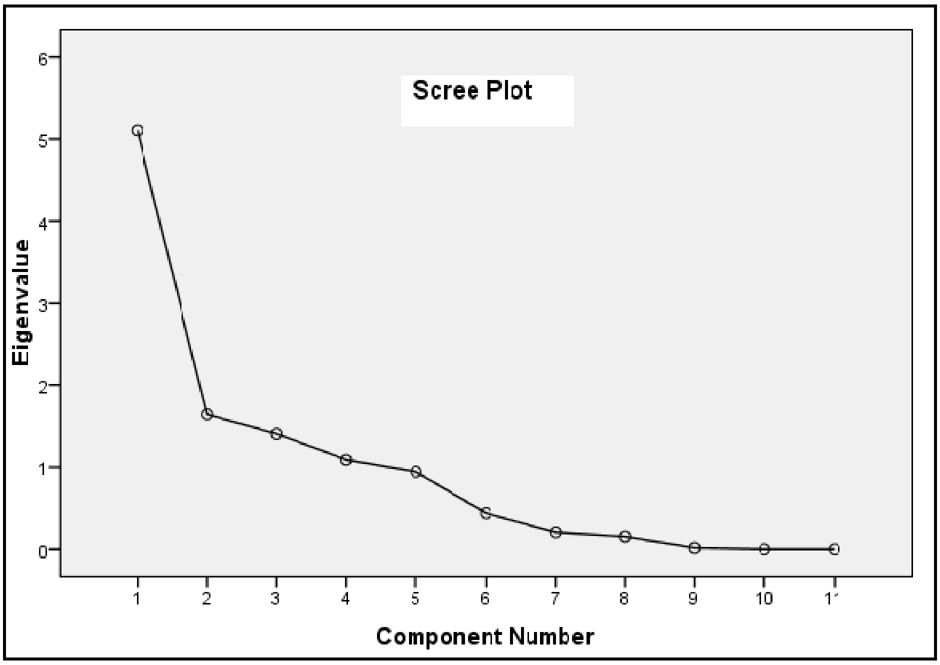

Fig 3 Scree plot of the Eigen values in post monsoon season

Cluster analysis

Pre monsoon season

In pre monsoon season, station 4, 7, 8, 5, 15, 6, 12, 16, 3 and 9 of cluster I are highly polluted areas. Factor 1 is major contributor in ground water pollution. Due to these reason stations of cluster I is more polluted than the stations of cluster II. Cluster II is corresponding to station 10, 11, 2, 1, 14 and 13 which are less polluted areas of Chrompet. These stations have higher loading in factor 2.

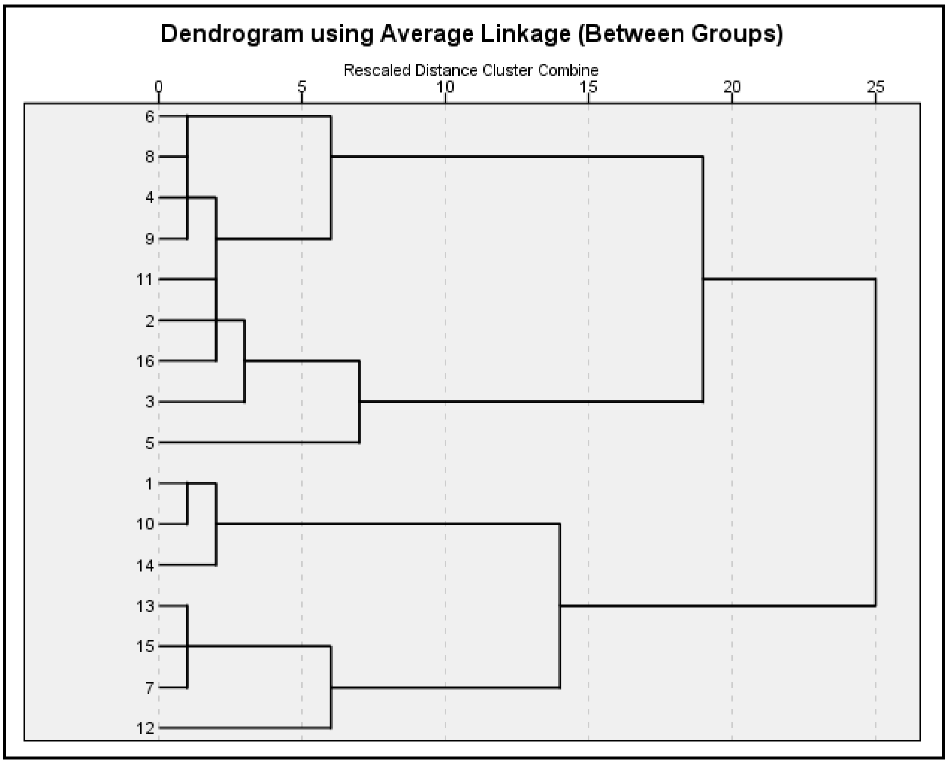

Fig 4.9 Dendorgram of the HCA of the groundwater quality of Pre monsoon

Post monsoon season:

In pre monsoon season, station 6, 8, 4, 9, 11, 2, 16, 3 and 5 of cluster I are highly polluted areas. Factor 1 is major contributor in ground water pollution. Due to these reason stations of cluster I is more polluted than the stations of cluster II. Cluster II is corresponding to station 1, 10, 14, 13, 15, 7 and 12 which are less polluted areas of Chrompet. These stations have higher loading in factor 2.

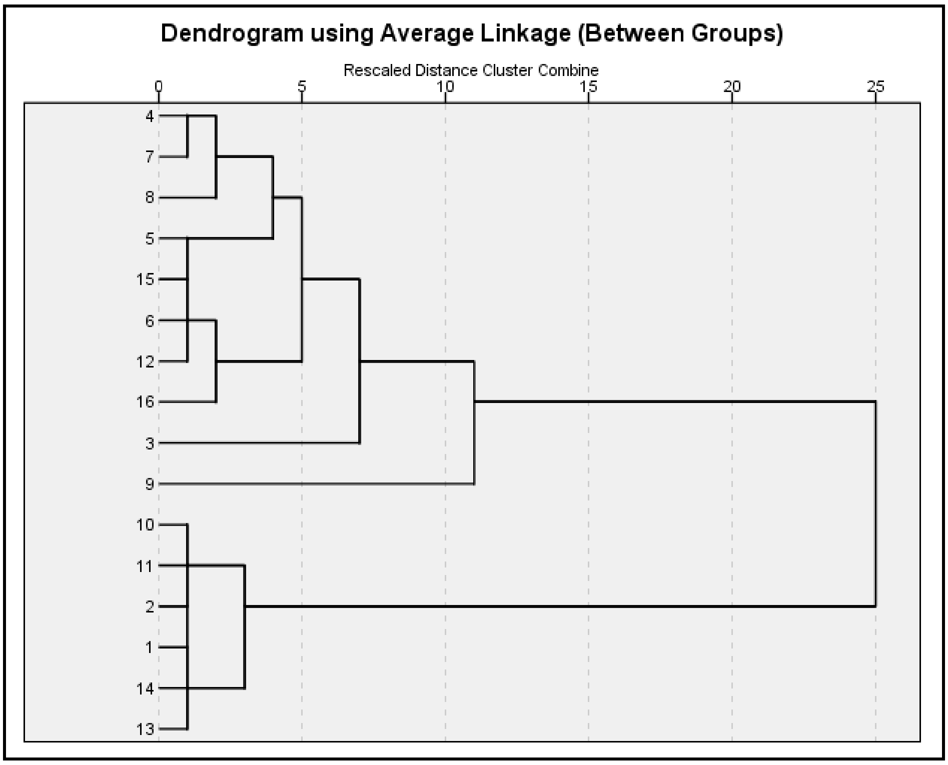

Fig 4.10 Dendorgram of the HCA of the groundwater quality of Post monsoon

Conclusion

The study showed that the analysis of hydro chemical data using the multivariate statistical techniques such as factor analysis and cluster analysis can give some information not available at first glance. Factor analysis is an effective means of manipulating, interpreting, and representing data concerning groundwater pollutants. Factor analysis converted the 11 parameters in to three factors in pre monsoon and four factors in post monsoon, which explained the data set with minimum loss of information. In pre monsoon the first factor termed as salinization factor, explained 50.475% of the total variance. The second factor can be called as alkalinity factor, which explained 16.001% of the total variance. Third factor is industrialization or anthropogenic factor, which explained 9.630% of the total variance. In post monsoon the first factor termed as salinization factor, explained 46.4% of the total variance. The second factor can be called as alkalinity factor, which explained 14.9% of the total variance. Third and fourth factor is industrialization or anthropogenic factor, which explained 12.79% and 9.630% respectively of the total variance. Hierarchical cluster analysis grouped 16 sampling stations of city into two clusters, i.e., highly polluted (HP) and moderately polluted (MP) stations, based on the similarity of water quality characteristics. It provides a useful classification of the groundwater sources in the study area that can be applied to the optimization of future spatial monitoring network with lower cost. Hence, this study illustrates that multivariate statistical methods are an excellent exploratory tool for interpreting complex water quality data sets and for understanding spatial variations, which are useful and effective for water quality management.

References

1. Asif Mahmood, Waqas Muqbool, Muhammad WaseemMumtaz and Farooq Ahmad, Application of Multivariate Statistical Techniques for the Characterization of Groundwater Quality of Lahore, Gujranwala and Sialkot (Pakistan), *Pakistan Journal Anal. Environment Chemistry* Vol. 12, 2011, 102-112.
2. S.J. Cobbina, F.A. Armah and S. Obiri, Multivariate Statistical and Spatial Assessment of Groundwater Quality in the Tolon-Kumbungu District, Ghana, *Research Journal of Environmental and Earth Sciences,* 4(1), 2012, 88-98.
3. A. Elci, and R. Potat, Assessment of the statistical significance of seasonal groundwater quality change in a Karstic aquifer system near Izmir-Turkey, *Journal of Environmental Monitoring Assessment*, 2010.
4. S. B. Kanade. and V. B. Gaikwad., A Multivariate Statistical Analysis of Bore Well Chemistry Data - Nashik and Niphad Taluka of Maharashtra, India, *Universal Journal of Environmental Research and Technology*, Vol. 1, 2011, 193-202.
5. H.C. Kataria, Manisha Gupta, Mukesh Kumar, Sandhya Kushwaha, Sherwati Kashyap, Sonal Trivedi, Rani Bhadoriya And Naval Kishore Bandewar , Study of Physico-chemical Parameters of Drinking Water of Bhopal city with Reference to Health Impacts, *Current World Environment,* 6(1), 2011,95-99 .
6. Kavitha Parmar and Vineeta Parmar, Evaluation of water quality index for drinking purpose of river Subernarekha in Singhum District, *International Journal of Environmental Science* 1, 2010.
7. P. Lilly Florence and A. Paul Raj, Ground water quality assessment of Gangavalli Taluk, Salem District, Tamil Nadu, India using multivariate statistical techniques, *Engineering Science and Technology*: *An International Journal* , ISSN: 2250-3498, 2013.
8. P. Lilly Florence, A. Paulraj and P. Ramachandramoorthy, Water Quality Index and Correlation Study for the Assessment of Water Quality and its Parameters of Yercaud Taluk, Salem District, Tamil Nadu, India, *Chemical Science Transactions*, 2012, 139-149.
9. T. Narmatha, A. Jeyaseelan, S.P. Mohan and V. Ram Mohan, Integrating Multivariate Statistical Analysis with GIS for Groundwater in Pambar Sub-Basin, Tamil Nadu, India, *International Journal of Geomatics And Geosciences* 2011.
10. Palanisamy and Kavitha, An assessment of the quality of groundwater in a textile dyeing industrial area in Erode city, Tamil Nadu, India, *Journal of Environmental Chemistry*, 2010, 1033-1099.
11. K. Ramesh and K. Seetha, Hydrochemical Analysis of Surface water and Groundwater in Tannery belt in and around Ranipet, Vellore district, Tamil Nadu, India, *International Journal of Research in Chemistry and Environment*, 2013, 36-47.

12. P. Ravikumar, Someshkar and Angamin, (2010), Hydrochemistry and evaluation of groundwater suitability for irrigation and drinking purposes in the Markandeya River basin, Belgaum district, Karnataka, *Journal of Environmental Monitoring Assessment*, 2010.
13. C.K. Singh and S. Mukherjee, Geochemical Assessment of Groundwater Quality Integrating Multivariate Statistical Analysis with GIS in Shiwaliks of Punjab, India, *School Of Environmental* Sciences 2007.
14. V. Sivakumar, M. Asaithambi, N. Jayakumar and P. Sivakumar, Assessment of the Contamination from the Tanneries & Dyeing Industries on to Kalingarayan Canal of Tamil Nadu, *International Journal of Chemical Technology* Research, 2010.
15. G. Tamma Rao, V.V.S. Gurunadha Rao and K. Ranganathan, Hydro geochemistry and groundwater quality assessment of Ranipet industrial area, Tamil Nadu, India, *Research Journal of Environmental and Earth Sciences*, 2013, 855–867.

*****