Use of Cluster Analysis-A Data mining tool for improved water quality monitoring of River Satluj

Neetu Arora
Department of Information Technology, Marwadi Education Foundation, Rajkot-360003, Gujarat
Email: neetu.arora@marwadieducation.edu.in

Amarpreet Singh Arora
Department of Environmental Science and Engineering, Marwadi Education Foundation, Rajkot-360003, Gujarat
Email: enviro_amar@yahoo.com

Siddhartha Sharma
Manager Environment, Amritsar Textile Processor Association, Amritsar-143001, Punjab
Email: siddharthas79@gmail.com

Dr. Akepati S. Reddy
School of Energy and Environment, Thapar University, Patiala-147004, Punjab
Email: siva19899@gmail.com

---ABSTRACT---

Rapid urbanization often witness deterioration of regional river quality. As part of the management process, it is important to assess the baseline characteristics of the river basin environment so that sustainable development can be pursued. Over recent years a huge library of data mining algorithms has been developed to tackle a variety of data-rich environmental problems.

Data mining is now becoming the most popular technique for handling huge amount of environmental or water quality data. Certain techniques such as Artificial Neural Networks, Clustering, Case-Based Reasoning and more recently Bayesian Decision Networks have found application in environmental modeling.

In the current research study data mining technique, cluster analysis (CA) was applied to a large environmental data set of chemical and micro-biological indicators of river water quality. The study was carried out by using long-term water quality monitoring data of river Satluj. The results obtained allowed detecting natural clusters of monitoring locations with similar water quality type and identifying important discriminant variables responsible for the clustering. This objective separation or clustering could lead to an optimization of river monitoring techniques and are believed to be valuable to water resources managers in understanding complex nature of water quality issues to improve water quality.

Keywords – **Cluster analysis, Data mining, river water quality. Optimization, Satluj watershed**

1. INTRODUCTION

Real hydrochemical data sets contain not only important information useful for quality assessment and/or treatment technology but also confusing noise. Mostly, measured variables are not normally distributed, often co-linear or autocorrelated, containing outliers, erroneous or nonsense values. In order to reveal mutual dependence or logical structures of data, there are several chemometric procedures generally called as data mining techniques. Some of them are based on the reduction of data dimensionality, such as principal component analysis [15, 16], factor analysis, cluster analysis [17, 18], independent component analysis [1], independent factor analysis [2], generative topographic mapping [3], etc.

The water environment quality issue is a subject of ongoing concerned with the development of economy in any country. Especially, in India, the water resources problems related to environmental degradation have increasingly been serious, because of the rapid industrialization and urban sprawl. Due to their roles in transporting domestic and industrial wastewater and non-point source pollutants from agricultural land in their vast drainage basins, rivers are among the most vulnerable water bodies to pollution. Anthropogenic influence (urbanization, industrial and agricultural activities, increasing consumption of water resources) as well as natural process (changes in precipitation inputs, erosion, weathering of crustal materials) degrade surface water quality and impair their use for drinking, industrial, agriculture, recreation or other purposes [4, 5]. In order to effectively manage and research river water environment, obtaining water environment quality parameter data is indispensable. Although the regular measurements needs doing much work, because of spatial and temporal variation of water environment quality, monitoring by regular measurements, which will provide a representative and reliable estimation of surface water quality, is necessary. The long-term monitoring for many profiles in different reach will generate a large and complex database, which needs a good approach to interpret [19]. The application of different multivariate statistical techniques, such as cluster analysis (CA), principal component analysis (PCA), factor analysis (FA), helps in the interpretation of complex data matrices to better understand the water quality and ecological status of the studied systems, allows the identification of possible factors that influence water environment systems and offers a valuable tool for reliable management of water resources [6, 7, 8]. In recent years, many studies related with these methods have been carried out. For instance, Boyacioglu and Boyacioglu [9] use the PCA and CA to classify the sampling sites and to identify the latent pollution source. Mendiguchía et al. [10] used the CA to divide a watershed to four zones with different water quality except using PCA to identify the main pollutants. According to the above researches, it can be concluded that these methods could be used to assess the relationships between variables and possible pattern in distribution of measured data. As a result, in the study, we mainly use CA to identify several zones with different water quality and PCA to find the most important factors that describe the natural and anthropogenic influences.

Usually, the trials are undertaken to assess river water quality or to optimize the monitoring procedure by: classifying sampling locations, revealing links between water quality parameters, identifying possible sources of pollution or modeling the contribution of the identified sources to the formation of the total concentration of the monitored chemical tracers.

In the present study, a large data matrix, obtained during a 7-year (April 2004–March 2011) monitoring program, is subjected to different multivariate statistical techniques to extract information about the similarities or dissimilarities between sampling sites, and the influence of possible source on the water quality parameters of River Satluj watershed. The specific objectives are to: (1) identify several zones with different water quality, (2) extract the parameters that are most important in assessing variations in river water quality of different zones, (3) find a good approach to assess the water quality reasonably. It can be helpful to the managers to understand the main pollutants of each cluster of sampling sites, and to take effective measures to manage the water resources, respectively.

2. Satluj River and its Basin

The Satluj river is one of the five great rivers of undivided Punjab. It originates in Tibetan Plateau, flows through the Himalayan Range in Indian state of Himachal Pradesh and enters in the plains of Punjab from the Shivalik hills near Nangal. It flows in east-west direction. Most of the river water is diverted for irrigation, municipal and industrial uses at the Nangal head-works and very little of the river water flows in the river beyond Nangal. Flow in the river includes the seepages and the water intentionally let into the river mainly due to the differences in hydal power generation at Bhakra and in water diversion at Nangal for meeting the varying irrigational water demands. Beyond Nangal the river flows across the Punjab plains for about 238 km and then joins the river Beas at the Harike wetland system, a manmade wetland system created from the construction of head-works for diverting the river water for agricultural, municipal and industrial uses [11]. In this 238 km stretch of the river, the Himachal Pradesh hills bordering the Punjab state and the agricultural plains and urban settlements of a major part of the Punjab state constitute the catchments of the Satluj river. Downstream to the confluence point with the Beas river, at the Harike head-works, the river water is again mostly diverted for irrigation, municipal and industrial uses.

Watershed area (between 31°45’N 74°57’E and 30°45’N 76°50’E) of the river in this stretch is about 10,880 km2. Out of which sub-watershed area for the main course of the Satluj river is 3053.18 km2 from Nangal head-works till Harike head-works. Eight monitoring locations within this sub-watershed (SAT-1 to SAT-8) are currently identified for monthly assessment of the water quality. Fig. 1 shows the Satluj watershed along with the sampling locations.

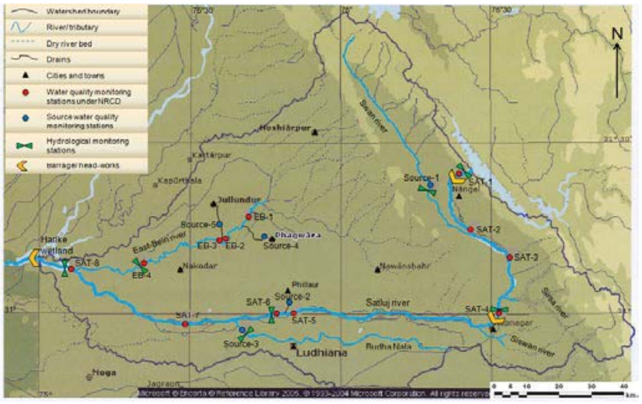

Figure 1: Map of Satluj river watershed showing monitoring stations.

3. Cluster analysis of water quality monitoring data

In the present study, optimization of the Satluj river water quality monitoring program was targeted through cluster analysis. The main objective was to reduce the number of monitoring stations (from the present 8), to reduce the frequency of sampling (from the present monthly sampling), and to minimize the number of parameters for which the samples should be tested. This optimization study effectively used the monitoring data accumulated during a 7 years period (April 2004-March 2011). The sampling strategy was designed in such a way to cover a wide range of determinants at key sites that accurately represent the water environment quality of the river system and account for tributary inputs that can have important impacts upon downstream water quality. Advanced statistical techniques (cluster analysis and principal component analysis) in combination with Water Quality Index (general purpose) approach were used to rationally optimize the monitoring program. Current study provides the details of the optimization study undertaken using cluster analysis.

Cluster Analysis (CA) groups the objects (cases) into classes (clusters) on the basis of similarities within a class and dissimilarities between different classes. The results of

CA help in interpreting the data and indicate the spatial and temporal patterns. In hierarchical clustering, clusters are formed sequentially by starting with the most similar pair of objects and forming higher clusters step by step [6, 12, 13].

CA was performed on the transformed water quality data sets by means of the Ward's method using squared Euclidean distance as a measure of similarity [14]. Monthly water quality data of the 8 monitoring stations on the Satluj River was subjected to cluster analysis. In order to interpret the cluster significance some degree of subjectivity was used. Final clustering was decided while giving due importance to the spatial and temporal consecutiveness. Spatial clustering was done first and the temporal data was grouped as per the spatial clustering prior to the temporal cluster analysis. Cluster significance was determined using the criterion of similarity factor greater than 0.8 and the most similar consecutive sampling sites in a cluster were considered as one Spatial Group (SG) and the most similar sampling months in a cluster were considered as one Temporal Group (TG). Based on the analysis, both sampling stations and sampling frequencies were rationalized.

Prior to CA, the descriptor variables (pollution indicators) were block standardized by range (autoscaling) to avoid any effects of scale of units on the distance measurements. In all clustering algorithms discussed in this paper, the squared Euclidean distance was used to measure similarity among clusters while Ward's method was used as an agglomeration technique.

Results obtained from the CA for spatial grouping are shown in Fig. 2.

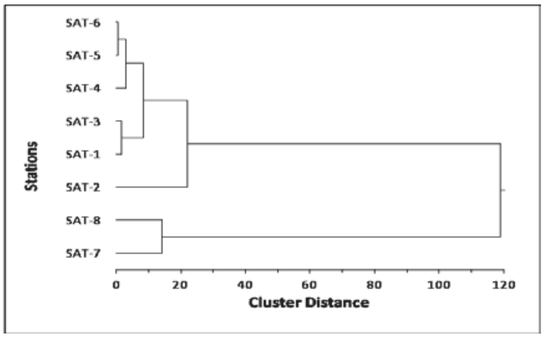

Figure 2: Cluster dendrogram showing similarity between different sampling stations of river Satluj.

The results indicate based on the similarities, four spatial groups (SGs) are possible. But when the condition of only adjacent stations along the river can be grouped together is applied, the existing 8 sampling stations can be divided into 6 SGs.

Table 1: Clustering strategy for development of Spatial Groups.

Station	Similarity Factor	Spatial Group
SAT-1	0.986	SG-1
SAT-2	0.815	SG-2
SAT-3	0.986	SG-3
SAT-4	0.976	SG-4
SAT-5	0.997	SG-5
SAT-6	0.997	
SAT-7	0.881	SG-6
SAT-8	0.881	

The similarities detected among the sampling stations and the SGs indentified on the basis of the similarities are presented in Table-1. The minimum similarity of 0.881 was observed among the two stations (SAT-7 and SAT-8) grouped into the SG-6.

CA was then applied for the 6 SGs independently for detecting temporal similarities and temporal grouping. The results obtained from the analysis are shown in Fig. 3(a) to Fig. 3(f). On the basis of the similarities among the 12 sampling months for each of the SG, temporal groups (TGs) were identified.

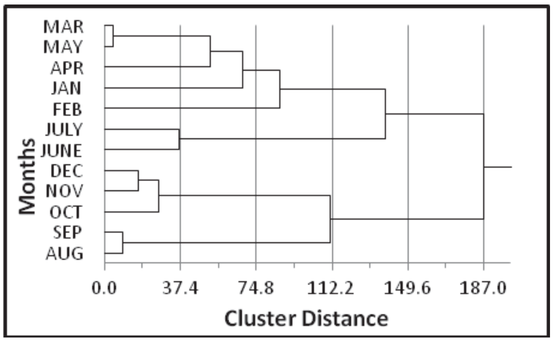

Figure 3(a): Cluster dendogram showing temporal grouping for SG 1

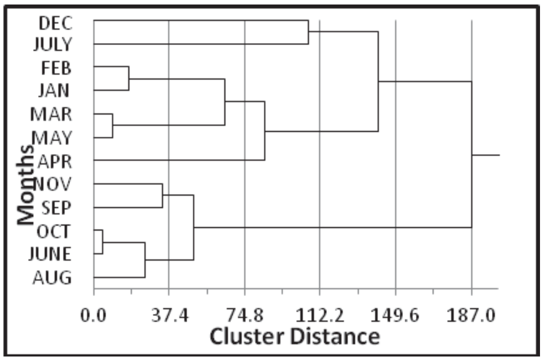

Figure 3(b): Cluster dendogram showing temporal grouping for SG 2

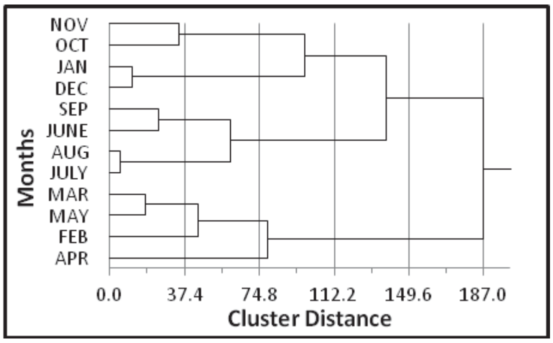

Figure 3(c): Cluster dendogram showing temporal grouping for SG 3

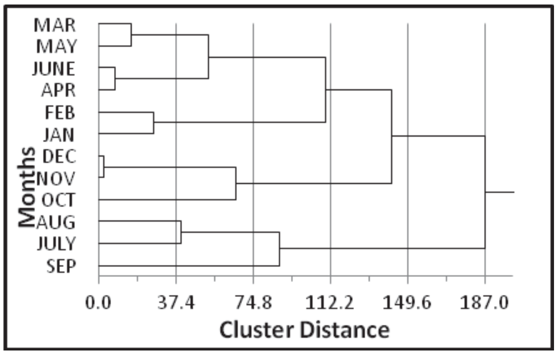

Figure 3(d): Cluster dendogram showing temporal grouping for SG 4

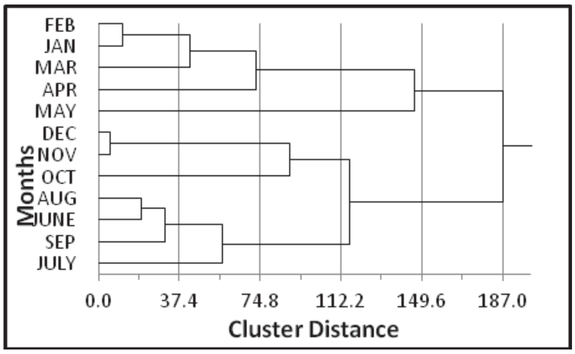

Figure 3(e): Cluster dendogram showing temporal grouping for SG 5

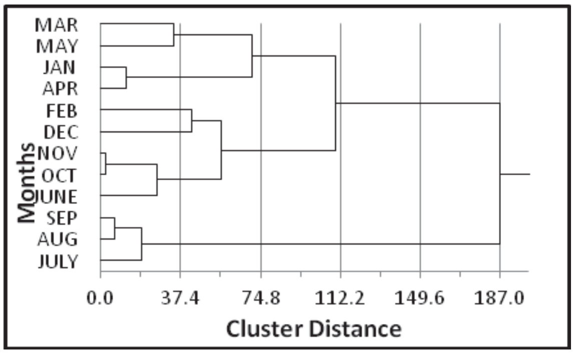

Figure 3(f): Cluster dendogram showing temporal grouping for SG 6

Table 2 provides the details of the temporal groups identified. Application of the condition that only the adjacent months of sampling events can be grouped, has marginally increased the number of TGs. Except in SG-5, close similarities were observed between March and May months in all the other 5 SGs. Since March and May are not two consecutive months, despite the similarities both the months were de-clustered.

From the cluster analysis, one can conclude that the number of sampling stations can be reduced from the present 8 to 6 stations. The future six monitoring stations can be S-1, S-2, S-3, S-4, S-5 or S-6 and S-7 or S-8. Similarly from the temporal cluster analysis one can conclude that the sampling frequency can be reduced to 8 months for SG-1, 11 months for SG-2, 9 months for SG-3, 10 months for SG-4, 10 months for SG-5 and 9 months for SG-6. Say for SG-1, the sampling can be avoided during June or July, August or September, and for two months during October to December. Table 3 provides the details on the reduced sampling stations and sampling frequencies for the 6 SGs. In summary, the sampling stations can be reduced from 8 to 6, and the total number of samplings over the year can be reduced to 57 from 96 (40.6% reduction).

4. Conclusion

The study offers a multivariate data mining strategy to assess the river water quality. The simultaneous performance of several classification and projection environ-metric methods makes it possible to detect not only the hidden factors responsible for the monitoring data structure for short- or long-term observation modes but to find some specific features of the locations of sampling, often different in short-term assessment as compared to the long-term one. In this case study, cluster analysis was used to evaluate spatial and temporal variations in surface water quality of Satluj river basin. Based on obtained information, it is possible to design an optimal sampling strategy, which could reduce the number of sampling stations, the frequency of sampling, the number of samples collected and associate costs and will also help to understand complex nature of water quality issues and determine priorities to improve water quality.

Factor Analysis and Principle Component Analysis will be effectively used in future studies to find inter-parameter associations existing between different pollutants. This data mining technique will further help in reducing the number of pollution parameters to be tested and subsequent cost of analysis. Also in future, the authors propose to use these data mining techniques for designing a water quality management system for Satluj River Basin. The finding of the current research study could be helpful to managers and government agencies in better water quality management for achieving the goal of sustainable use of the water resources.

5. Acknowledgements

The authors would like to thank the National River Conservation Directorate (NRCD), Ministry of Environment and Forests, Government of India, New Delhi, for providing technical and financial support for this research programme and Thapar University, Patiala for providing the laboratory facilities to undertake the task of testing and analysis.

References

[1] P. Comon, Independent component analysis, a new concept? *Signal Process., 36*(3), 1994, 287-314.

[2] H. Attias, Independent factor analysis, *Neural Computing, 11*(4), 1998, 803-851.

[3] C.M. Bishop, M. Svensen and C.K.I. Willams, GTM: The generative topographic mapping, *Neural Computing, 10*(1), 1998, 215-234

[4] S.R. Carpenter, N.E. Caraco, D.L. Correll, R.W. Howarth, and V.H. Smith, Nonpoint pollution of surface waters with phosphorus and nitrogen, *Ecological Applications*, *8*(3), 1998, 559–568.

[5] H.P. Jarvie, B.A. Whitton, and C. Neal, Nitrogen and phosphorus in east coast British rivers: speciation, sources and biological significance, *Science of the Total Environment*, *210*, 1998, 79-109.

[6] V. Simeonov, J.A. Stratis, C. Samara, G. Zachariadis, D. Voutsa, A. Anthemidis, M. Sofoniou, and T. Kouimtzis, Assessment of the surface water quality in northern Greece, *Water Research, 37*(17), 2003, 4119-4124.

[7] P. Praus, Water quality assessment using SVD-based principal component analysis of hydrological data, *Water SA, 31*(4), 2005, 417–422.

[8] S. Shrestha, and F. Kazama, Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan, *Environmental Modelling & Software, 22*(4), 2007, 464–475.

[9] H. Boyacioglu, and H. Boyacioglu, Water pollution sources assessment by multivariate statistical methods in the Tahtali Basin, Turkey, *Environmental Geology, 54*, 2007, 275–282.

[10] C. Mendiguchía, C. Moreno, M.D. Galindo-Riaño, and M. García-Vargas, Using chemometric tools to assess anthropogenic effects in river water: A case study: Guadalquivir River (Spain), *Analytica Chimica Acta, 515*(1,5), 2004, 143–149.

[11] S.K. Jain., A. Sarkar, and V. Garg, Impact of declining trend of flow on Harike wetland, India, *Water Resources Management, 22*, 2008, 409-421.

[12] Y. Zhang, F. Guo, W. Meng, X. Wang, Water quality assessment and source identification of Daliao river basin using multivariate statistical methods. *Environmental Monitoring and Assessment, 152*, 2009, 105-121.

[13] H. Razmkhah, A. Abrishamchi and A. Torkian, Evaluation of spatial and temporal variation in water quality by pattern recognition techniques: A case study on Jajrood River (Tehran, Iran). *Journal of Environmental Management, 91*, 2010, 852-860.

[14] K.P. Singh, A. Malik, and S. Sinha, Water quality assessment and apportionment of pollution sources of Gomti river (India) using multivariate statistical technique: a case study, *Analytica Chemica Acta, 538* (1-2), 2005, 355-374.

Books:

[15] BK. Lavine, *Clustering and classification of analytical data* (In: Meyers RA (ed.) Encyclopaedia of Analytical Chemistry, John Wiley & Sons, Chichester,2000).

[16] I.T. Jollifee, *Principal component analysis* (2nd edn. Springer-Verlag, New York, 2002).

[17] ER. Malinowski, *Factor analysis in chemistry* (2nd edn. John Wiley & Sons, New York, 1991).

[18] ER. Malinowski and DG. Howery, *Factor analysis in chemistry* (John Wiley & Sons, New York. 1980).

[19] D. Chapman, *Water quality assessment* (In: Chapman D. On behalf of UNESCO, WHO and UNEP. London: Chapman & Hall, 1992).

Authors Biography

Neetu Arora is presently working in the Department of Information Technology, Marwadi Education Foundation, Rajkot, Gujarat. She is currently doing research on Application of Data Mining techniques for extracting useful Environmental information. At present she is focusing on developing algorithms for extracting useful information concerning river water quality monitoring regimes.

Amarpreet Singh Arora is presently working as Assistant Professor in the Department of Environmental Science and Engineering, Marwadi Education Foundation, Rajkot, Gujarat. He has published, till date, eight research papers in peer reviewed international and national journals and conferences. He obtained his PhD degree from Thapar University, Patiala and specializes in the area of water resources management, sustainable urban stormwater management, wastewater treatment, river quality monitoring and watershed scale river basin management.

Siddhartha Sharma is presently working as Manager, Environment (Amritsar Textile Processor Association, Amritsar, Punjab). He has 3 years of industrial experience along with 7 years of research experience. He is about to complete his PhD degree (Thesis submitted) from Thapar University, Patiala and specializes in river water quality management.

Dr. Akepati S. Reddy is currently the Associate Professor and Head of School of Energy and Environment, Thapar University, Patiala. He has 22 years of teaching experience along with 20 years of research experience. He has till date, 30 publications (8 in international journals and 22 in various conferences). He obtained his M.Tech. Degree from IIT Bombay and PhD from Punjabi University, Patiala and specializes in the area of Environmental Technology, Industrial waste minimization through source reduction, recycling, reuse and resource recovery.

Table 2: Clustering strategy for development of Temporal Groups (TGs) for different SGs.

	SG-1		SG-2		SG-3		SG-4		SG-5		SG-6	
Months	**Similarity Factor**	**Temporal Group**	**Similarity Factor**	**Temporal Group**	**Similarity Factor**	**Temporal Group**	**Similarity Factor**	**Temporal Group**	**Similarity Factor**	**Temporal Group**	**Similarity Factor**	**Temporal Group**
Jan	0.635	$G_{1,1}$	0.907	$G_{2,1}$	0.940	$G_{3,1}$	0.857	$G_{4,1}$	0.939	$G_{5,1}$	0.937	$G_{6,1}$
Feb	0.536	$G_{1,2}$	0.907		0.763	$G_{3,2}$	0.857		0.939		0.773	$G_{6,2}$
Mar	0.976	$G_{1,3}$	0.949	$G_{2,2}$	0.905	$G_{3,3}$	0.914	$G_{4,2}$	0.774	$G_{5,2}$	0.816	$G_{6,3}$
Apr	0.721	$G_{1,4}$	0.548	$G_{2,3}$	0.579	$G_{3,4}$	0.956	$G_{4,3}$	0.611	$G_{5,3}$	0.937	$G_{6,4}$
May	0.976	$G_{1,5}$	0.949	$G_{2,4}$	0.905	$G_{3,5}$	0.914	$G_{4,4}$	0.218	$G_{5,4}$	0.816	$G_{6,5}$
June	0.804	$G_{1,6}$	0.977	$G_{2,5}$	0.868	$G_{3,6}$	0.956	$G_{4,5}$	0.893	$G_{5,5}$	0.859	$G_{6,6}$
July	0.804		0.430	$G_{2,6}$	0.973	$G_{3,7}$	0.788	$G_{4,6}$	0.693	$G_{5,6}$	0.898	$G_{6,7}$
Aug	0.951	$G_{1,7}$	0.863	$G_{2,7}$	0.973		0.788	$G_{4,7}$	0.893	$G_{5,7}$	0.965	
Sep	0.951		0.817	$G_{2,8}$	0.868	$G_{3,8}$	0.532	$G_{4,8}$	0.836	$G_{5,8}$	0.965	
Oct	0.857	$G_{1,8}$	0.977	$G_{2,9}$	0.816	$G_{3,9}$	0.644	$G_{4,9}$	0.527	$G_{5,9}$	0.987	$G_{6,8}$
Nov	0.911		0.817	$G_{2,10}$	0.816		0.986	$G_{4,10}$	0.971	$G_{5,10}$	0.987	
Dec	0.911		0.430	$G_{2,11}$	0.940	$G_{3,1}$	0.986		0.971		0.773	$G_{6,9}$

Table 3: Chart matrix for the suggested sampling frequency on the basis of cluster analysis

Months	**SG-1**	**SG-2**	**SG-3**	**SG-4**	**SG-5**	**SG-6**	**Sampling Frequency**
January	$G_{1,1}$	$G_{2,1}$	$G_{3,1}$	$G_{4,1}$	$G_{5,1}$	$G_{6,1}$	3
February	$G_{1,2}$		$G_{3,2}$			$G_{6,2}$	6
March	$G_{1,3}$	$G_{2,2}$	$G_{3,3}$	$G_{4,2}$	$G_{5,2}$	$G_{6,3}$	6
April	$G_{1,4}$	$G_{2,3}$	$G_{3,4}$	$G_{4,3}$	$G_{5,3}$	$G_{6,4}$	6
May	$G_{1,5}$	$G_{2,4}$	$G_{3,5}$	$G_{4,4}$	$G_{5,4}$	$G_{6,5}$	6
June	$G_{1,6}$	$G_{2,5}$	$G_{3,6}$	$G_{4,5}$	$G_{5,5}$	$G_{6,6}$	6
July		$G_{2,6}$	$G_{3,7}$	$G_{4,6}$	$G_{5,6}$	$G_{6,7}$	3
August	$G_{1,7}$	$G_{2,7}$		$G_{4,7}$	$G_{5,7}$		6
September		$G_{2,8}$	$G_{3,8}$	$G_{4,8}$	$G_{5,8}$		4
October	$G_{1,8}$	$G_{2,9}$	$G_{3,9}$	$G_{4,9}$	$G_{5,9}$	$G_{6,8}$	4
November		$G_{2,10}$		$G_{4,10}$	$G_{5,10}$		3
December		$G_{2,11}$	$G_{3,1}$			$G_{6,9}$	4
Sampling Frequency	8	11	9	10	10	9	**57**