DATA MINING OF SANGANER TEHSIL, JAIPUR (RAJASTHAN) WITH CLUSTERING TECHNIQUES OF GROUNDWATER CONTAMINATION

Mehta Anurika[1], Jain Nupur[2], Duggal Rakesh[3]

[1]*Ph.D. Scholar*
Department of Chemistry, Poornima University, Jaipur, India,
mehta.anu25@gmail.com

[2]*Associate Professor*
Department of Chemistry, Gurukul Institute of Engg. & Tech., Kota, India,
jain_nupur2010@yahoo.in

[3]*Director*
Poornima Group of Institutions, Jaipur, India,
rakesh@poornima.org

Abstract

Water samples were collected from forty different villages of Sanganer tehsil for groundwater analysis. The samples so collected from different sources like wells, tube-wells and hand-pumps in three phases of year 2014 during Pre - monsoon, Monsoon and Post - monsoon were analyzed for physico-chemical parameters like pH, EC, TDS, TH, Cl^-, SO_4^{2-}, F^- and NO_3^- etc. The statistical Hierarchical Cluster Analysis (HCA) method classified 40 groundwater samples into three clusters of mixed water (Group 1), blended water (Group 2) and highly blended water (Group 3) and is based on similarity of groundwater quality characteristics demonstrating the usefulness of multivariate statistical analysis. The results obtained appear to be of importance as they will be helpful for monitoring and managing ground water pollution in terms of water quality. Spatial and temporal data sets of hydrological processes are often very large and difficult to analyze and display but present analysis attempts to throw light on groundwater quality, sources of groundwater contamination, groundwater quality variation and its spatial distribution. Clustering has been done using different algorithms such as hierarchical based algorithms for providing an evolutionary algorithm for clustering starting from data mining mechanism and highlighting important facets in the context of clustering algorithms, namely, hierarchical based algorithm.

Index Terms: *Data Mining, Groundwater, Contamination, Hierarchical Cluster Analysis, Water quality.*

1. INTRODUCTION

Groundwater is valuable natural resource for drinking water supply and irrigation in India. It gets contaminated either naturally or by fertilizers used in agriculture, constructional, domestic and industrial wastes. Therefore groundwater quality management is necessary for better quality of life. The analyzed physico-chemical parameters of study area are in large data set and have been studied using data mining as it is the process of extracting information from large data sets using algorithms and techniques drawn from the field of statistics, machine learning and Data Base Management Systems. [1] Data mining and its techniques such as cluster analysis have been used in study area where people are affected by contaminated groundwater. An attempt has been made to identify the areas of Sanganer tehsil which are affected with the contaminated groundwater, using clustering hierarchical based algorithm. Through application of this method (HCA), the number of sampling stations, variables or both of these factors can be optimized and this optimization leads to upgrading of monitoring networks. [2]

2. MATERIAL AND METHODS

2.1 Study area

Sanganer Tehsil is attached with main city of Jaipur with geographical coordinates of in North and in East. It lies between 26°49'N to 26°51' N latitude and 75°46'E to 75°51' E longitude. It is widely known for the industry of handmade papers, sanganeri printing as well as for the Jain temples. The total population of Sanganer tehsil is 573171 as per census 2011. There are about 142 villages in Sanganer tehsil.

2.2 Water sampling

A total forty samples were collected from wells, tube-wells or hand-pumps from different villages of Sanganer Tehsil during pre-monsoon period (June), monsoon period (August) and post-monsoon period (October) of year 2014 and details are summarized in Table-1 along with GIS map of sampling stations in Fig-1. Before sampling, the water was left to run from the source for ten minutes and then water sampling was done in the laboratory conditions. Temperature, pH, electrical conductivity, total dissolved solids, salinity were measured on site using potable meter (PCS Tester35 Multi-parameter). All other parameters were analyzed according to the standard methods of APHA [3] and compared with W.H.O. standards. [4]

Table-1: Source & location of samples of different villages of Sanganer Tehsil

Sample No.	Sampling Source	Village
S1	Hand Pump	Asawala
S2	Hand Pump	Bagru
S3	Tube Well	BagruRawan
S4	Hand Pump	Baksawala
S5	Well	Bamoriya
S6	Well	Bar kaBalaji
S7	Hand Pump	Beelwa
S8	Hand Pump	Bhankrota
S9	Tube Well	Bhatawala
S10	Hand Pump	Dayalpura
S11	Hand Pump	Durgapura
S12	Tube Well	Goner
S13	Tube Well	Govindpura
S14	Tube Well	Hajiwala
S15	Hand Pump	Heerapura
S16	Hand Pump	Jagannathpura
S17	Tube Well	Jaranwala
S18	Hand Pump	Khetapura
S19	Hand Pump	Khori
S20	Hand Pump	Kishorpura
S21	Hand Pump	Lakhawas
S22	Well	Laxmipura No. 1
S23	Hand Pump	Mahapura
S24	Tube Well	Mahel
S25	Hand Pump	Manoharpura
S26	Hand Pump	Mohanpura
S27	Hand Pump	Muhana
S28	Tube Well	Nevta
S29	Hand Pump	Pratapnagar
S30	Tube Well	Ramchandrapura
S31	Tube Well	Ramsinghpura
S32	Hand Pump	Sanganer
S33	Tube Well	Seemliya
S34	Tube Well	Shikarpura
S35	Hand Pump	Sirani
S36	Tube Well	Sitapura
S37	Tube Well	Sukhdeopura
S38	Hand Pump	Surajpura
S39	Tube Well	Teelawas
S40	Tube Well	Vatika

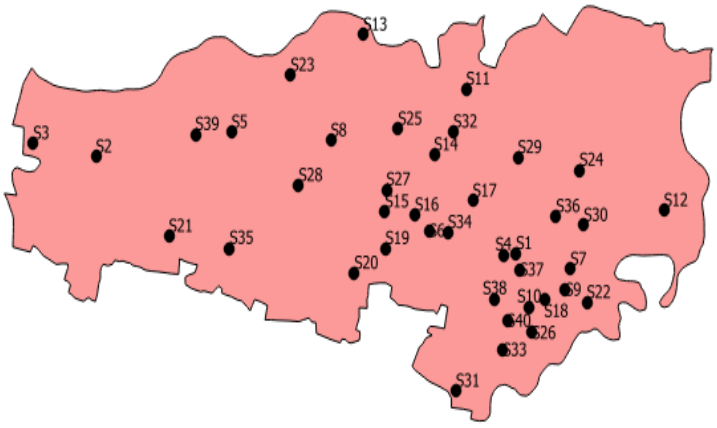


Fig-1: GIS map of Sanganer tehsil sampling stations

2.3 Data Mining Analysis

Data mining, also known as Knowledge Discovery in Database (KDD) refers to extraction of useful information from data in databases. [5] The Fig-2 (a) shows data mining steps used in the study:

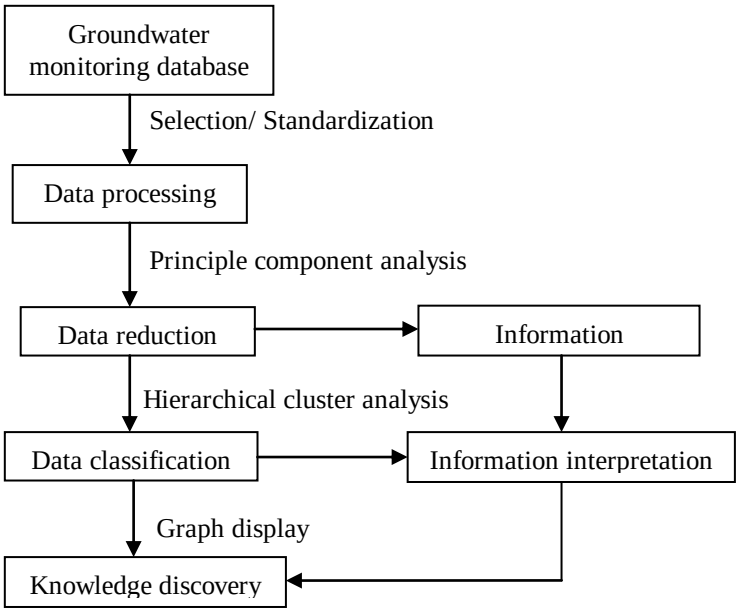


Fig-2 (a): Knowledge discovery processes

The knowledge discovery in databases process comprises of a few steps leading from raw data collections to some form of

new knowledge. [6] The process consists of the steps shown in [Fig-2 (b)]:

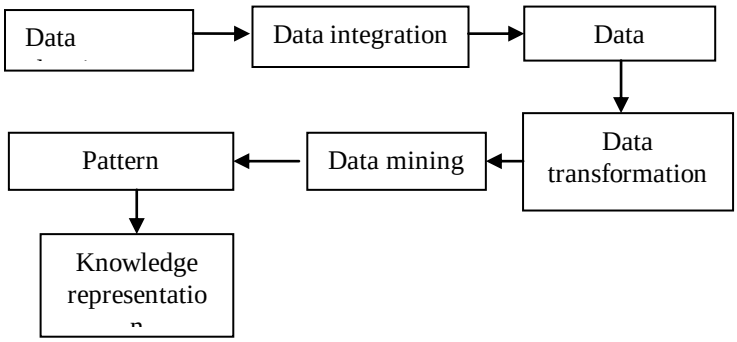


Fig-2 (b): Various steps of data mining

2.4 Cluster Analysis

In clustering, Ward method with Euclidean distant is used to measure the similarity between various data sets. Cluster analysis is the technique of data mining, which forms groups from available large data based on their similarities to each other.

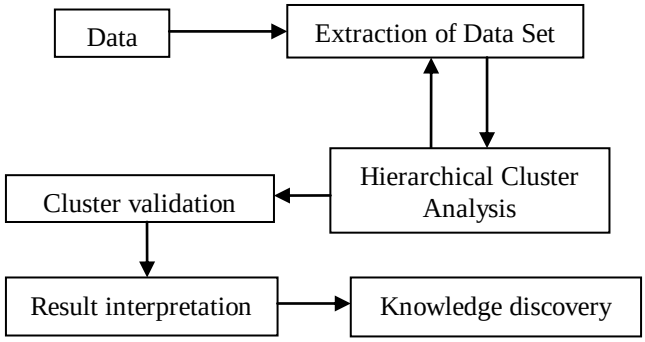


Fig-3: Clustering procedure [7]

Fig-3 shows the different steps of cluster analysis. Hierarchical cluster analyses (HCA) are effective methods of manipulating, interpreting and representing data concerning groundwater contaminants and geochemistry [8] and have been applied to data generated from study area of Sanganer tehsil as marked in GIS representation in Fig-1.

3. RESULTS AND DISCUSSION

In the present study, HCA has been employed on the standardized data using Ward method with Euclidean distances to measure the similarity of data and represented into dendrogram plot. All the physico-chemical characteristics were used as variables to show the spatial heterogeneity among the sampling stations as a result of sequence in their relationship and the degree of contamination. Dendrograms are based on sampling stations of classified 40 monitoring sites in the Sanganer tehsil and are grouped into three groups (Group-1, Group-2 and Group-3) based on similarities of water quality characteristics.

The group classifications varied significantly, because the sites in these groups had similar features and natural backgrounds that were affected by similar sources. Accordingly, forty sampling station clusters classified into three groups, Group-1, Group-2 & Group-3 in pre-monsoon, monsoon and post-monsoon periods respectively are represented in [Fig-4 (a), (b) & (c)].

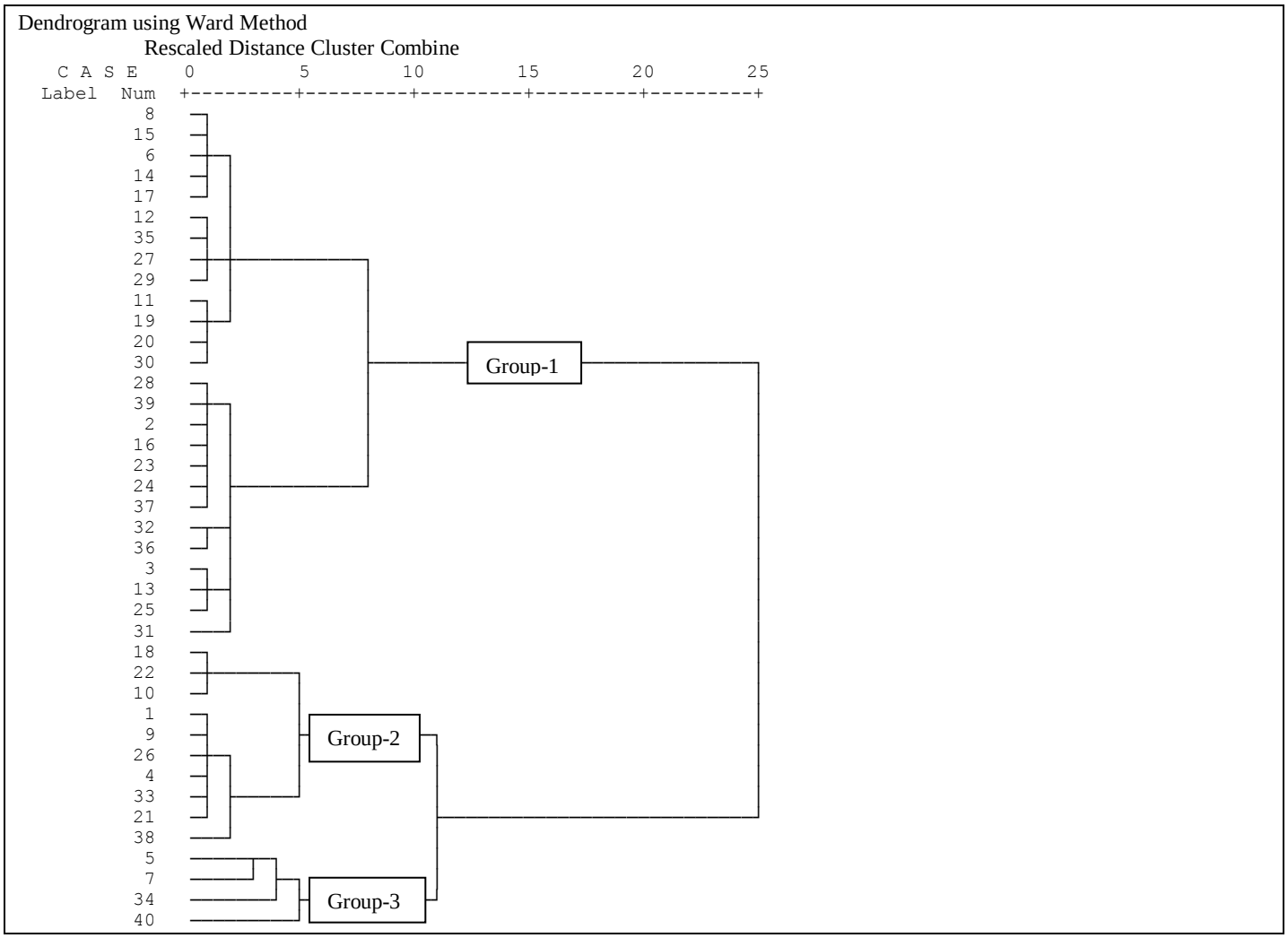


Fig-4 (a): Dendrogram of sampling stations by HCA in pre-monsoon period

Group-1 (G1) includes 26 samples from S2, S3, S6, S8, S11, S12, S13, S14, S15, S16, S17, S19, S20, S23, S24, S25, S27, S28, S29, S30, S31, S32, S35, S36, S37 and S39 sampling stations and it occupies 65%in pre-monsoon period [Fig-4 (a)] whereas in monsoon period G1 includes 24 samples S1, S2, S3, S6, S8, S11, S12, S13, S14, S15, S17, S19, S20, S23, S24, S27, S28, S29, S30, S32, S35, S36, S37, and S39 only and it occupies 60% [Fig-4 (b)]. For post-monsoon period, 20 samples S2, S6, S8, S11, S12, S14, S15, S16, S17, S19, S20, S23, S24, S27, S28, S29, S30, S35, S37 and S39 fall in this range and occupies 50% in this period [Fig. 4 (c)]. This type of water (G1) is relatively fresh with a mean EC of 970.04µS/cm, 749.75µS/cm and 946.30µmhos/cm in pre-monsoon, monsoon and post-monsoon periods respectively, which is the characteristic of mixed water (Mg^{2+}-Ca^{2+}-HCO_3^--Cl^-) combination. This group is basically bicarbonate and chloride dominated and also has low concentration of sulphate.EC seems to be a major distinguishing factor, which increases with concentrations increasing in all major ions in following the order: G1, G2 & G3 (Table-2). [9]

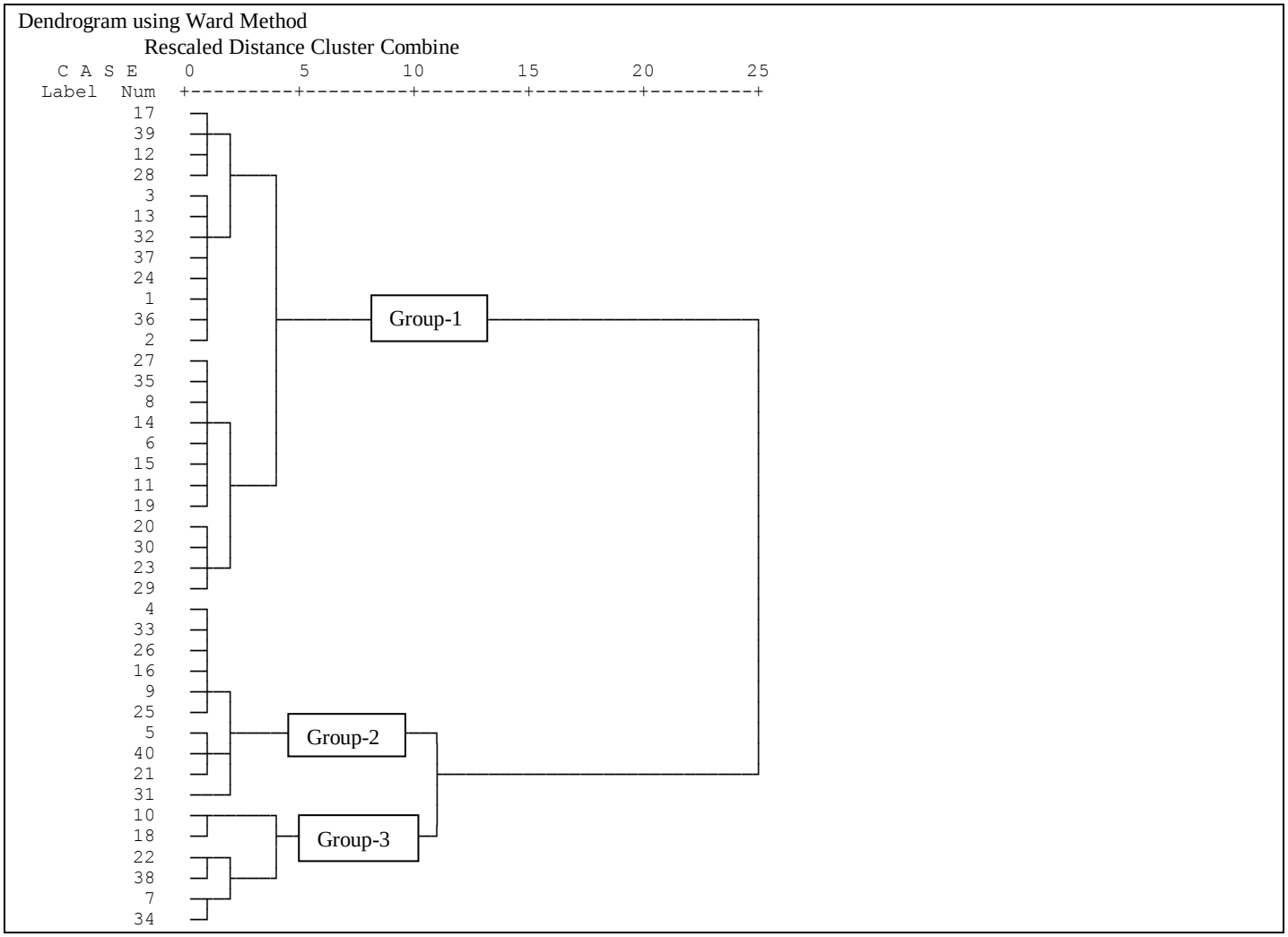


Fig-4 (b): Dendrogram of sampling stations by HCA in monsoon period

Group-2(G2) is represented by 10 samples collected from stations S1, S4, S9, S10, S18, S21, S22, S26, S33 & S38 and it occupies 25% of the water samples [Fig-4 (a)] while 10 samples of stations S4, S5, S9, S16, S21, S25, S26, S31, S33& S40 occupy 25% of the water samples [Fig-4 (b)] and 14 samples S1, S3, S4, S9, S13, S21, S22, S25, S26, S31, S32, S33, S36 & S38 occupy 35% of the water samples [Fig-4 (c)] in pre-monsoon, monsoon and post-monsoon periods respectively. EC in pre-monsoon, monsoon and post-monsoon periods was found to be 2106.50µS/cm, 1527.40µS/cm and 1818.79µmhos/cm (Table-2) respectively, which is the characteristic of blended water (Mg^{2+}-Ca^{2+}-HCO_3^--Cl^-). Chloride content is also high with respect to bicarbonate concentration.

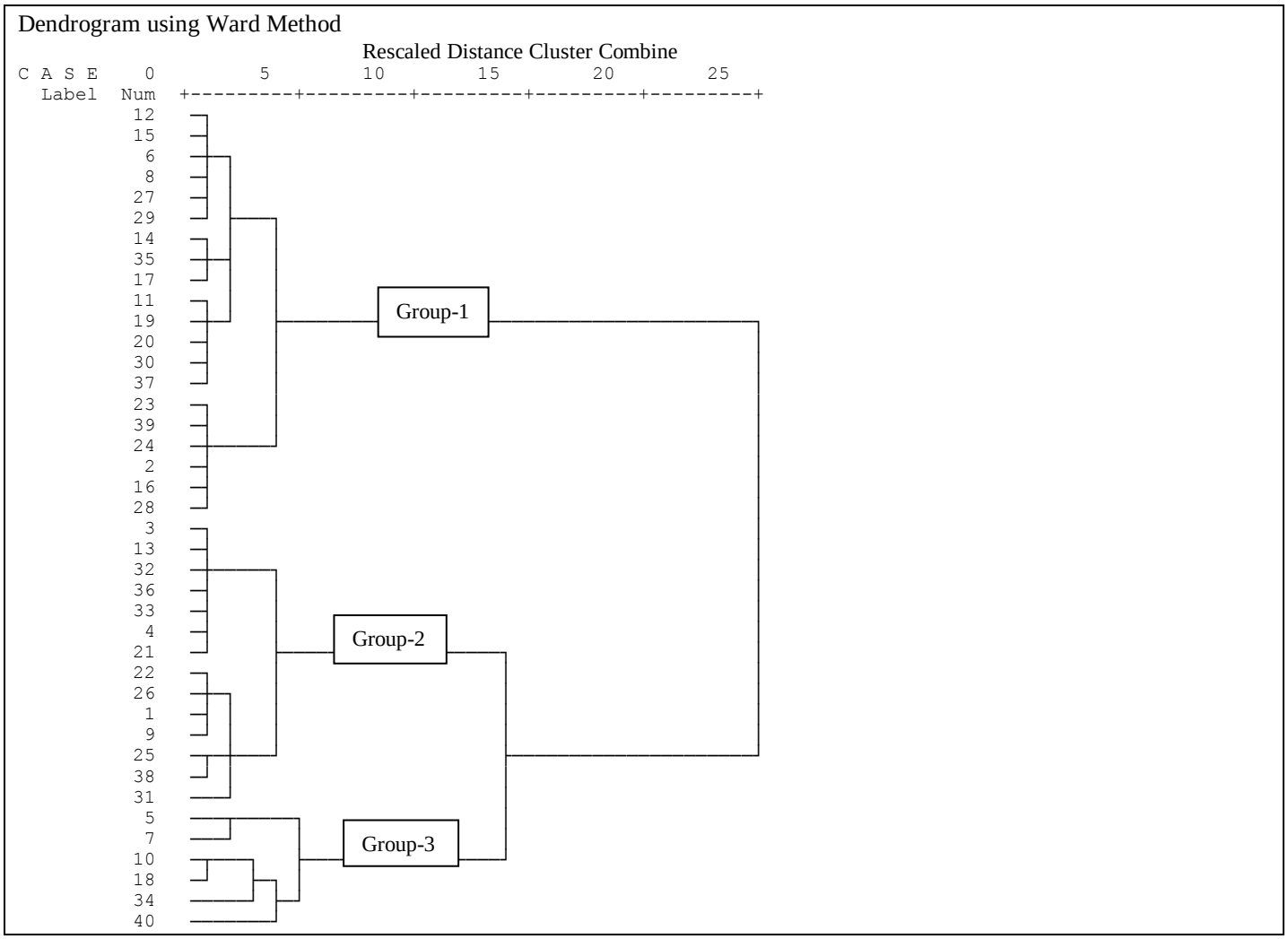


Fig-4 (c): Dendrogram of sampling stations by HCA in post-monsoon period

Group-3(G3) is composed of four samples S5, S7, S34 and S40 only and relates to 10% of water samples [Fig-4 (a)] having EC 3425.00µS/cm in pre-monsoon period, six samples S7, S10, S18, S22, S34 and S38 relate to 15% of the water samples [Fig-4 (b)] having EC 2658.33 µS/cm in monsoon period and six samples S5, S7, S10, S18, S34 and S40 relating to 15% of the water samples [Fig-4 (c)] having EC 3248.67µmhos/cm in post-monsoon period, which is the characteristic of high blended water (Mg^{2+}-Ca^{2+}-HCO_3^--Cl^-) composition. Chloride content is too high with high concentration of bicarbonate (Table-2).

Table-2: Mean parameter values of 3 principle groups determined from HCA in different periods

Parameters	Pre-monsoon period	Monsoon period	Post-monsoon period

	Group-1 (n=20)	Group-2 (n=14)	Group-3 (n=6)	Group-1 (n=20)	Group-2 (n=14)	Group-3 (n=6)	Group-1 (n=20)	Group-2 (n=14)	Group-3 (n=6)
EC(µS/cm)	970.04	2106.50	3425.00	749.75	1527.40	2658.33	946.30	1818.79	3248.67
pH	8.18	7.88	7.80	7.82	7.74	7.55	8.33	8.34	7.99
HCO_3^-(mg/L)	276.31	134.80	442.25	223.46	220.20	182.33	305.60	250.43	382.00
CO_3^{2-}(mg/L)	13.96	17.70	19.00	14.58	17.50	17.33	24.40	17.43	34.33
Ca^{2+}(mg/L)	56.69	36.44	159.00	46.33	48.20	57.27	66.96	68.03	98.47
Mg^{2+}(mg/L)	17.06	26.04	61.80	22.19	21.29	48.64	20.18	35.28	49.28
Cl^-(mg/L)	133.38	297.35	513.15	112.63	191.20	230.50	138.30	248.50	468.67
SO_4^{2-}(mg/L)	40.88	66.85	114.88	30.50	48.20	82.17	45.85	55.93	106.67
Na^+(mg/L)	114.61	170.80	389.00	89.00	150.70	110.67	133.80	153.79	329.00
K^+(mg/L)	3.08	3.30	2.50	2.13	2.70	2.33	3.35	3.36	2.50

HCA has also been employed on different physico-chemical variables analyzed for forty groundwater samples in pre-monsoon, monsoon and post-monsoon periods. Clusters based on eighteen variables are also classified into three groups; Group-1, Group-2 and Group-3 in all three periods and are shown in Fig-5 (a), (b) & (c). These groups are grouped on similarities of variables.

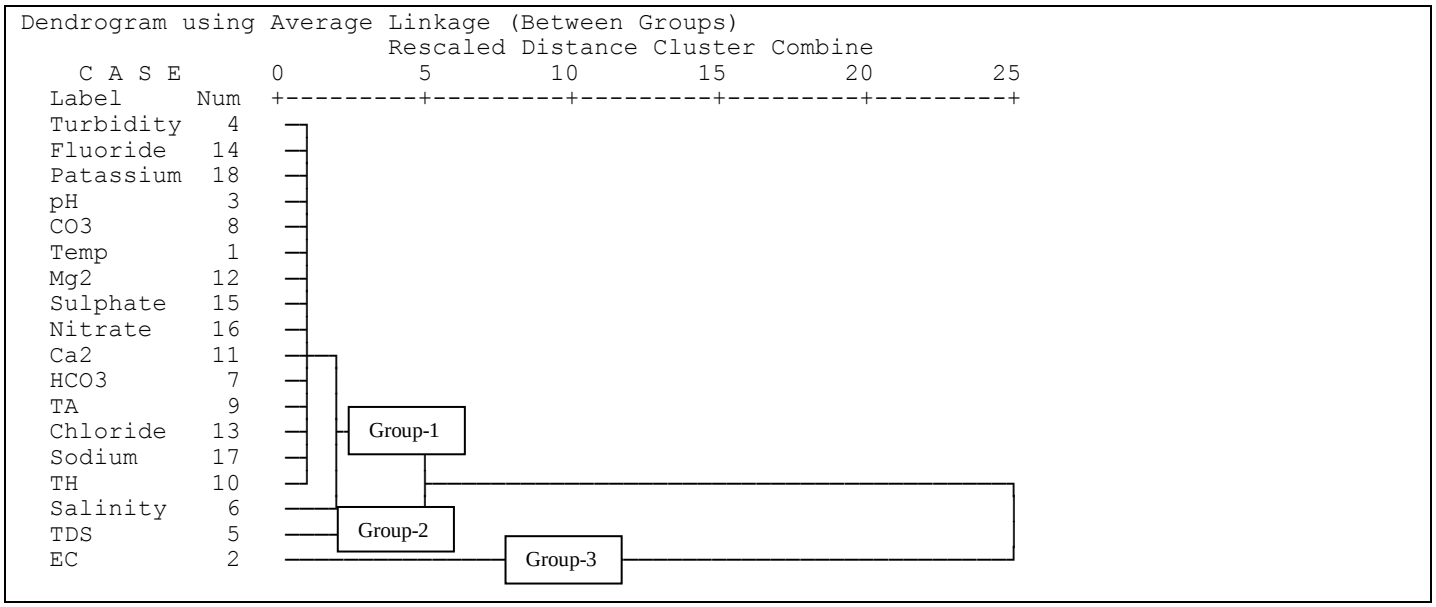
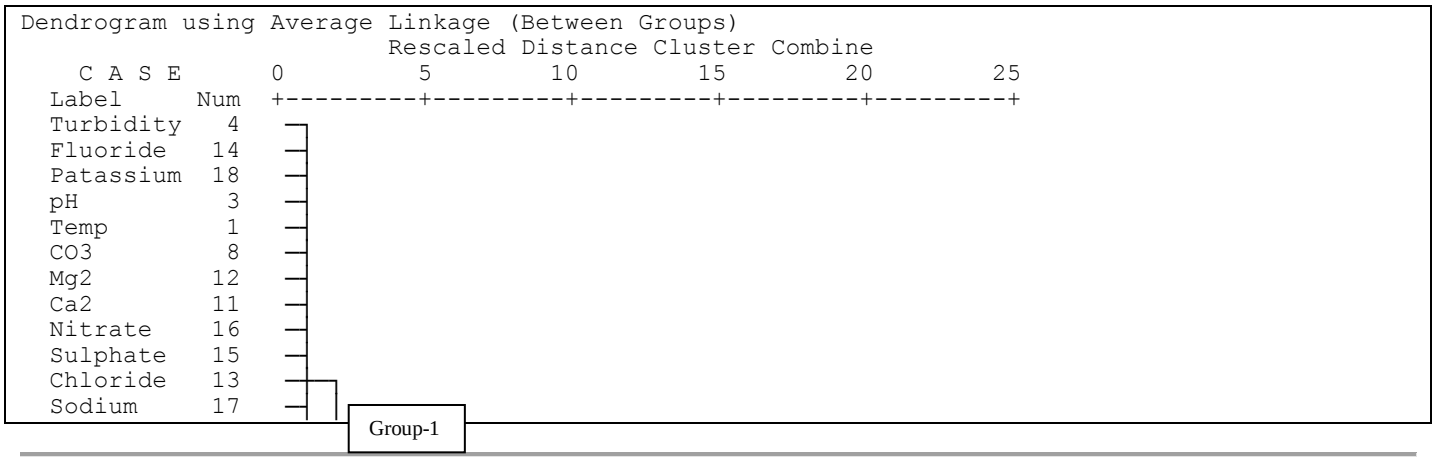


Fig-5 (a): Dendrogram of variables by HCA in pre-monsoon period

Most of the variables were classified in Group-1 with good correlation between pH, temperature, turbidity, SO_4^{2-}, Ca^{2+}, Mg^{2+}, HCO_3^-, CO_3^{2-}, TA, TH, Cl^-, F^-, NO_3^-, salinity, Na^+ & K^+ with EC and TDS in all three periods [Fig-5 (a), (b) & (c) and Table-3].



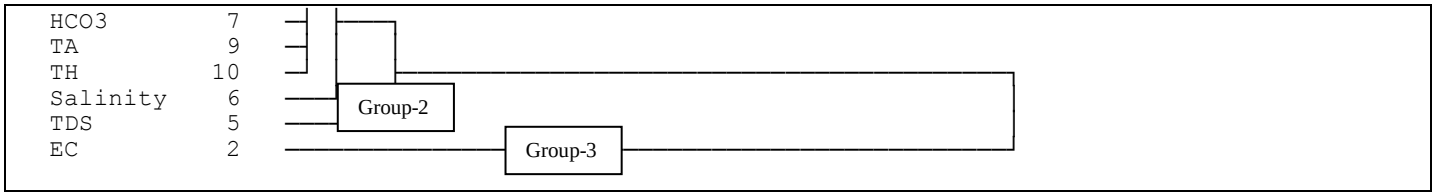


Fig-5 (b): Dendrogram of variables by Hierarchical cluster analysis in monsoon period

Group-2 & Group-3 includes only one parameter TDS & EC respectively in all three periods as shown in Fig-5 (a), (b) & (c). The possible salt combinations $CaSO_4$, NaCl and mixed Ca^{2+}-Na^{+}-HCO_3^{-} are probably derived from weathering of rock salts and irrigation return flow.

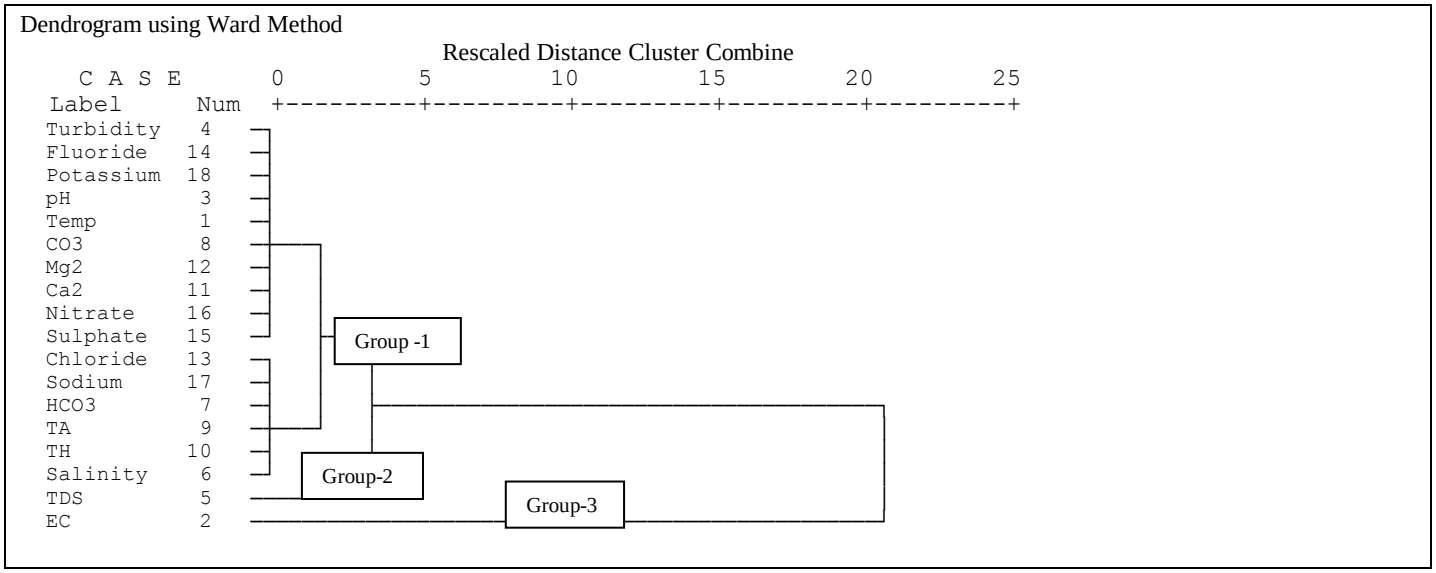


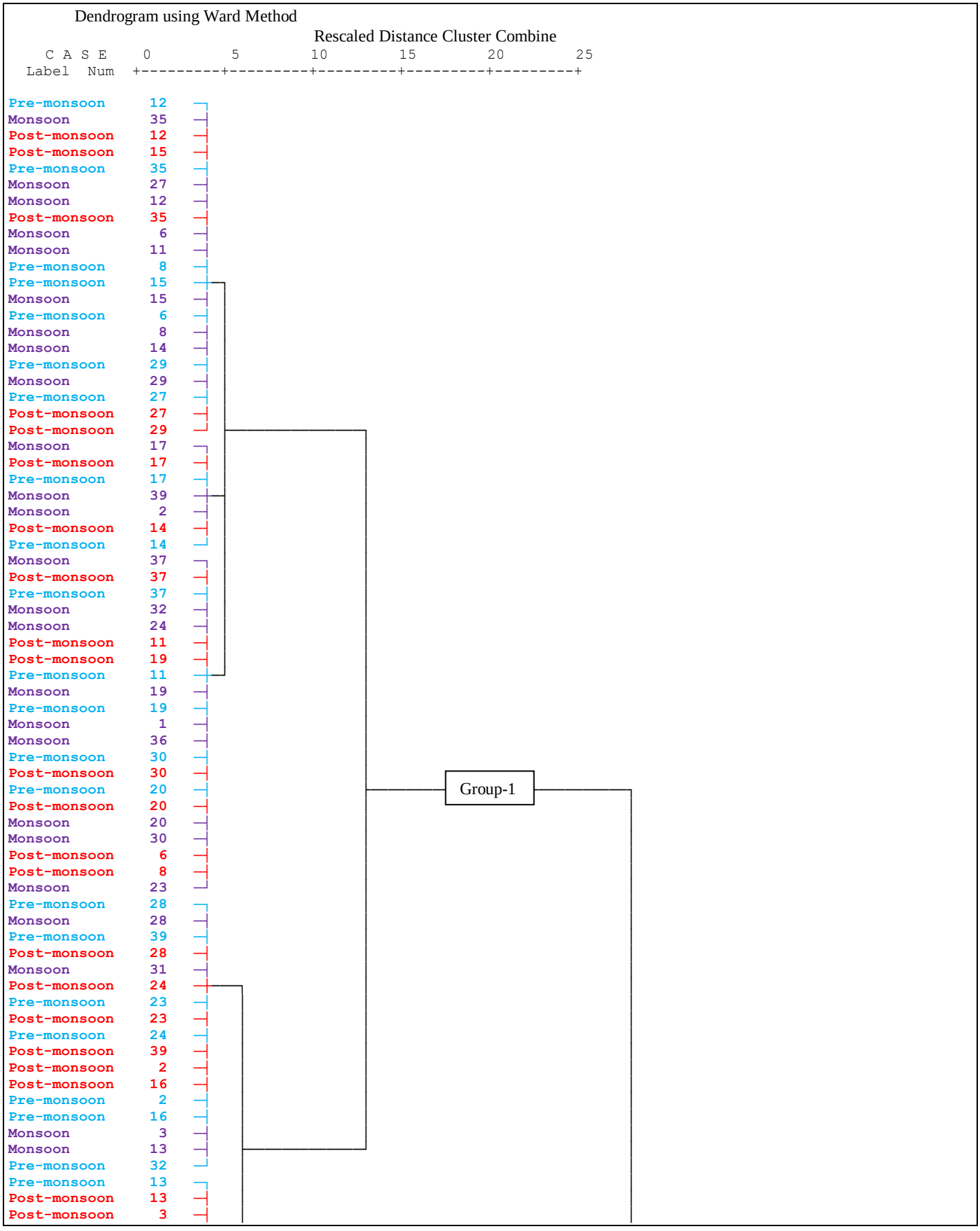
Fig-5 (c): Dendrogram of variables by HCA in post-monsoon period

According to cluster analysis, EC appears to be a major parameter on the basis of both sample station cluster groups and variable cluster groups, which affects all other groups and related parameters.

Table-3: Group of Clustered stations and variables of pre-monsoon, monsoon and post-monsoon periods

Group	Sampling stations			Variables		
	Pre-monsoon	Monsoon	Post-monsoon	Pre-monsoon	Monsoon	Post-monsoon
G-1	S2, S3, S6, S8, S11 to S17, S19, S20, S23 to S25, S27 to S32, S35 to S37 and S39	S1 to S3, S6, S8, S11 to S15, S17, S19, S20, S23, S24, S27 to S30, S32, S35 to S37, and S39	S2, S6, S8, S11, S12, S14 to S17, S19, S20, S24, S27 to S30, S35, S37 and S39	Turbidity, F-, K+, pH, CO32-, Temp., Mg2+, SO42-, NO3-, Ca2+, HCO3-, TA, Cl-, Na+, TH, Salinity		
G2	S1, S4, S9, S10, S18, S21, S22, S26, S33 & S38	S4, S5, S9, S16, S21, S25, S26, S31, S33 & S40	S1, S3, S4, S9, S13, S21, S22, S25, S26, S31 to S33, S36 & S38	TDS		
G3	S5, S7, S34 and S40	S7, S10, S18, S22, S34 and S38	S5, S7, S10, S18, S34 and S40	EC		

Similarities are observed between different sample stations for pre-monsoon, monsoon and post-monsoon periods in Table-3 and Fig-6. Similarities also indicate that samples are highly contaminated in post-monsoon period as compared to pre-monsoon and monsoon periods. Variable groups show similarities in all three periods, in which EC emerges out to be a major parameter.


Dendrogram using Ward Method
Rescaled Distance Cluster Combine
C A S E
Label Num
0
5
10
15
20
25
Pre-monsoon 12
Monsoon 35
Post-monsoon 12
Post-monsoon 15
Pre-monsoon 35
Monsoon 27
Monsoon 12
Post-monsoon 35
Monsoon 6
Monsoon 11
Pre-monsoon 8
Pre-monsoon 15
Monsoon 15
Pre-monsoon 6
Monsoon 8
Monsoon 14
Pre-monsoon 29
Monsoon 29
Pre-monsoon 27
Post-monsoon 27
Post-monsoon 29
Monsoon 17
Post-monsoon 17
Pre-monsoon 17
Monsoon 39
Monsoon 2
Post-monsoon 14
Pre-monsoon 14
Monsoon 37
Post-monsoon 37
Pre-monsoon 37
Monsoon 32
Monsoon 24
Post-monsoon 11
Post-monsoon 19
Pre-monsoon 11
Monsoon 19
Pre-monsoon 19
Monsoon 1
Monsoon 36
Pre-monsoon 30
Post-monsoon 30
Pre-monsoon 20
Post-monsoon 20
Monsoon 20
Monsoon 30
Post-monsoon 6
Post-monsoon 8
Monsoon 23
Pre-monsoon 28
Monsoon 28
Pre-monsoon 39
Post-monsoon 28
Monsoon 31
Post-monsoon 24
Pre-monsoon 23
Post-monsoon 23
Pre-monsoon 24
Post-monsoon 39
Post-monsoon 2
Post-monsoon 16
Pre-monsoon 2
Pre-monsoon 16
Monsoon 3
Monsoon 13
Pre-monsoon 32
Pre-monsoon 13
Post-monsoon 13
Post-monsoon 3
Group-1

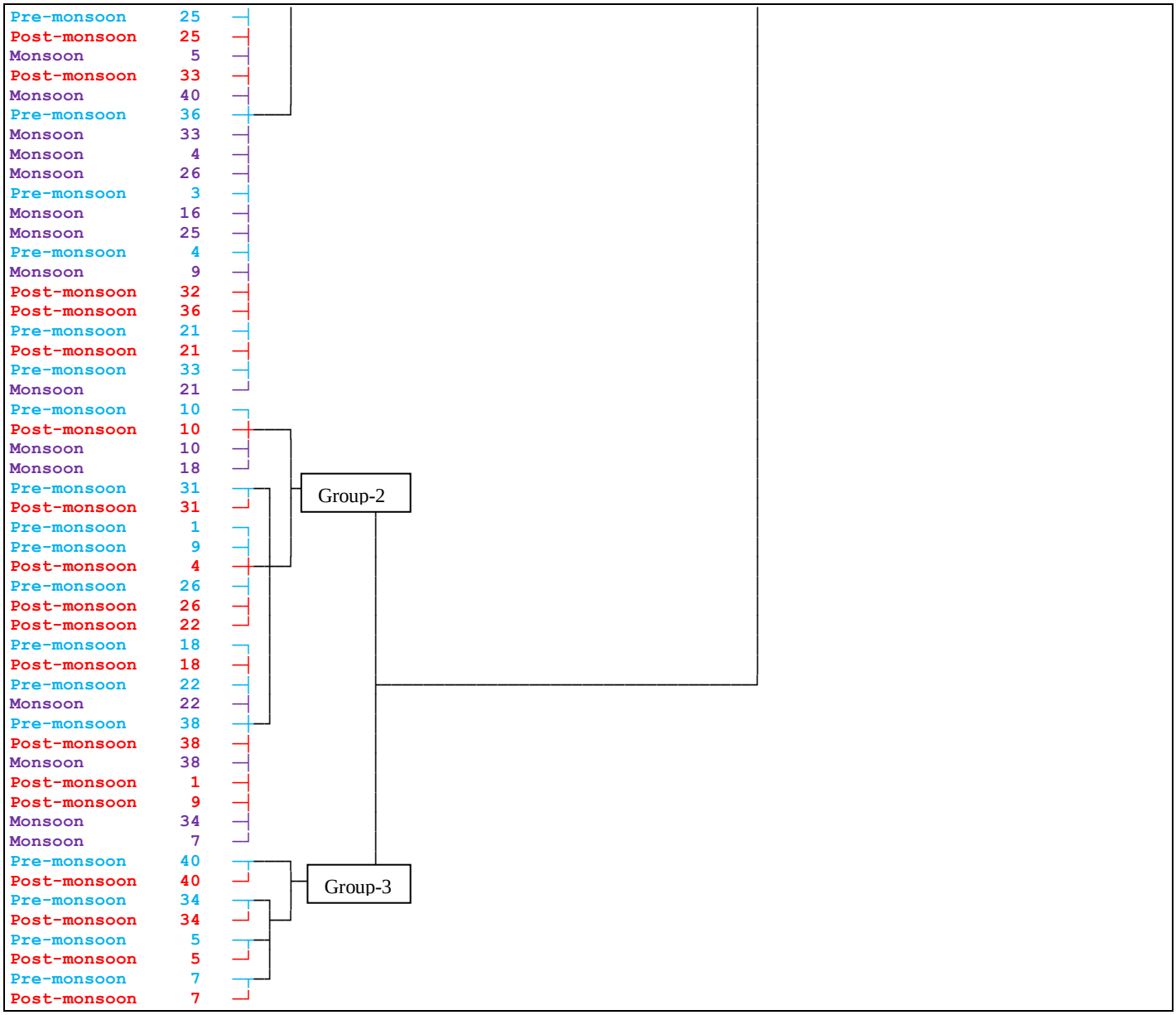


Fig-6: Dendrogram of variables by Hierarchical cluster analysis in monsoon period

As shown in Fig-6 sampling stations of Sanganer tehsil in pre-monsoon, monsoon and post-monsoon periods are clustering into three groups. Group-1 includes 89 sample stations and concerns 74.17% of the water samples, Group-2 having 23 sample stations occupies 19.17% and Group-3 includes 8 sample stations with a concern for 6.66% of the water samples. Pre-monsoon and post-monsoon period samples are very much contaminated as compared to monsoon period as no sample station of monsoon period is found in Group-3.

4. CONCLUSION

The study focused on similarities and dissimilarities in the group by cluster analysis technique of data mining and samples are classified in three major water types by HCA analysis in both sample stations clusters and in variable clusters, as mixed water, blended water and highly blended water. In groundwater samples in all dendrogram of clusters, EC seems to be a major factor which is characteristic of blended water (Mg^{2+}-Ca^{2+}-HCO_3^--Cl^-) composition. Chloride content is high with respect to bicarbonate concentration. However, fluoride levels in some samples do raise questions about potable drinking water. Hydrochemical analyses also indicate that parameters in some samples fall above recommended limits of W.H.O. and are thus less suitable or unsuitable for domestic purposes. In a nut shell, it can be concluded that overall quality of water has to be ensured for making it safe for drinking purpose in context to Sanganer tehsil.

REFERENCES

[1]. T. Balasubramanian and R. Umarani, "Clustering: An Analysis Technique in Data Mining for Health Hazards of High Levels of Fluoride in Potable Water", *IJCSET*, Vol. 2 (4), pp. 1113-1117, April 2014.

[2]. H. Hosseinimarandi, Md. Mahdavi, H. Ahmadi, B. Motamedvaziri and A. Adelpur, "Assessment of Groundwater Quality Monitoring Network Using Cluster Analysis, Shib-Kuh Plain, Shur Watershed, Iran", *Journal of Water Resource and Protection*, Vol. 6, pp. 618-624, April 2014.

[3]. APHA methods 3111: Standard methods for the examination of water and waste water, *American Public Health Association, Washington*, DC, 2005.

[4]. WHO, Guidelines for drinking water quality, Recommendations, *World Health Organization*, Geneva, 1, pp. 188, 1996.

[5]. Ting-Nien Wu and Chiu-Sheng Su, "Application of Principal Component Analysis and Clustering to Spatial Allocation of Groundwater Contamination", *Fifth International Conference on Fuzzy Systems and Knowledge Discovery, IEEE, Computer Society*, pp. 236-240, 2008.

[6]. Kamakshaiah K., R. Seshadri, "Ground Water Quality Assessment using Data Mining Techniques", *International Journal of Computer Applications*, Vol. 76 (15), pp. 39-45, August 2013.

[7]. R. Xu and D. Wunsch, "Survey of Clustering Algorithms", *IEEE transactions on neural networks*, Vol. 16 (3), pp. 645-678, May 2005.

[8]. R. Purviya, H. L. Tiwari and S. Mishra, "Application of Clustering Data Mining Techniques in Temporal Data Sets of Hydrology: A Review", *International Journal of Scientific Engineering and Technology*, Vol. 3 (4), pp. 359-363, April 2014.

[9]. L. Belkhiri, A. Boudoukha, L. Mouni and T. Baouz, "Multivariate statistical characterization of groundwater quality in Ain Azel plain, Algeria, *African Journal of Environmental Science and Technology*, Vol. 4 (8), pp. 526-534, August 2010.