International Journal of Geology, Earth & Environmental Sciences ISSN: 2277-2081 (Online)
An Open Access, Online International Journal Available at http://www.cibtech.org/jgee.htm
2015 Vol. 5 (2) May-August, pp. 64-73/Khan and Abbasi
Research Article

SPATIAL DISTRIBUTION OF HYDROCHEMICAL DATA USING HIERARCHICAL CLUSTER ANALYSIS – A CASE STUDY

***Taqveem Ali Khan[1] and Mohammad Adil Abbasi[2]**
[1]Department of Geology, A.M.U.Aligarh-202002, India
[2]Abdul Kareem Khan Inter College, Amroha, Jyotiba Phole Nagar, U.P. India
**Author for Correspondence*

ABSTRACT
Water quality is determined by chemical analysis; the data is used for various purposes such as classification, analysis, and correlation. The graphical interpretation –a basic tool in hydrochemical studies - is one of the methods used for summarizing and presenting hydrochemical data. The statistical analysis is used to compile and evaluate the data. In this paper the performance of the graphical and statistical methods used to classify hydrochemical data is tested and compared. This includes Collins diagram, bar diagram, pie diagram, Stiff diagram, trilinear diagram, and dendrogram - hierarchical cluster analysis (HCA). All the methods were discussed and compared as to show their viability in making a simple and straight forward presentation of the hydrochemical data. It has been observed that the graphical methods used in visual interpretation of hydrochemical data have their own limitations in comparison to the multivariate statistical technique – HCA. The HCA produces the most useful grouping informations. This information gives viable information when transformed into a GIS format. The results show that the integrated technique of HCA and GIS is useful in understanding the spatial heterogeneity of hydrchemical characters of an area.

Keywords: *Cations, Anions, Graphical Presentation, GIS, Collins Diagram, Piper Trilinear Diagram*

INTRODUCTION
The quality of groundwater is as important as its quantity. Water being a universal solvent its purity cannot remain intact. Hydrogeochemical studies require a large amount of information related to groundwater chemistry. Number of chemical variables is responsible for the evaluation of the complex nature of the groundwater of an area. There is a need for comparing the origin and distribution of groundwater masses having different geochemical attributes. The hydrochemical facies are frequently compared by means of graphical representations (Back, 1961; Zaporozec, 1972), such as Stiff diagrams (Stiff, 1951) and trilinear diagrams (Piper, 1944).
There are number of conventional graphical methods of data analysis available for the interpretation and presentation of chemical analysis results (Back, 1961; Matthess, 1982; Hem, 1989), e.g. histograms, trilinear, semi-logarithmic diagrams and many others (Lloyd and Heathcote, 1985). The advantages and disadvantages of graphical methods and statistical analysis of hydrochemical data has been studied by many worker (e.g., Dalton and Upchurch, 1978; Güler *et al.*, 2002).
Water quality is a multivariate concept. It is not defined by any single constituent. Because of the occurrence and reoccurrence of related mineral species among flow path, water chemistry variables typically are correlated among each other. Consequently a variable by variable analysis ignores relevant information (Riley *et al.*, 1990). Multivariate analysis has been applied to hydrochemical data to a limited extent in the past (e.g. Thomas, 1977; Symader and Thomas, 1978; Keet, 1982; Khan 2008). Multivariate analysis is a method of choice for analyzing hydrochemical data, particularly when the data is supported and interpreted with a thorough knowledge of the hydrogeology of an area.
The paper highlights the advantage of a statistical tool i.e., hierarchical cluster analysis relative to conventional graphical methods of data representation. The objective of this paper is to show the viability of the hierarchical cluster analysis data when transformed into a GIS format to show the spatial distribution of groundwater quality in the area. The integrated approach discussed here can be used in other areas also.

MATERIALS AND METHODS

Results of 18 water quality analysis samples are used for the present study. SPSS 13 software is used for the statistical analysis. Surfer 9 software is used for preparing the spatial distribution map.

The widely used Collins bar diagram, pie diagram, stiff diagram and Piper trilinear diagram are plotted as a representative case for graphical presentation of hydrochemical data. Pearson coefficient is done besides the case in point Hierarchical cluster analysis (HCA), The cluster method groups samples into distinct population (clusters) that may be significant in the hydrogeological, hydrochemical context, as well as from the statistical point of view (Guler *et al.*, 2002).

Cluster analysis datasets were identical to those typically used for the construction of Piper diagrams' this will enable a direct comparison of results from the two methods.

The dendrogram is plotted showing water clusters in the area. These clusters are plotted on the map so as to show the spatial distribution of groundwater quality in the area.

Study Area

The study area forms the south eastern part of the Bulandshaher district and bounded in the east by the Ganga River and in the west by the Nim River (Figure 1). It is spread over an area of 267.78 sq. kms. It lies between 28^{o} $5^{'}$ and 28^{o} $22^{'}$ N latitudes and between 78^{o} $15^{'}$ to 78^{o} $27^{'}$ $30^{'}$ E longitudes. The Ganga - Nim sub - basin falls under the subtropical climatic zones of India. During summer temperature rises to 45^{o} C and falls to 4^{o} C during winters. The average rainfall in the area is 618.97 mm.

Three types of soils have been found in the area: sandy loam, sandy to loam, and loam to clay loam. Near the track of river Ganga the soil type is sandy loam. It is saline in nature, consequent to shallow groundwater table. The salt efflorescence appears to be a common feature of the entire tract which is locally called "usar".

Sandy loam to loam type of the soil covers the major portion of the upland tract. The soil varies in color from light brown to deep brown and the texture of the soil is sandy to good quality loam. These soils are well drained.

Clay to clay loam type of soil is only in a small of the portion on left side of the lower Ganga canal.

The studies of well drilled by Oil and Natural Gas Commission and Central Ground Water Board in search of oil and gas, and water in the Ganga basin decipher the sub – surface geology as consisting of alternate spurs and depressions. Bundelkhand granite formed the basement in the Ganga basin. This granitic massif underwent structural dislocation sometimes during the pre – Vindhyan time generating thereby the two prominent basins viz, east U.P shelf and west U.P shelf, where upper Vindhayans were deposited. Since upper Proterozoic to the lower Tertiary, it underwent erosion and during the Pliocene period, Neogene Siwaliks were deposited, which was later on followed by the deposition of Quaternary sediments.

Hydrogeologically, there occurs three tiers aquifer system down to the depth of 120 m.b.g.l. The aquifer material consists of fine through medium to coarse sand. The groundwater occurs under phreatic condition in shallow aquifers and semi confined to confined conditions in the deeper aquifers. The water table in the area ranges between 0.76 to 15.40 m.b.g.l. The elevation of water table ranges from189 to 169 meters. The groundwater flow in the area is form northwest to southeast in direction with some local variations. On the right bank of the river Ganga the general flow is towards the east i.e., towards the river Ganga, where as in the western part, it is from north –east to south-east.

The hydraulic gradient ranges from 0.3 m/km to 4 m/km. The gentle hydraulic gradient was observed all over the area except near the Ganga bank where it is steep (4m/km) which may possibly due to the predominance of low permeability zones.

Synopsis of Hydrochemical Analysis

The groundwater of the sub – basin is mildly alkaline in nature where pH ranges between 6.71 – 7.81. The electrical conductivity of the water sample ranges between 173 – 1419 micromhos/cm at 25^{o} C. Chloride concentration in the groundwater samples of the area varies from 16 – 625 mg/l. Save at two location water is suitable for drinking purposes. Sulphate concentration varies from 4 – 519 mg/l. Except at one location the water is with in the permissible limit of 250 mg/l (IC.MR., 1975; W.H.O., 1984) Sodium is

found to range between 4.98 to 216 mg/l. Save the location number 17 water samples were found within the limit. Potassium varies between 2mg/l to 172mg/l. The calcium ranges between 39 to 151 mg/l. Magnesium was found to vary from 28 to 207mg/l. Total hardness varies from 104 to 760mg/l. Total dissolve solids range between 233 to 281mg/l. In all the groundwater of shallow aquifer of the area that is used for drinking and other household purposes is potable, hard, and alkaline in reaction and slightly - mineralized (Khan and Abbasi, 2003).

RESULTS AND DISCUSSION

Results

Graphical Methods

The bar diagram most widely used is that of Collins (1923). The bar has two vertical columns whose heights are proportional to the total concentration of cations in the left column and of anions in the right column, both in meq/l (Figure 2). The Collins diagram is an effective tool in the oral presentations of water composition (Zaporozec, 1972). Pie diagram presents the relative major ion composition in percent milliequivalents per liter. The Stiff pattern is a polygon that is created from three parallel horizontal axes extending on either side of a vertical zero axes (Stiff, 1951). In this diagram, cations are plotted on the left of the axes and anions are plotted on the right, in units of milli-equivalents per liter (meq/l). The Stiff diagram is usually plotted without the labeled axis and is useful in making visual comparison of waters with different characteristics. The patterns tend to maintain its shape upon concentration or dilution, thus visually allowing us to trace the flow paths on maps (Stiff, 1951).

Piper trilinear diagram is widely used for visualization and classification of hydrochemical data. Piper trilinear diagram has three parts: a cation triangle, an anion triangle, and a central diamond shaped field. Several attempts were made to plot the relative percentages of cations and anions on triangular and rectangular diagrams (e.g., Langelier and Ludwig, 1942; Romani, 1981; Chadha, 1999; Ray and Mukherjee, 2008). Trilinear diagram (Figure 3) allows determining the groundwater facies in the study area. In the anion triangle the range of facies is from chloride to sulphate - bicarbonates. The cation distribution shows a range of facies from sodic to magnesium – calcic. Finally, projection of the data in the central diamond shaped field indicates that there is progression from sodium chloride facies to calcium – magnesium bicarbonate facies.

Statistical Methods

Pearson Correlation Coefficient

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 Hirsh, 2002). The data were statistically computed using correlation coefficient in order to indicate the sufficiency of one variable to predict the other (Davis, 1986).

The correlation matrix is useful in delineating the association between variables showing overall coherence of the data set and indicating the participation of the individual chemical parameters in several factors, a fact commonly occurred in hydrochemistry. The results of Pearson correlation coefficient analysis are given in the Table 1. The variables having coefficient value (r) > 0.5 are considered significant. Perusal of the table indicates that sodium is positively correlated with bicarbonate, chloride and sulphate. Potassium is positively correlated with chloride, sulphate and sodium. Calcium has a positive relation with chloride, sulphate, sodium and potassium. EC is positively correlated with bicarbonate, and chloride is positively correlated sulphate. The negative relation is found to exist between EC and pH, EC and carbonate, EC and potassium. Bicarbonate is negatively correlated with carbonate, where as magnesium is negatively correlated with bicarbonate and sulphate. The variation in relationship indicates the complexity of the quality of groundwater and effectiveness of Pearson’s coefficient in deciphering the hydrochemical data.

Research Article

Hierarchical Cluster Analysis

An exploratory analysis is carried out to find the similarities in the dataset. This exploratory analysis phase consist of grouping the samples into clusters. The analysis defines groups of samples with similar hydrochemical characteristics. The clustering analysis consists of five steps.

(1) Select appropriate level of measurement.
(2) Transform and standardize the variables.
(3) Select the appropriate distance or similarity measure.
(4) Select the clustering algorithm.
(5) Carry - out the analysis and interpret the data.

In the present study hierarchical cluster analysis is used to group the data into clusters and a dendrogram is prepared (Figure 4) using SPSS 13 Software. Hierarchical clustering joins the most similar observation and then successfully the next most similar observation. The levels of similarity at which observations are merged are used to construct a dendrogram (Chen *et al.*, 2007). The Euclidean distance is represented on the horizontal axis of the dendrogram. It gives similarity between two clusters.

Dendrogram prepared forms five groups of the hydrochemical data. Samples belonging to each group (Table 2) and their chemical characteristics are elaborately discussed here.

Group A (EC, Cl, Na K type) consist of samples from location no 3,14,7,13,18,8,2,9,12,16, and 6 (Figure 1). The pH ranges from 6.71 to 7.77. Lowest pH value water is reported in this group. Electrical conductivity values ranges between 426 to 680 micromhos/cm. The values of EC are higher than the values of EC in the other groups. So, the water of this group can be categorized as EC dominated. The chloride, sodium and potassium are in close range which manifest that the water type is chloride, sodium and potassium type.

Presence of sodium and chloride manifest high salt present in the vadose zone. Recharge water while moving through vadose zone dissolve salts that finally add to the groundwater. Weathering of silicate minerals contribute to the presence of high sodium and chloride.

Group B (Na, K, Ca, Mg type) consist of the samples no 4, 10, 17. pH ranges from 7.52 to 7.81. More than 7 pH value manifest the water type is a bit alkaline in nature. Here EC ranges between 173 and 257. So the water of low EC is categorized by this group. Carbonate and silicate weathering supply Ca and Mg in the groundwater. The water type is dominated by sodium, potassium, calcium and magnesium which are in close range in this group.

Group C (CO_3. HCO_3, Na, Ca type) comprises of samples from location no 1 and 5. pH values are 7.22 and 7.33. Again the water type is alkaline in nature. EC values vary from 799 to 945. The water type of this group is carbonate, bicarbonate, sodium and calcium dominated.

The presence of bicarbonate in the water is from the dissolution of carbon dioxide from the soil by biochemical activity.

Group D (CO_3, Cl, Ca, Mg type) and E have one sample each. Group D has relatively high chloride, carbonate, calcium and magnesium concentration. The water type is low in sulphate.

Group E (CO_3. HCO_3, Cl, Na, Ca Mg) has highest values of all variables. EC is 1419 micromhos/cm highest in the area. However pH is 6.98 which happen to be at the lower side suggesting acidic type of water. Carbonate, bicarbonate, chloride, sodium, calcium and magnesium are relatively higher than that of other samples.

Discussion

Graphical interpretation of groundwater quality is limited because (1) there are a finite number of variables that can be considered, (2) the variables are generally limited by convention to major ions, and (3) spurious relationships may be introduced by use of certain of the procedures. The spurious results are an artifact of the use of closed number systems to compute the trilinear diagram (Dalton and Upchurch, 1979).

Chayes (1960, 1971) has shown that by converting numbers to proportions with a constant sum, i.e., 100%, the numbers may no longer be independent. For example, in calculating the anion proportions for the trilinear diagram, chloride, sulfate, and bicarbonate plus carbonate must sum to 100 percent.

Table 1: Pearson correlation matrix

	pH	EC	CO_3^{2-}	HCO_3^-	Cl^-	SO_4^{2-}	Na^+	K^+	Ca^{2+}	Mg^{2+}
Ph	1									
EC	-.483	1								
CO_3^{2-}	.493	-.114	1							
HCO_3^-	-.467	**.717**	-.368	1						
Cl^-	.253	.223	.018	.404	1					
SO_4^{2-}	.278	.100	.231	.265	**.803**	1				
Na^+	.213	.429	.116	**.545**	**.883**	**.715**	1			
K^+	.438	-.197	.214	.201	**.790**	**.726**	**.648**	1		
Ca^{2+}	.131	.098	.108	.204	**.832**	**.729**	**.659**	**.633**	1	
Mg^{2+}	.199	.071	.278	-.068	.169	-.132	.172	.115	.270	1

Table 2: Cluster groups and their members

Group	Member (Sample No)
A	3,14,7,13,18,8,2,9,12,16,6
B	4,10,17
C	1,5
D	15
E	11

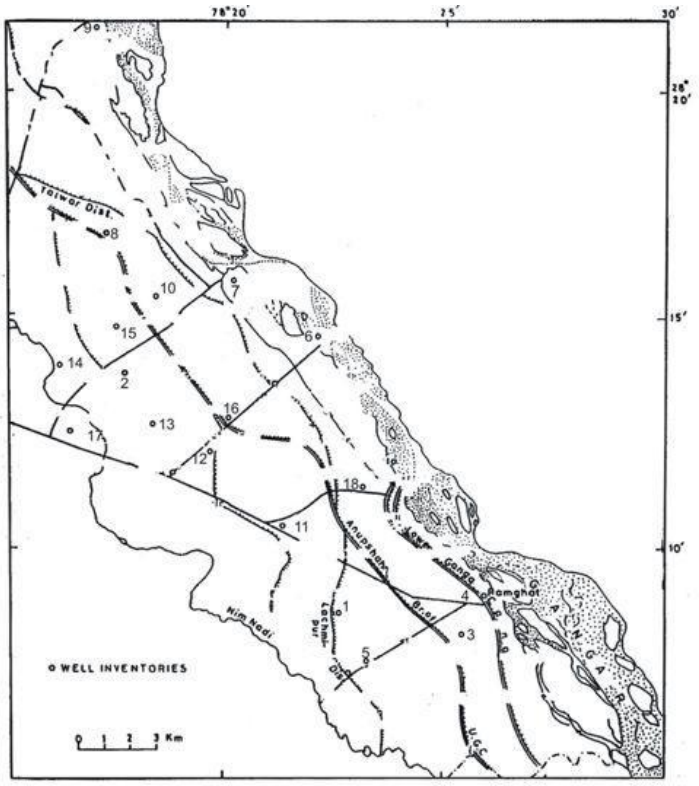

Figure 1: Base Map

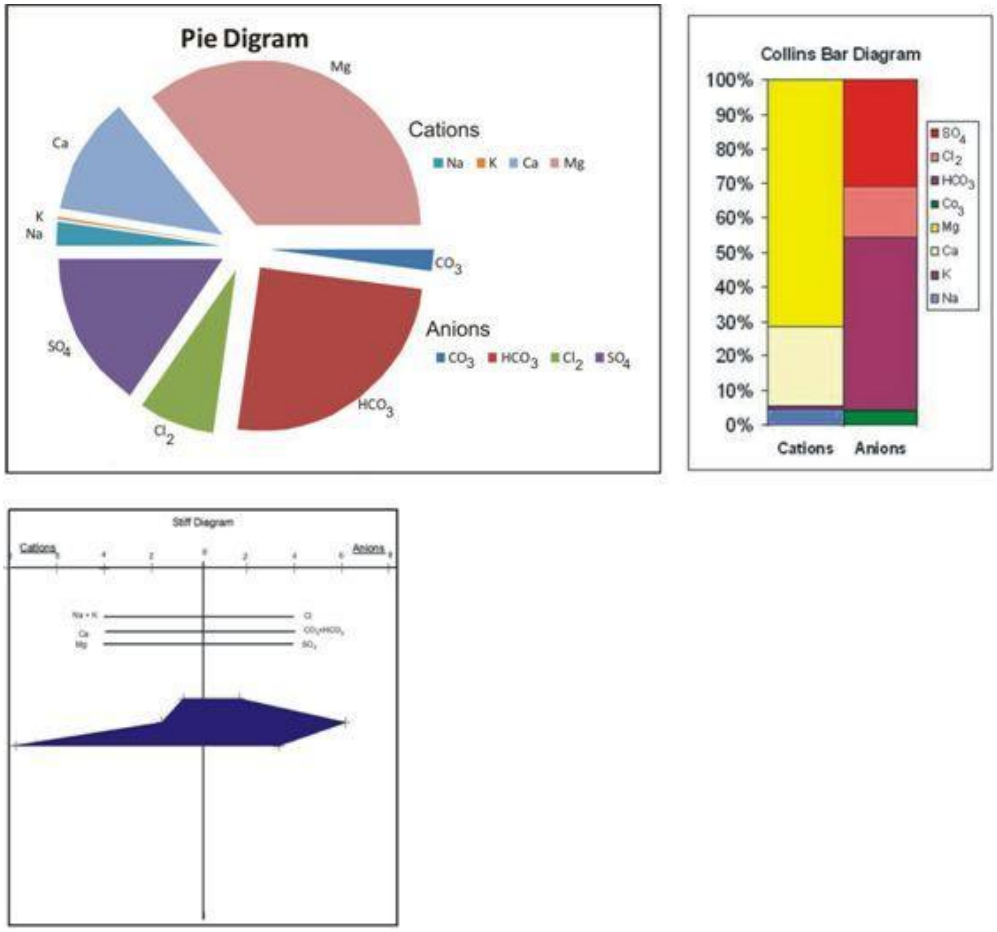

Figure 2: Stiff Collin bar diagram

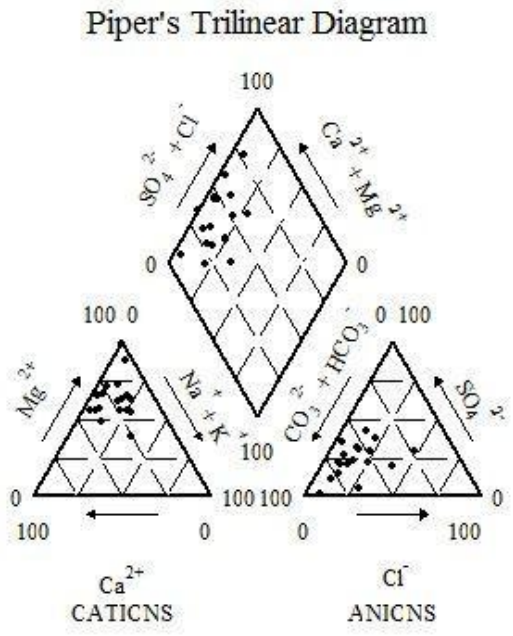

Figure 3: Piper trilinear diagram

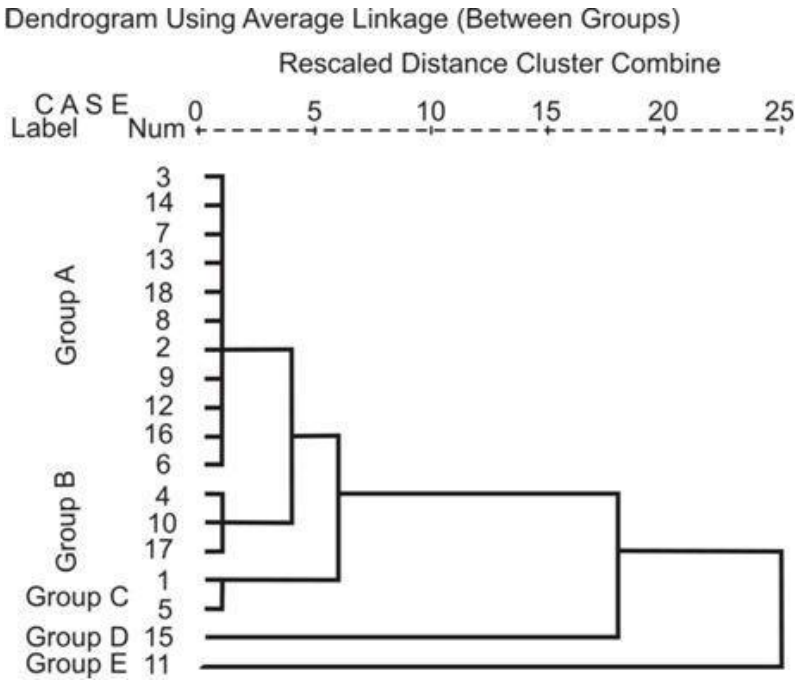

Figure 4: Dendrogram

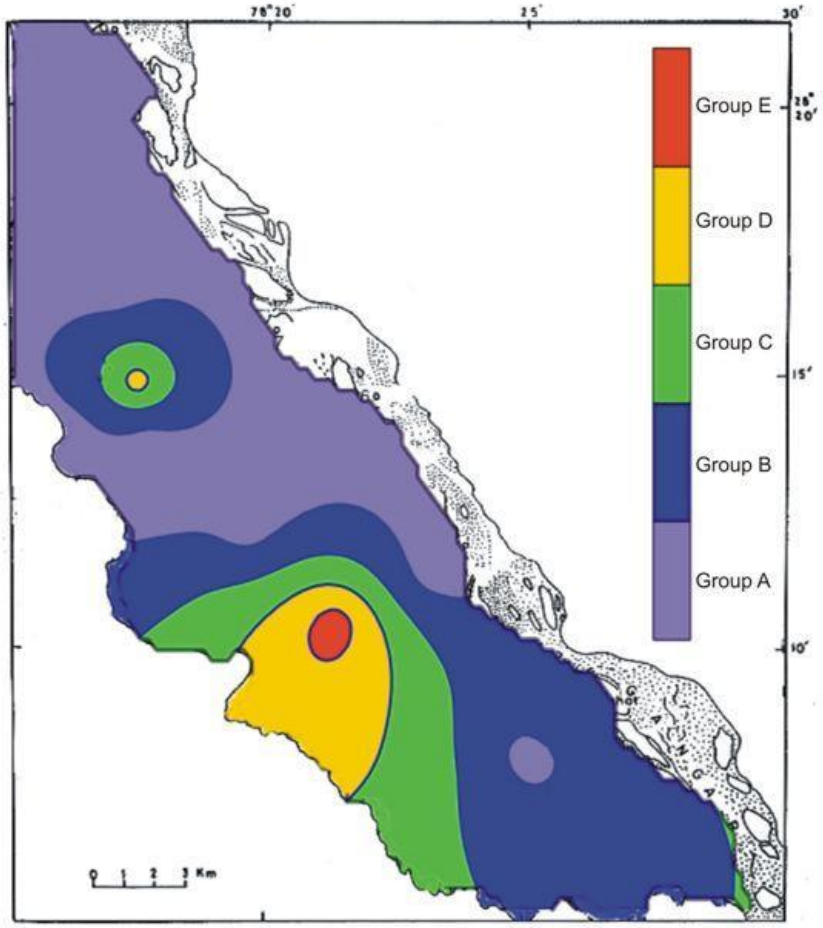

Figure 5: Spatial distribution

Therefore, any increase in concentration in one or two of the variables forces an apparent, but perhaps false, decrease in third variable. Clearly, this response complicates interpretation of hydrochemical facies from trilinear diagrams (Dalton and Upchurch, 1979). Trilinear diagram limits the visualization of geochemical facies of the groundwater as a whole. The resultant facies can not be plotted to show the spatial variation of the water type in the area. Trilinear diagram reduces the water type of the whole area into a single unit.

The efficiency and semi-objective nature of the statistical techniques makes these techniques superior to the graphical methods in order to group samples based on water chemistry data (Guler *et al.,* 2002). For the integrated approach the reduced or the clustered data is to be brought into the GIS format. This integrated approach will help in depicting the spatial distribution of hydrochemical analysis data in one view. The cluster - formed groups (Table 2) are plotted on the base map of the area. Each group is assigned a different value ranging from 1 to 5 corresponding to the group it belongs. The map is digitized using Golden Surfer 9 software. A GIS map is thus prepared showing clustering groups distribution over the entire area (Figure 5). The map is unique in the sense that the results of the hierarchical cluster analysis are brought in the GIS format. The map thus prepared depicts the spatial distribution of clustered water type in the area. Perusal of the map shows that the area parallel to the river Ganga is of Group A and B type only. No other group water – type touches the river Ganga. Half of the map encompassing northern portion of the area comprises group A type of water. The central part of the northern region has in all three types of water i.e. group B, C and D. But this only covers relatively small field as shown in the figure. About 50% of the southern half of the study area consists of group B type of water which is high in sodium, potassium, calcium and magnesium. This can be said a cation region because of the presence of cations - in close concentration. This is the region through which lower Ganga Canal traverses and the water table is shallow. In the middle of this group B a small portion has group A type of water. Group C, D and E are in quick succession almost parallel to the western flank of lower Ganga canal. Thus, high salt content in this region can be attributed to shallow water table and deposition of more salts in the soil due to evaporation.

Conclusion

The Figure 2 shows that Collins, Stiff, and pie diagrams are of one sample only. So, it is not practically feasible to produce and sort multiple number of figures, one for each example, in order to sort and classify large data sets. The choice of similarity would be based on the evaluation of the analyst, which is highly subjective (Guler *et al.,* 2002). The graphical methods used to group the samples is not efficient and produce biased results. But these method are effective in visual representing of the dataset.

Hydrochemical data from the study area has been analyzed using Pearson's correlation coefficient and hierarchical cluster analysis. These methods are used to see if there exist any relation between different variables and in how many groups the water of the area can be clustered. The analysis shows that some variables are positively correlated with other. That is the rise of one variable will envisage the rise of the other variable. While on the other hand negative relation exists between some variables that is the fall in one will result in the rise of the other. The hierarchical cluster analysis has in all clustered the water of the area in five groups. Each group shows a distinct dominance of select variables. The spatial variability observed in the composition of the major ions provide insight into aquifer heterogeneity and connectivity, as well as the physical and chemical processes controlling water chemistry. The techniques discussed in this paper has advantages and disadvantages in clustering and displaying water samples using chemical and physical parameters. The graphical methods like Collin, pie Stiff and trilinear diagrams provide valuable and rapidly accessible information regarding the chemical composition of water samples e.g., the relative proportion of the cations and anions. But these techniques have some limitations when used alone. These techniques are not useful in producing distinct grouping of samples as there is no inherent means to discriminate the groups or to test the degree of similarity between samples in a group. On the other hand the HCA technique is more efficient as it offers a semi objective – graphical clustering procedure called dendrogram. An integrated approaches seems to offer a method that holds the advantages while minimizing the limitations of either approach

The study, further, demonstrates that the moderate or a large set of hydrochemical data can be reduced in groups on the basis of similarities among the variables by using hierarchical cluster analysis. And when these groups are plotted and overlaid on base map a discernible spatial distribution of the hydrochemical data is ascertained potently.

This method has an edge over methods showing distribution of individual variables because these maps evades the rest of the variables. And if a number of variable distribution maps are prepared and overlaid in a GIS environment to form a single map then it becomes too knotty to be read. Thus, the method elaborated here provides the researcher with quantitative information about the groundwater quality data base that they would not be able to infer by judgment and experience alone.

REFERENCES

Back W (1961). Techniques for mapping of hydrochemical facies. US Geol Surv Prof Paper **424-D** 380–382.

Chadha DK (1999). A proposed new diagram for geochemical classification of natural waters and interpretation of chemical data. *Hydrogeology Journal* **7**(5) 431–439.

Chayes F (1960). On correlation between variables of constant sum. *Journal of Geophysics Research* **65** 4185-4193.

Chayes F (1971). *Ratio Correlation* (Univ. of Chicago Press) Chicago 99.

Chen K, Jiao JJ, 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.

Collins WD (1923). Graphic representation of analyses. *Industrial & Engineering Chemistry* **15** 394.

Dalton GM and Upchurch SB (1978). Interpretation of Hydrochemical facies by factor analysis. *Ground Water* **16**(4) 228-233.

Davis JC (1986). *Statistics and Data Analysis in Geology*, 2nd edition (John Willey and Sons) New York.

Guler CG, Thyne JE, McCray and Turner AK (2002). Evaluation of graphical and multivariate statistical methods for classification of water chemistry data. *Hydrogeology Journal* **10** 455 – 474.

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

Hem JD (1989). *Study and Interpretation of the Chemical Characteristic of Natural Water*, 3rd edition. US Geological Survey of Water Supply Paper 2254.

ICMR (1975). *Manual of Standards of Quality for Drinking Water Supplies*, 2nd edition. I.C.M.R., New Delhi.

Keet BA (1982). Regional ground water survey in county kildare, Ireland Technical Report, Institute of Earth Science Amsterdam, The Netherland.

Khan TA (2008). Cluster analysis and quality assessment of groundwater in and around Aligarh city, U., India. *Proceeding All India Seminar on Advances in Environmental Sciences and Technology* 109- 114.

Taqveem Ali Khan and Abbasi MA (2003). Hydrochemical studies of shallow groundwater in the Dibai Block of Bulandshahar Distt, U.P., *Pollution Research* **22**(4) 503-506.

Langelier WF and Ludwig HF (1942). Graphical methods for indicating the mineral character of natural waters. *American Water Works Association Journal* **34** 335–352.

Lloyd JW and Heathcote JA (1985). *Natural Inorganic Hydrochemistry in Relation to Groundwater, an Introduction* (Clarendon Press) Oxford.

Matthess G (1982). *The Properties of Groundwater* (Wiley) New York.

Piper AM (1944). A graphical Procedure in the geochemical interpretation of water analyses. *Transactions of the American Geophysical Union* **25** 914-923.

Ray RK and Mukherjee R (2008). Reproducing the Piper trilinear diagram in rectangular roordinates. *Groundwater* **46**(6) 893–896.

Riley JA, Steinhorst RK, Winter GV and Williams RE (1990). Statistical analysis of the hydrochemistry of groundwater in Columbia River basalts. *Journal of Hydrology* **119** 245 – 262.

Romani S (1981). A new diagram for classification of natural waters and interpretation of chemical analysis data. *In Proceedings of Quality of Ground Water International Symposium,Noordwijkerhout. Studies in Environmental Science 17, Amsterdam* (The Netherlands: Elsevier).
Stiff HA Jr (1951). The interpretation of chemical water analysis by means of patterns. *Journal of Petroleum Technology* **3**(10) 15-16.
Thomas W (1977). Schwermetalle in flubsedimenten klassifizierung und bewertung mit methoden der multivariaten statistic. Verhandlungen der Gesellscharft fur okologie, Keil.
Symader W and Thomas W (1978). *Interpretation of Average Heavy Metal Pollution in Flowing Waters and Sediment by Means of Hierarchical Grouping Analysis Using Two Different Error Indices*, Catena **5** 131 -144.
WHO (1984). Guidelines for drinking water quality, W.H.O, Geneva.
Zaporozec A (1972). Graphical interpretation of water quality data. *Groundwater* **10**(2) 32-43.