Volume 66, Issue 4, 2022
Journal of Scientific Research
of
The Banaras Hindu University

Agglomerative and Divisive hierarchical cluster analysis of groundwater quality variables using open-source tools over YSR district, AP, India

Jagadish Kumar Mogaraju[*1]

[*1]Academic Counselor, Indira Gandhi National Open University, jagadishmogaraju@gmail.com

Abstract: **Groundwater quality variables like F, Total Hardness (TH), Total Alkalinity (TA), Total Dissolved Solids (TDS), SO_4, SAR, NA, EC, Cl, Ca, Mg, and pH were tested with Hierarchical clustering analysis (HCA) to identify the groupings or clusters that exist in the dataset. The dataset is subjected to Agglomerative and divisive hierarchical clustering. The observations were scaled to compare variables systematically. The clustering structure was determined using an agglomerative coefficient. Agglomerative approaches like complete, average, single, and ward are tested using agglomerative coefficients. The ward approach best suits the dataset to investigate a strong clustering structure. The agglomerative coefficient obtained is 0.8666752, and the divisive coefficient is 0.8371531. The entanglement score attained was 0.26, demonstrating a good alignment with nominal entanglement. The principal component analysis resulted in two main components contributing 54.8% and 18.2% explainable variance. The variables that are prominent in each PC are investigated and reported. The gap statistic and average silhouette method are used to know the optimal number of clusters. Open-source software like R/ R studio is used for this analysis. This work concludes that clustering analysis is essential to understand the groundwater quality variables better.**

Index Terms: **Hierarchical Cluster Analysis, Groundwater quality, Dendrogram, Principal Component Analysis, Entanglement**

I. INTRODUCTION

Groundwater is a genuine gift from nature to humanity. It is a rare and valuable product on the edge of a paradigm change(A. Elubid et al., 2019). While numerous attempts were made to conserve this natural resource, this argument for saving and restoring it is not uncommon(Prasad et al., 2019). The YSR District (Kadapa) is situated between 13º 43' and 15º 14' north latitude and 77º55' and 79º29' east longitude in Andhra Pradesh state of India(Vaidyanadhan, 1962). The geographical area is estimated to be around 15379 square kilometers. The Pennar River has five major tributaries that feed this area. Kunderu, Sagileru, Chitravati, Cheyyair, and Papagni are prominent. Pennar River is a perennial river that flows primarily from the northwest to the southeast(Vittala et al., 2005). The drainage pattern observed is dendritic and parallel, with sub-dendritic nature(Vaidyanadhan, 1962). Because the drainage is primarily parallel to subparallel, there is structural control. Black cotton, Red earth, Red sandy, and Red loamy soils abound in the research region. In 2012, the total forest area was estimated to be 500961 hectares(Balram et al., 2013). There were approximately 222099 hectares of barren and uncultivated land. Turmeric, bajra, jawar, cotton, groundnuts, and paddy are the main crops grown in this area. Alluvium, shale, sandstone, limestone, quartzite, dolomite, phyllite, schist, and granite-gneisses are the most common rocks found in this area. In the pre-monsoon season, the maximum depth to water level is around 17.35 m bgl, compared to 14.57 m bgl in the post-monsoon season. The aquifer's storage coefficient has been around 1.3 x 10-5. The peak transmissivity is around 884.7 sq.m/day, and the off-peak transmissivity is about 1.51 sq.m/day. The net groundwater availability is approximately 105039 ham(Ramakrishna Reddy et al., 2000). The hill ranges are oriented east to west or northwest to southeast. The prominent hill ranges in the region are the Lankamalai, Palakonda, Yerramalai, Nallamalais, and Velikonda. The highest point in the area was 1108 meters above sea level(Singh et al., 2019).

Geomorphologically, this area can be separated into three basic units. The kind of soil, slope factor, and relief all contribute to this classification. Fluvial, denudational, and structural landforms are the three categories(Nagaraju et al., 2016). Alluvial plains are a type of fluvial landform. Alluvial plains dominate fluvial landforms in Bazada and near major rivers. These flood plains were characterized as high-yielding zones or places. Bazada is primarily found in the foothills. Shallow aquifers appeared to have formed. Residual hills,

pediment zone, pediment-inselberg complex, and pediplains are examples of denudational landforms(Radhakrishna et al., 2007). The weathered pediplains have a moderate amount of groundwater. The pediment inselberg complex and shallow weathered pediplains have limited groundwater prospects. The landforms are intermontane valleys, linear ridges, Mesa/Buttee, cuesta, structural, and hills. Most landforms are found in the eastern section(Sreedhar & Nagaraju, 2017). Groundwater recharge can be substantial in the intermontane and valleys. The study area has various rock types, which date from the early Proterozoic and late Archean periods(Singh et al., 2019). Other rocks periodically replaced these rock types from the Dharwarian epoch(Anand et al., 2003). Dykes and dolerite cut through the majority of these rocks. The Kurnool group and Cuddapah supergroup have superimposed most of the older rocks. These rocks are from the upper and middle Proterozoic periods, respectively. On the previously denudated surfaces of the older rocks, a significant depression has formed. The pebbles are dispersed throughout the surrounding districts. Clay, silt, sand, and gravel create alluvium, which dominates the riverbeds. The quartz reefs and dykes have connected Dharwar and Archean. Migmatite, granite-gneiss, granodiorite, and granite are part of an Archean peninsular gneissic complex(.M et al., 2018). These rock types dominate the southern part of the research region. Rocks of varied ages dominate this area. Migmatites and gneisses are archean crystalline rocks. In these rocks, porosity is essentially non-existent. Secondary porosity is seen in some cases caused by fracturing and weathering(Gowd et al., 2019).

These processes increased the likelihood of groundwater being present in this area. Fractures and joints were evident in semi-confined and weathered settings favorable to groundwater. The weathered zone is approximately 10 meters thick. Groundwater was observed in the weathered site using excavated and bore wells with a diameter of about 6 meters. In contrast, deep and shallow bore wells are in the fractured zone. Fractured zones are found at intervals ranging from 8 to 145.80 m bgl. In addition, there were possible cracks at depths ranging from 20 m to 100 m bgl. The fractured zones affect the bore well yields, ranging from 1 to 3 lps. The fragmented zones reached from 0.1 to 4.9 lps(Goswami et al., 2021). Dolomites, limestones, quartzites, and shales dominate the Kadapa-Kurnool formations. The weathered section is roughly 10 m bgl thick(Kale et al., 2020). Most dug wells have dried up due to sustained groundwater pressure, as seen during dry years. The thickness of alluvium found near river courses ranges from 1 to 20 m bgl. The filter point wells reached a depth of 15 meters(Sreedhar & Nagaraju, 2017).

.

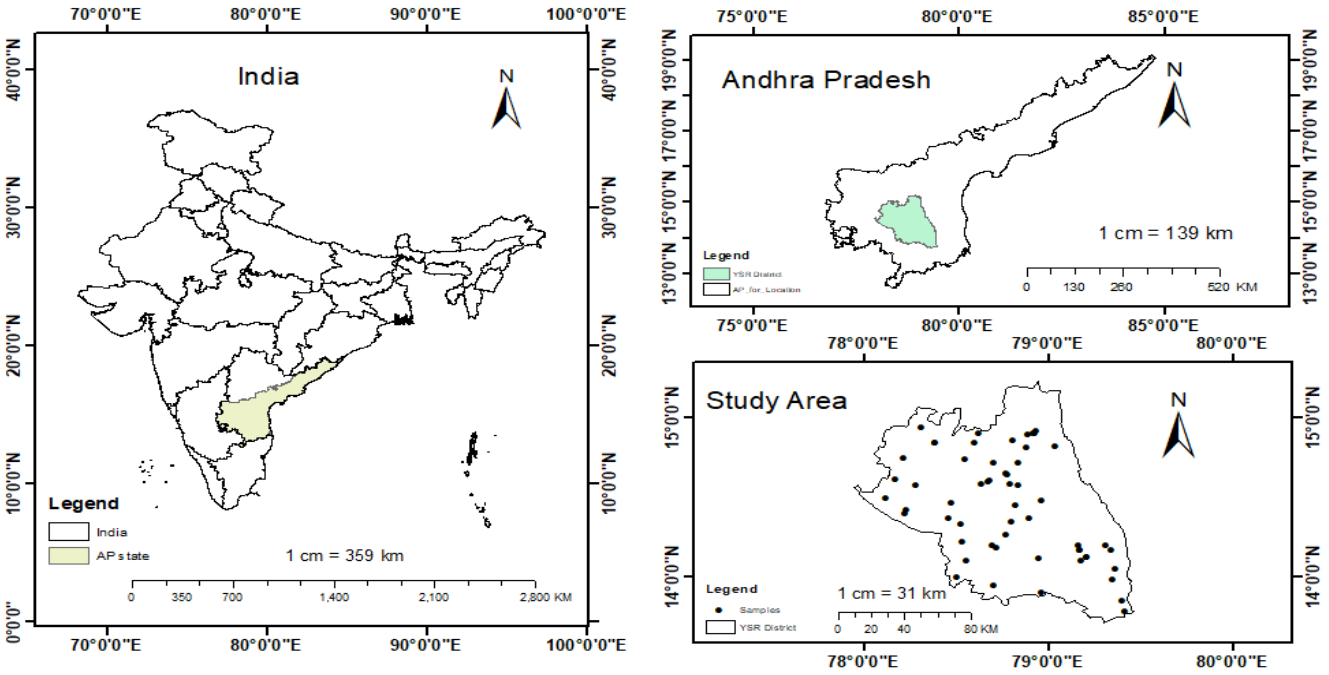

Fig. 1.Study area

II. Materials and Methods

A. Data

The datasets needed for this analysis are collected from the Central Groundwater Board (CGWB), Government of India. The datasets are available at http://cgwb.gov.in/GW-data-access.html and http://indiawris.gov.in/. The location map and data are given in Figure 1.

B. Hierarchical Cluster Analysis

Hierarchical clustering is an approach to compensate for the need for k-means clustering(Daughney et al., 2012). This clustering method can find groupings likely to persist in the

dataset. The cluster numbers need not be pre-specified in the hierarchical cluster analysis. Instead, we can generate a dendrogram, a tree-based representation of the observations. Then, the observation datasets are subjected to hierarchical cluster analysis using R software. This study used R packages such as tidyverse, cluster, factoextra, and dendextend(*R: The R Project for Statistical Computing*, n.d.). Hierarchical clustering can be classified into two forms, namely 1. Agglomerative clustering and 2. Divisive hierarchical clustering. Agglomerative nesting, or AGNES, is a type of agglomerative clustering. Every object in this category is classified as a leaf or a single element cluster. Two clusters with similar characteristics can be united into a relatively large cluster at each phase, referred to as a node. The operation will be repetitive until all observed points form a single cluster entity, referred to as a root. A dendrogram can be obtained using this method of analysis. DIANA, or Divise analysis, is a divisive hierarchical clustering method. We can see a top-down strategy in this analysis. It's possible to think of it as the inverse of AGNES. A single cluster can contain all of the observed items. Every iteration step in this approach ensures that the heterogeneous cluster is sorted into two. This round continues until all observed items are assembled in their native cluster. Next, small clusters can be investigated using agglomerative clustering.

Distance measurements such as Manhattan distance, Euclidean distance, and others are used to investigate the dissimilarity of the observations. Cluster agglomeration techniques can be used to examine any distinction between two clusters. These procedures are called linking methods. Complete linkage clustering is the same as maximum linkage clustering. We can use this method to look at pairwise dissimilarities between two clusters. The distance between the clusters under investigation can be taken to equalize the most considerable dissimilarity value found for each cluster. We can make compact clusters using this approach. Single linkage clustering is the simplest form of linkage clustering. This approach allows us to create long, loose clusters. Mean linkage clustering is a variation of average linkage clustering. We can use this approach to examine the pairwise dissimilarities between the two clusters. The average of the dissimilarity obtained between the two groups must be considered. By evaluating the centroids of two clusters, centroid linkage clustering may be used to examine their dissimilarity. Ward's minimal variance method will reduce the overall variation within a cluster. Cluster couples with the shortest distance will merge(Bouteraa et al., 2019).

The vertical axis' height represents the observation rate of dissimilarity and similarity. The rate of similarity can be inversely related to height. The height of the incision determines the number of clusters formed. By cutting the dendrograms, we can identify the sub-groups. The dendrogram generates argument borders around clusters. A scatter plot is a valuable tool for visualizing these groups. A comparison was made between Ward's technique and the entire linkage. The entanglement metric will be between 1 and 0, with 1 indicating whole entanglement and 0 indicating no entanglement and good alignment. The Elbow, Average Silhouette, and Gap statistic approaches created the best clusters(Rahbar et al., 2020).

C. Principal Component Analysis

The observed dataset was subjected to principal component analysis or PCA, which determined principal components. These fundamental elements are required to comprehend the data. PCA decreases the dimensionality of the observed data, which can be considered data reduction via feature extraction(Yang et al., 2015). The local variables will represent the variability, and PCA can prevent multicollinearity. Every variable is centered at zero to perform PCA, eliminating specific difficulties with a variable's scale(Aly et al., 2015). The magnitude of a variable can affect PCA, and the results are based on the variables that were previously scaled individually. PCA can provide a low-dimensional space to visualize our data with greater variation clarity. In p-dimensional space, ‘n’ observations can be accommodated. PCA linearly unifies the ‘p’ features, which results in dimensions. The magnitude of the variables seen in the dataset is significantly varied; thus, we can correlate.

III. Results

Nine principal components were obtained after subjecting the dataset to principal component analysis. The first principal component (PC) yielded 58% of explained variance , followed by second PC (18.2%), third PC (12.6%), fourth PC (5.8%), fifth PC (3.7%),sixth PC (2.3%), seventh PC (1.3%), eight PC (0.9%), and ninth PC with 0.3% explained variance. The main contributing PCs, such as the first and second PC, were considered (Figure 9). The First PC, EC, TDS, Cl, Na, SO_4, SAR, NO_3, Mg, TH, and TH, were contributing variance. In the second PC, Ca, TH, pH, RSC, Mg, F, SAR, TA, HCO_3, and K mainly contributed to the variance. A scree plot explaining the variance is given in figure 10.

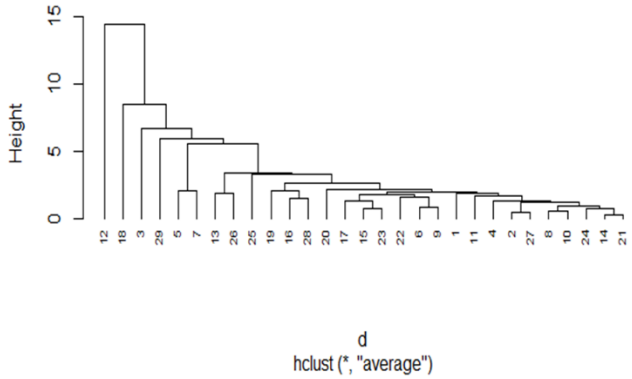

Fig. 2. Dendrogram using average method

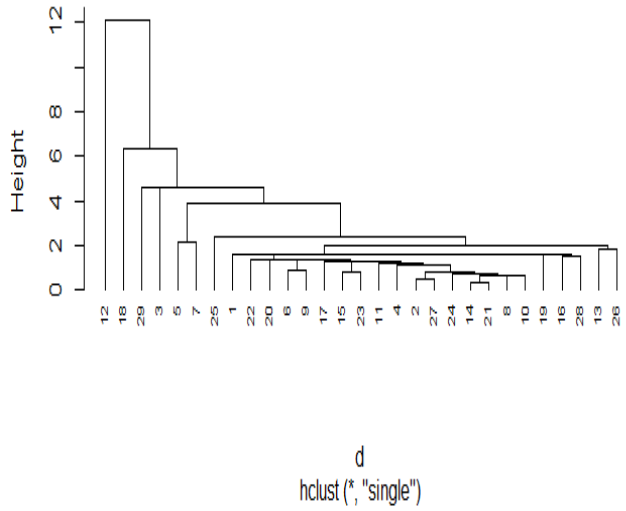

Fig. 3. Dendrogram using single method

The observed data is scaled to make proper comparisons between the variables. Complete, average, single, and ward agglomeration approaches were compared using the agglomerative coefficient. 0.8344129 (average technique), 0.8356995 (single method), 0.8396585 (complete method), and 0.8666752 (ward method) are the agglomeration coefficients generated (Figure 2,3,4,5,6). The cluster plot reflecting four clusters is shown in figure 7. The Ward approach is selected to find a stronger clustering structure since it has a greater agglomeration coefficient. The leaf of the dendrogram represents one observation. Similar observations were grouped into different branches and so on. The gap statistic and average silhouette method were used to know the optimal number of clusters (Figures 11 and 12). The entanglement score obtained from this dataset is 0.26 (Figure 8). This showed minimal entanglement and hence good alignment.

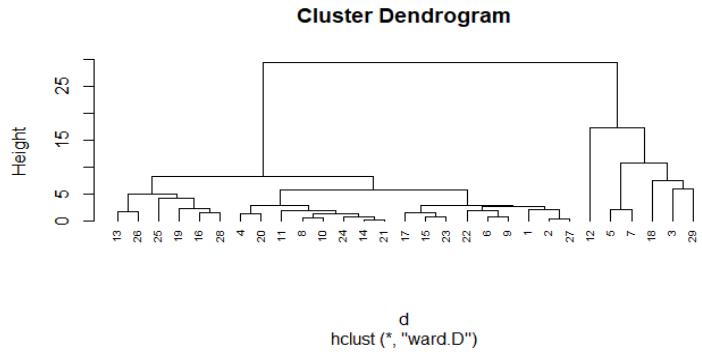

Fig. 4. Dendrogram using ward method

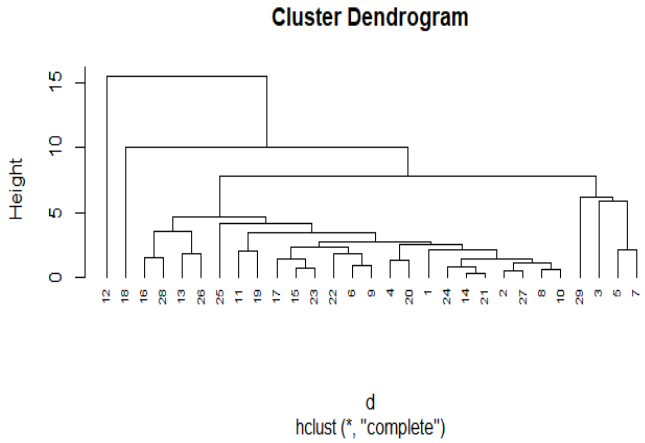

Fig. 5. Dendrogram using complete method

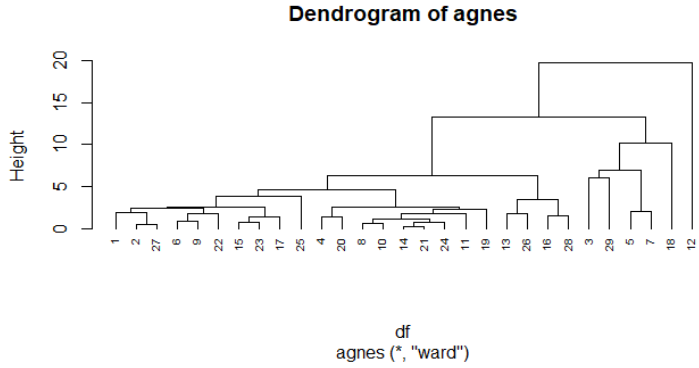

Fig. 6. Dendrogram using AGNES

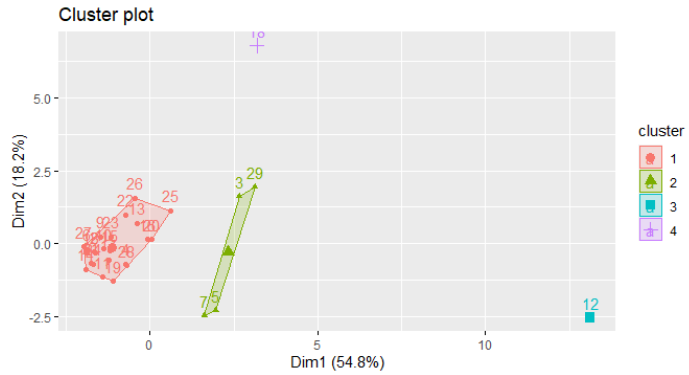

Fig. 7. Cluster plot

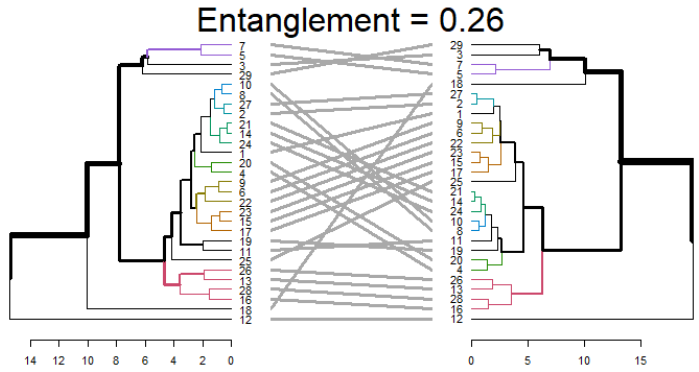

Fig. 8. Entanglement

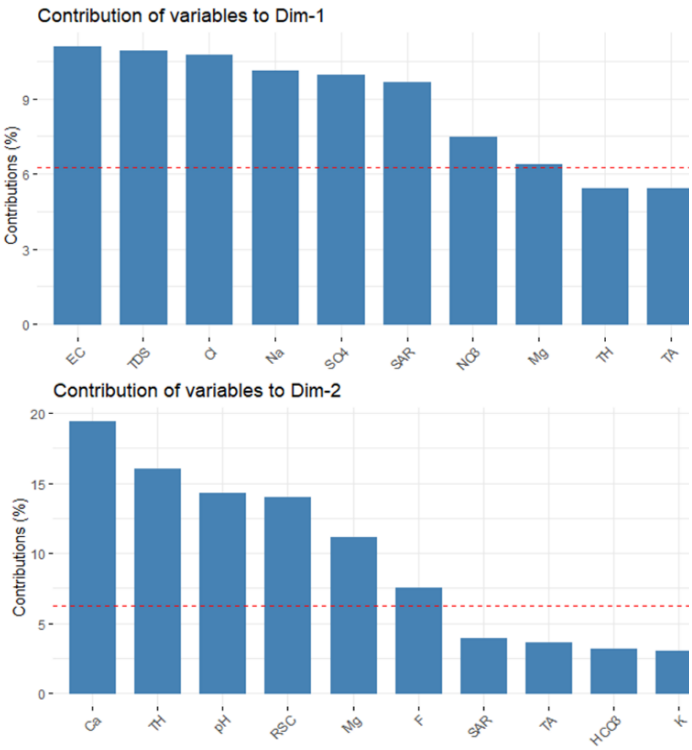

Fig. 9. PC1 and PC2

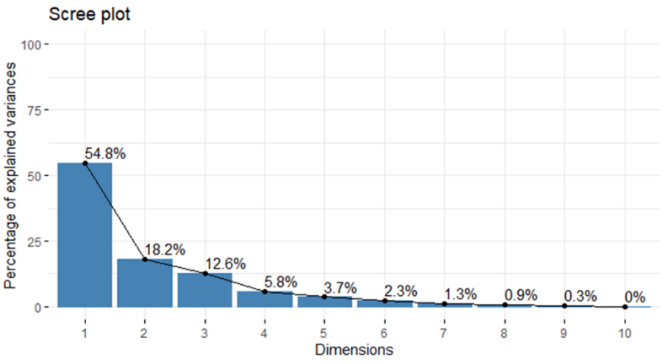

Fig. 10. Scree plot

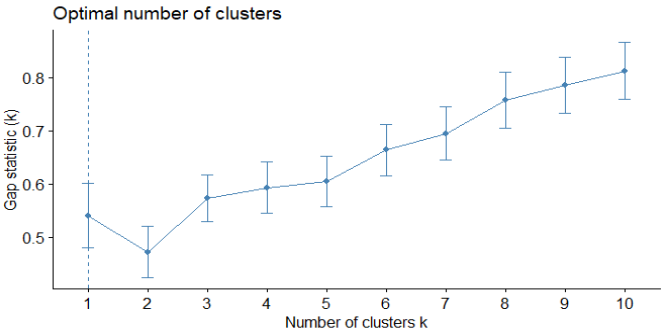

Fig. 11. Optimal number of clusters (Gap statistic)

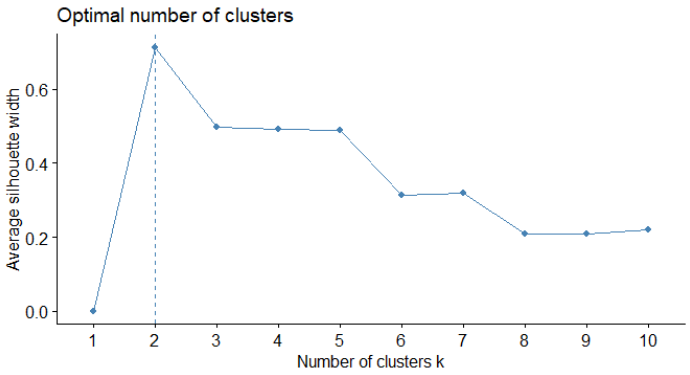

Fig. 12. Optimal number of clusters (Average silhouette)

CONCLUSION

The principal components were extracted from the dataset. It is observed that the groundwater quality variables EC, TDS, Cl, Na, SO_4, SAR, NO_3, Mg, TH and TA contribute much to the variance. Agglomeration approaches like complete, average, single, and ward were tested with the dataset. It is found that the ward method is suitable. Hence, the ward method is used to know the clustering structure.A minimal entanglement score of 0.26 reflects good alignment due to limited entanglement. This work is done using open-source software such as R and R studio, with free packages available over the internet.This work fills the gap of cluster analysis of the groundwater datasets with no extra costs incurred for purchasing software.

ACKNOWLEDGMENT

The author would like to thank the Ministry of Water Resources, River Development, and Ganga Rejuvenation. Government of India.

REFERENCES

A. Elubid, B., Huang, T., H. Ahmed, E., Zhao, J., M. Elhag, K., Abbass, W., & M. Babiker, M. (2019). Geospatial Distributions of Groundwater Quality in Gedaref State Using Geographic Information System (GIS) and Drinking Water Quality Index (DWQI). *International Journal of Environmental Research and Public Health*, *16*(5), 731. https://doi.org/10.3390/ijerph16050731

Aly, A. A., Al-Omran, A. M., & Alharby, M. M. (2015). The water quality index and hydrochemical characterization of groundwater resources in Hafar Albatin, Saudi Arabia. *Arabian Journal of Geosciences*, *8*(6), 4177–4190. https://doi.org/10.1007/s12517-014-1463-2

Anand, M., Gibson, S. A., Subbarao, K. V., Kelley, S. P., & Dickin, A. P. (2003). Early Proterozoic Melt Generation Processes beneath the Intra-cratonic Cuddapah Basin, Southern India. *Journal of Petrology*, *44*(12),2139–2171. https://doi.org/10.1093/petrology/egg073

Balram, R., Ramanaiah, S., & Harinath, V. (2013). *Geo Environmental Studies of Ganganeru River Basin, Kadapa District, Andhra Pradesh, India*. *2*(5), 7.

Bouteraa, O., Mebarki, A., Bouaicha, F., Nouaceur, Z., & Laignel, B. (2019). Groundwater quality assessment using multivariate analysis, geostatistical modeling, and water quality index (WQI): A case of study in the Boumerzoug-El Khroub valley of Northeast Algeria. *Acta Geochimica*, *38*(6), 796–814. https://doi.org/10.1007/s11631-019-00329-x

Daughney, C. J., Raiber, M., Moreau-Fournier, M., Morgenstern, U., & van der Raaij, R. (2012). Use of hierarchical cluster analysis to assess the representativeness of a baseline groundwater quality monitoring network: Comparison of New Zealand's national and regional groundwater monitoring programs. *Hydrogeology Journal*, *20*, 185–200. https://doi.org/10.1007/s10040-011-0786-2

Goswami, S., Tiwari, R. P., Maurya, V. K., Natarajan, V., Saravanan, B., Bhatt, A. K., & Verma, M. B. (2021). Exploration for concealed fracture controlled uranium mineralization: A case from Shivaramapuram-Nutankalva tract in basement granitoids, south of Cuddapah basin, Andhra Pradesh, India. *Journal of Geochemical Exploration*, *226*, 106710. https://doi.org/10.1016/j.gexplo.2020.106710

Gowd, S. S., Krupavathi, C., & Reddy, Y. S. (2019). Hydrogeochemical Studies of Chennur Mandal, Kadapa District, Andhra Pradesh, India using Geospatial Techniques. *International Journal of Research*, *6*(3), 565–575. https://journals.pen2print.org/index.php/ijr/article/view/17277

Kale, V. S., Saha, D., Patrabnis-Deb, S., Sai, V. V. S., Tripathy, V., & Patil-Pillai, S. (2020). *Cuddapah Basin, India: A Collage of Proterozoic Subbasins and Terranes*. https://doi.org/10.16943/ptinsa/2020/49820

.M, R., Kottala, R., & Badapalli, P. (2018). *Recognition and Mapping of Structural Guides for Barytes Mineral Exploration in Parts of Kadapa District using Remote Sensing and GIS*. 30–36.

Nagaraju, A., Sreedhar, Y., Thejaswi, A., & Dash, P. (2016). Integrated Approach Using Remote Sensing and GIS for Assessment of Groundwater Quality and Hydrogeomorphology in Certain Parts of Tummalapalle Area, Cuddapah District, Andhra Pradesh, South India. *Advances in Remote Sensing*, *5*(2), 83–92. https://doi.org/10.4236/ars.2016.52007

Prasad, M., Reddy, B. M., Sunitha, V., Reddy, M. R., & Reddy, Y. S. (2019). Inventory data on the sinkhole occurrences from Proterozoic Cuddapah Basin, India. *Data in Brief*, *25*, 104054. https://doi.org/10.1016/j.dib.2019.104054

R: The R Project for Statistical Computing. (n.d.). Retrieved August 5, 2021, from https://www.r-project.org/

Radhakrishna, T., R, K., & G, B. (2007). Mafic dyke magmatism around the Cuddapah Basin: Age constraints, petrological characteristics and geochemical inference for a possible magma chamber on the southwestern margin of the basin. *Journal of the Geological Society of India*, *70*.

Rahbar, A., Vadiati, M., Talkhabi, M., Nadiri, A. A., Nakhaei, M., & Rahimian, M. (2020). A hydrogeochemical analysis of groundwater using hierarchical clustering analysis and fuzzy C-mean clustering methods in Arak plain, Iran. *Environmental Earth Sciences*, *79*(13), 342. https://doi.org/10.1007/s12665-020-09064-6

Ramakrishna Reddy, M., Janardhana Raju, N., Venkatarami Reddy, Y., & Reddy, T. V. K. (2000). Water resources development and management in the Cuddapah district, India. *Environmental Geology*, *39*(3), 342–352. https://doi.org/10.1007/s002540050013

Singh, Th. D., Manikyamba, C., Tang, L., Ganguly, S., Santosh, M., Subramanyam, K. S. V., & Khelen, A. C. (2019). Phanerozoic magmatism in the Proterozoic Cuddapah Basin and its connection with the Pangean supercontinent. *Geoscience Frontiers*, *10*(6), 2239–2249. https://doi.org/10.1016/j.gsf.2019.04.001

Sreedhar, Y., & Nagaraju, A. (2017). Groundwater quality around Tummalapalle area, Cuddapah District, Andhra Pradesh, India. *Applied Water Science*, *7*(7), 4077–4089. https://doi.org/10.1007/s13201-017-0564-y

Vaidyanadhan, R. (1962). Effect of Uplift and Structure on Drainage in the Southern Part of Cuddapah Basin. *Journal of Geological Society of India (Online Archive from Vol 1 to Vol 78)*, *3*(0), 70–85. http://ischolar.info/index.php/JGSI/article/view/68209

Vittala, S. S., Govindaiah, S., & Gowda, H. H. (2005). Evaluation of groundwater potential zones in the sub-watersheds of north pennar river basin around Pavagada, Karnataka, India using remote sensing and GIS techniques. *Journal of the Indian Society of Remote Sensing*, *33*(4), 483. https://doi.org/10.1007/BF02990733

Yang, Q., Zhang, J., Wang, Y., Fang, Y., & Martin, J. D. (2015). Multivariate Statistical Analysis of Hydrochemical Data for Shallow Ground Water Quality Factor Identification in a Coastal Aquifer. *Polish Journal of Environmental Studies*, *24*(2), 769–776. https://doi.org/10.15244/pjoes/30263

***