International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 9, No. 5, October, 2022, pp. 1369-1377
Journal homepage: http://iieta.org/journals/mmep

Identification of Potential Quality of Groundwater Using Improved Fuzzy C Means Clustering Method

Naga Durga Satya Siva Kiran Relangi[1*], Aparna Chaparala[2], Radhika Sajja[3]

[1] Department of CSE, ANU College of Sciences, Acharya Nagarjuna University, Guntur 522510, A. P., India
[2] Department of CSE, RVR & JC College of Engineering (Autonomous), Guntur 522019, A. P., India
[3] Department of Mechanical Engineering, RVR & JC College of Engineering (Autonomous), Guntur 522019, A. P., India

Corresponding Author Email: kiran.cse@anits.edu.in

https://doi.org/10.18280/mmep.090527

Received: 23 June 2022
Accepted: 12 October 2022

Keywords:
artificial neural networks clustering analysis, Davies Bouldin index, silhouette score, weighted arithmetic water quality index

ABSTRACT

The groundwater quality assessment gained more attention among the water quality management stations and researchers. The conventional water quality index method and artificial neural network models are used to assess groundwater. But these models are inadequate to handle data with uncertainty. In this work, we propose an improved Fuzzy C Means clustering method to identify the homogeneous clusters with respect to groundwater quality. For this purpose 1020 groundwater samples data with 7 physiochemical parameters of the year 2019 are collected from West Godavari, Andhra Pradesh, India. The effectiveness of the proposed clustering method is evaluated with two standard clustering methods namely K-means and Fuzzy C Means. The initial selection of the number of clusters and cluster centers determines the success of both the conventional K Means and Fuzzy C Means clustering methods. The proposed improved Fuzzy C Means method identifies the optimal number of clusters based on the water index value. The proposed improved Fuzzy C Means clustering method is implemented on the groundwater data set. The performance is computed with the help of the silhouette score and Davies Bouldin Index. The proposed clustering method outperforms with the existing K Means and Fuzzy C Means with silhouette score of 0.857 and Davies Bouldin Index value of 0.502 when the number of clusters are 4.

1. INTRODUCTION

Groundwater is one among the natural resource and is generally used by many people for drinking purpose. Apart from the drinking it can be used in manufacturing industries, irrigation. Across the globe majority of the people depends on the groundwater for drinking, agricultural and domestic purposes. If the groundwater is primarily used for drinking purpose then quality needs to be given a high priority due to contamination by various factors. In many sampling locations the groundwater is polluted by anthropogenic activities [1-3]. Groundwater quality assessment, clustering and predicting the quality of the groundwater is required to effectively deal with groundwater pollution. Therefore this study is aimed to identify the homogeneous clusters for potable purpose.

The evaluation of groundwater quality is basically viewed as a clustering and classification problem also. Because present water quality assessment criteria aren't standardized, there's a lot of interest in unsupervised approaches. Clustering of groundwater is the process of identifying the homogeneous clusters with respect to groundwater quality. Identifying the homogeneous clusters is based on the closeness of groundwater data. There are various types of clustering methods that can be grouped by partition, hierarchical and clustering of large data sets. The K Means, hierarchical and Fuzzy C Means method are the very popular methods to identify homogeneous clusters in the data set. The water quality management stations employ these clustering models frequently to identify the homogeneous clusters. The K Means clustering method is also known as hard clustering and FCM is called as soft clustering technique. The hard clustering model assign each data sample to only one cluster but by soft clustering method each data sample is belongs to multiple clusters basing on the membership values between the data sample and cluster centroid.

Till date, many methods have been used to the assess groundwater quality, including conventional water quality assessment methods, Artificial Neural Networks (ANN) and Fuzzy logic. ANN are self-organizing, self-learning, non-linear processes that can handle the system which is difficult to be handled by the conventional water quality assessment methods. ANN is well suited for classifying and predicting groundwater quality due to its beneficial treatment properties such as non-linearity, parallelism, addiction tolerance, learning and generalization capabilities, distributed associations, fault tolerance capabilities, and applicability to complex problems. In the past, researchers applied ANN and also performed comparative study of ANN models in groundwater quality assessment [4-6]. The problem with the ANN is that it can handle the crisp data but groundwater data is associated with uncertainty and the fuzzy logic theory is the best way to present the groundwater data. The models built on top of the fuzzy theory explores the fuzzy nature exist in the data Hosseini-Moghari et al. [7]. Therefore in this research work we used soft clustering method to identity the homogenous regions from the groundwater data. Further, we improved the performance of the Fuzzy C Means clustering method by incorporating the global best water quality index value while computing the objective function. The improved Fuzzy C Means clustering method identifies the homogeneous clusters of the groundwater data more accurately and supports the results obtained are comparable to those reported previously by Mohammadrezapour et al. [8].

The rest of the paper is organized as follows. The literature survey is presented in Section 2. The background details about the clustering metrics Silhouette Score, Davies Bouldin Index and the conventional water quality index method are presented in Section 3. The materials and methodology we implemented in this work are presented in Section 4. Section 5 presents the results produced by the 3 clustering methods. Lastly, in section 6 the conclusions and future extension of this research work are discussed.

2. LITERATURE SURVEY

There are various groundwater quality assessment methods that are widely used by groundwater monitoring stations and researchers Bui et al. [9] improved the prediction rate of water quality indices through novel hybrid machine learning algorithms. The conventional water quality index method and artificial neural networks are the most predominantly used groundwater quality assessment approaches. Over the years various methods are developed to assess the water quality including both surface and groundwater presented in [10, 11]. But the problem with the conventional water quality assessment methods is there are no thumb rule to decide the total number of groundwater quality parameters and their range values as discussed in Lumb et al. [12]. Zhang et al. [13] proposed an improved WQI method based on Criteria Importance Through Inter-criteria Correlation weighting method. According to this improved method 94.12% of the groundwater is suitable for consumption i.e. for potable purpose. Mainly the nitrite, nitrate, and fluoride are the parameters those dilute groundwater. The studies presented below discusses the applicability of the conventional and other methods for groundwater management. Aouiti et al. [14] applied various water quality assessment methods to assess the quality of groundwater for diverse uses. According to their research work the improved water quality index method is efficient than the other groundwater quality assessments methods.

According to El-Zeiny and Elbeih [15] there exist 4 commonly used conventional water quality assessment methods viz. National Sanitation Foundation Water Quality Index, Canadian Council of Ministries of the Environment Water Quality Index, Oregon Water Quality Index and Weighted Arithmetic Water Quality Index. Poonam et al. [16] each technique has its own merits and demerits. Among the 4 methods Weighted Arithmetic Water Quality Index method is often used to evaluate the groundwater quality. Balamurugan et al. [17] applied WQI to study the potentiality of groundwater for potable and irrigation purposes in two seasons post monsoon and pre monsoon seasons of Sarabanga River region, Tamil Nadu, India. Their research reports that, in both seasons, the WQI value for groundwater revealed that 74.5 sq km and 37.24 sq km of the area were not suitable for potable purpose. The irrigation index value states that groundwater suitable for irrigation purpose.

Udeshani et al. [18] determined the groundwater quality by using the WQI method. Their research objective is to identify the relationship between incidence of Chronic Kidney Disease and drinking water quality. The application of WQI method reports that about 50% of the groundwater belongs to poor water category. This study found that groundwater quality has a direct impact on the underlying cause of Chronic Kidney Disease in the area. Asadi et al. [19] applied the WQI and irrigation water quality index methods to assess the groundwater for drinking and irrigation uses. The WQI and IWQI indexes' trends show that groundwater quality has been declining over time. In the past, the WQI method is used by several researchers to estimate the excellence of the groundwater [20-24]. Wu et al. [25] done a comprehensive literature survey carried out on various WQI methods to find the quality of groundwater. In their research they also used the fuzzy logic theory for assessing the groundwater. Their research concluded that fuzzy logic method is well suited to assign the weights to the groundwater quality parameters. The studies mentioned below discusses the conjunction of the water quality index method and unsupervised clustering methods in groundwater quality management. Wunsch et al. [26] used the Self Organizing Feature Map combined with DSL2 algorithm to automatically drive the optimal clusters in the clustering process. Developed a framework to identify the homogenous regions in the groundwater data. Hierarchical clustering is also another popular clustering model which identifies the homogeneous clusters based on the similarity and presents the clusters as dendograms.

Mohammadrezapour et al. [8] applied genetic algorithm to find the optimal number of clusters in K Means and Fuzzy C Means methods. The optimal number of clusters in the both algorithms are quite different. The number of clusters in K Means is 5 and in FCM is 6. Further, the homogeneity of the clusters formed by the 2 clustering methods are investigated with Levene homogeneity test the results reveals that the Fuzzy C Means model performs well in terms of the mean squared error, the mean squared error of the Fuzzy C Means model is 0.0000392 and for the K Means model it is 0.0000412. The mean squared error of the Fuzzy C Means model is low when compared with the K Means model which in turn depicts that Fuzzy C Means clustering method identifies the homogeneous clusters better than the K Means method.

Obeidat and Awawdeh [27] performed a comprehensive study of the groundwater assessment through conventional groundwater quality assessment and multivariate statistical methods. This study reports that 46% of the samples belongs to excellent water category, 50% of the groundwater samples belongs to good water category and only 4% of the samples belongs to poor category. Therefore from this study it is noticed that 96% of the groundwater is suitable for aquatic purpose.

Suleiman et al. [28] assessed the groundwater quality by applying the Karl Pearson's correlation coefficient to explore the impact of physical and chemical parameters which leads to groundwater pollution. Thereafter, the researchers applied hierarchical clustering method to group the groundwater samples depending on the contamination concentration and to recognize the factors which leads to groundwater contamination. The Karl Pearson's coefficient reveals the groundwater parameters TDS, EC, Ca, Cl, SO4, Na, Mg and TH influenced water contamination in several locations from the study area, the hierarchical cluster analysis forms three significant groups, namely cluster 1, cluster 2, and cluster 3, which were categorised as lower contaminated regions, moderately contaminated areas, and higher contaminated areas, respectively. Hence, groundwater from cluster 3 is unsuitable for potable purpose. The major sources of groundwater pollution are anthropogenic and industrial activates.

BS and Raman [29] has proposed a modified WQI method to assess the groundwater excellence. The modified water quality index method is used to assign the relative weights to all the parameters used in the water quality index determination. According to B. S and Raman the modified method is an efficient method to assess the groundwater when compared with the other WQI approaches. In the researcher Devi [30] studied the quality of groundwater in the Regions of Kadapa District in Andhra Pradesh. Applied K-means, K-Mediods and Hierarchical clustering methods to group the regions of water samples based on the water quality. Further, applied the outlier analysis to identify the other nontrivial patterns. Their research reports that groundwater belongs to excellent, good and poor types.

The quality and suitability of groundwater for drinking in the vicinity of a shallow, unconsolidated Quaternary aquifer were investigated. The researchers used the combination of conventional groundwater quality assessment methods coupled with fuzzy logic to determine the groundwater quality. The results of their research reflected that the groundwater in the study area is fresh to neutral and is classed as very hard due to Total Hardness. Finally, it is concluded that the groundwater is in good condition and can be used for drinking purposes [31].

Oorkavalan et al. [32] the researchers studied the suitability of the groundwater for potable, agricultural and domestic uses according to the WHO guidelines. As there are the conventional methods cannot predict the groundwater quality. Hence, the researchers applied clustering method to assess the quality of the groundwater. This study reports that the potential of the multivariate statistical methods to assess the groundwater type.

Limited research has been done on the applicability of Fuzzy C Means clustering method for studying underground water quality. Initially Fuzzy C Means was put forth by Dunn in 1974 and extended by Bezdek et al. in 1984. For example, to find out the soil content the pollutants in the marine sediments, Chang and Chang have applied K Means and FCM. Their research is Fuzzy clustering gives acceptable results for the following reasons: Unsafe boundaries between clusters and overlaps Between classes.

Güler et al. [33] studied the distribution of groundwater chemistry. Applied Fuzzy C Means clustering method and GIS which in turn identifies 4 homogeneous groups. Further, applied PCA to analyze the impact of natural and anthropogenic processes in creating the homogeneous groups. The results of this study are promising and suggest that the combination of these methods is suitable for groundwater management. Goyal and Gupta [34] uncovered the homogenous rainfall regime in India's northeastern region. The results revealed that the FCM technique outperformed K-means in identifying homogeneous zones. Caniani et al. [35] the authors studied the hierarchical classification of groundwater pollution through fuzzy logic. Their study revealed that fuzzy logic is an objective and useful tool for environmental planning.

From the literature it is observed that clustering method often used in groundwater quality management but most of the researchers applied the existing clustering models. In this work we propose an improved Fuzzy C Means soft clustering model basing on the intra class and inter cluster distance measures to identify the homogeneous clusters. The performance of the proposed clustering model is evaluated using average Silhouette score and Davies Bouldin Index.

3. BACKGROUND

Clustering is the process of identifying the homogenous regions based on the similarity measures. Most commonly used similarity measures are Euclidean distance, Manhattan distance, Minkowski distance and Jaccard distance. Euclidean distance always selects the shortest path to take as a linear line since similarity is determined by the smallest distance between two data points. Other criteria, however, might not necessarily produce the shortest distance. Experimental data is continuous data, and the optimal representation for continuous data is Euclidean. Based on the nature of the experimental data, the Euclidean measure was selected as a similarity metric. Hence, we used the Euclidean distance to find the similarity among the groundwater samples to identify the homogenous regions. The Euclidean distance is mathematically expressed as follows.

$$d(i,j) = \sqrt{(p_{i1} - p_{j1})^2 + \cdots + (p_{in} - p_{jn})^2} \tag{1}$$

Identification of homogeneous clusters can be done by using different clustering models viz. K-means, Self Organizing Feature Map, Hierarchical, DBSCAN, Fuzzy C Means etc. After forming the homogeneous clusters the validation is made by using the clustering metrics like viz., Silhouette, Clinks Harabasz, Davies Bouldin Index, Homogeneity, Fowlkes Mallows and V measure etc.

The average distance between one data point and others within the cluster and the average distance among different clusters is termed as the Silhouette Score [36]. The silhouette score is named as S and S is computed as given below.

$$S= \frac{(b-a)}{maximum\ (a,b)} \tag{2}$$

where, *a* and *b* represents the mean intra-cluster distance and the mean inter-cluster distance respectively. The score of a is computed as the average distance between a point in a class and all other points in the same class. The score *b* is computed as the average separation between a sample and every other point in the next cluster. The Silhouette score can be anywhere from -1 to 1. The computed value of S closer to 1 indicates that a sample is better clustered and if it is closer to − 1 the sample should be categorized into another cluster.

Similarly Davies Bouldin Index (DBI) is another measure to verify clustering validity by clustering method. With the maximum inter cluster distance, the characteristic similarities between each cluster tends to be small and as a result the difference among the clusters are more pronounced. With the minimum intra cluster distance, higher degree characteristic similarity can be found for each data member Surono and Putri [37]. An optimal clustering scheme should have minimal DBI (close to 0). DBI can be calculated by using the following steps.

Step1: Sum of Square Within-cluster (SSW). To determine the cohesion in the ith cluster is to calculate the value of the Sum of Square Within-cluster (SSW).

$$SSWi = \frac{1}{m_i}\sum_{j=1}^{m} d(x_j, c_i) \quad (3)$$

Step 2: Sum of Square Between-cluster (SSB). By finding out the value of the sum of squares within the cluster, one can determine the amount of cohesion in the i^th^ cluster.

$$SSBi; j = d(c_i, c_j) \quad (4)$$

Step 3: Each cluster owns a ratio value which can be measured by executing the following equation.

$$R_{i,j} = \frac{SSWi + SSBi,j}{SSB\,i,j} \quad (5)$$

Step 4: Davies Bouldin Index can be computed by using the below equation and can be given as follows.

$$DBI = \frac{1}{K}\max_{i \neq j} R_{i,j} \quad (6)$$

3.1 Water quality index

Kiran et al. [38] used the WAWQI to determine the groundwater quality. To measure groundwater quality, we must first assign a weight to each of the groundwater quality parameters based on their relative importance for drinking purposes. Eq. (7) is used to establish the relative weight of the seven parameters and the quality rating scale of each parameter is determined by using Eq. (8). The groundwater quality subindices of each parameter are calculated by multiplying the relative weight of each parameter by the quality rating scale of each parameter as shown in Eq. (9). Finally, we may use Eq. (10) to calculate the WQI of each sample.

$$rw_i = \frac{W_i}{\sum_{i=1}^{7} W_i} \quad (7)$$

$$Q_j = {}^{PC_j}\!/_{S_j} * 100 \quad (8)$$

$$SI_j = rw_i * Q_j \quad (9)$$

$$WQI = \sum_{i=1}^{7} SI_i \quad (10)$$

In the above equations W_i, rw_i represents the weight and relative weight of the 7 groundwater parameters (as reported in earlier section, Deepa and Venkateswaran [3] and Vidyalakshmi et al. [10]), Qj denotes the quality rating scale of the 7 parameters, and Sj is the national drinking water quality standard, according to BIS 2012 [39]. Each parameter's concentration is represented by PC_j. SIj and WQI stands for sub index and water quality index, respectively.

4. MATERIALS & METHODS

We gathered 1020 groundwater samples data set from RWS & S laboratory, Narsapuram, West Godavari District, Andhra Pradesh, India, for this study. All of the samples were subjected to 7 physical, chemical, and microbiological criteria. In (Table 1) the statistical information about the parameters is presented. We had applied 3 unsupervised learning models to identity the homogeneous groups based on the similarity among the groundwater parameters data. The 3 models namely K-means, Fuzzy C Means and improved Fuzzy C Means takes the groundwater parameters data set as input and forms the homogenous groups. The 3 clustering models takes the number of clusters as an additional input to form the clusters. The trial and error technique employed to find the optimal number of clusters. Thereafter, we applied WAWQI method to check the suitability of each cluster for potable purpose. We used Python version 3.7.2 to implement the above stated clustering methods. The performance of the clustering model for ground water samples concerning the quality is being evaluated against variable number of clusters. In this work the performance of the clustering methods is being evaluated through the metrics like Silhouette Coefficient, Davies-Bouldin Index approaches. The data samples are being cluster by deploying the K-Means, Fuzzy C Means and improved Fuzzy C Means clustering methods over multiple iterations. The clustering method relay on the feature set associated with the dataset on ground water quality which includes pH, Temp, Conductivity, BOD, Nitrate+Nitrite, Fecal Coliform and Total Coliform for the clustering process.

Clustering is the process of forming homogeneous groups based on the similarity among the members in the data group, means that members in the group are similar to other members in the same group and dissimilar to the members in the other group. Clustering methods are broadly divided into 2 groups they are hard and soft clustering methods. By the hard clustering methods one member is assigned to one cluster only but with soft clustering methods, the member or sample should be assigned to multiple clusters.

Table 1. Statistics of the groundwater quality parameters

Name of the parameter	Desirable limits recommended by BIS/ICMR	Statistical values computed for the groundwater parameters							
		Min	Max	Mean	Std	25%	50%	75%	Total samples
Ph	6.5-8.5	2.6	9.1	7.43	0.85	7.3	7.5	7.8	1020
Cond (µS)	300	37	36593	758.61	1951.7	294	531	970	1020
BOD (mg/l)	<5.0	0.3	27.8	9.59	7.74	2.2	8.8	13.5	1020
Nitrate+Nitrate (mg/l)	0-45	0	14.85	3.2	2.6	1.5	2.69	4.77	1020
FC (100 mL)	1.00-1.5	2	36250	5876.2	10439	24	130	5000	1020
TC(100 mL)	1.00-10.00	5	75000	13511.1	21972.1	82	343	26750	1020
Temp	-	17	31.1	24.8	3.12	22	24.5	25.8	1020

K-means algorithm: The K Means clustering method is frequently used to cluster the groundwater data to identify the homogeneous clusters. MacQueen is developed the K-means algorithm in 1967. The first step is to define the number of cluster and the cluster centroid randomly. Initially we used 2 as the number of clusters. In the next step the data belonging to each cluster is identified based on the similarity measure between the cluster center and each data groundwater data sample. The Euclidian distance is the similarity measure used to find the distance between the cluster center and the other groundwater data samples. After assigning the data samples to cluster the cluster centers are updated this same procedure repeated until there is no change in the cluster centers as compared with the previous cluster centers.

Fuzzy C Means clustering algorithm: The Fuzzy C Means clustering is a good choice when the data possesses vagueness or uncertainty. The Fuzzy C Means clustering algorithm is put forth by Dunn and Bezdek in 1981. The soft clustering algorithm is advantageous than the other hard clustering algorithms like K-means, Self Organizing Map, Hierarchical Clustering etc., because it considers the membership degree of each data point. The Fuzzy C Means algorithm is presented below. Fuzzy C Means clustering is used in the groundwater studies to identify the homogeneous clusters. As its name implies it works based on the fuzzy logic. Unlike hard clustering Fuzzy C Means clustering assign each groundwater data sample to several clusters at the same time that the sum of the membership degree of each groundwater sample must be equal to *1* and mathematically given below.

$$\sum_{i=1}^{c} u_{ir} = 1 \; \forall r = 1, \dots\dots, n \qquad (11)$$

In the above equation c represents the number of clusters and u_{ir} is the total membership in the i^th^ cluster. The parameter n denotes the unlabelled groundwater data samples. The number of clusters c must be initialized first by the user. The number of cluster c is the key parameter which influence the clustering partitions in this work we use average silhouette coefficient value to fix the value of *c*. The fuzziness parameter q is chosen in the range of (1.5-2.5) by Pal and Bezdek in 1995. The modified objective function is presented in Eq. (14).

1. First each sample is assigned to every cluster with a random membership value.
2. Calculate the new centre of each cluster by using the following equation

$$v_{ij} = \frac{\sum_{r=1}^{n} u_{ir}^{q}\, x_{rj}}{\sum_{r=1}^{n} u_{ir}^{q}} \qquad (12)$$

In the above equation u_{ir} is the membership value from the r^th^ sample to i^th^ cluster, and x_{rj} is the value of the r^th^ variable in k^th^ sample. The varible *q* denotes the fuzziness cofficient.

3. Once new cluster centres are obtained, next calculate the degree of membership of each unlabelled sample to the newly formed cluster centre in each cluster and is given in the following equation. The Euclidean distance is used to find the membership of the unlabelled sample to the cluster centre.

$$u_{ij} = \frac{(d_{ir}^{2})^{-1/(q-1)}}{\sum_{k=1}^{c} (d_{ir}^{2})^{-1/(q-1)}} \qquad (13)$$

In the above equation d$_{ir}$ is the distance from the r^th^ data sample to centre of i^th^ cluster.

4. The proposed objective function is for variable *j* is given below with respect to fuzzy coefficient *q*, in order to establish denser clusters we compute the water quality index of each sample there after find the distance between each data point and every other member in the cluster based on their water quality index values.

$$J = \sum_{i=1}^{c} \sum_{r=1}^{n} u_{ir}^{q} d_{ir}^{2} = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ir}^{q} ||x_r - v_i||^2 + u_{ir}^{q} ||\alpha * x_r wqi - wqi_{ri}||^2 \qquad (14)$$

where, $G_{best} = \sum_{r=1}^{n} max\,(x_r wqi)$ and $\alpha = ||G_{best} - x_r wqi||$.

In the above Eq. (14) $x_r wqi$ is the water quality index value of the r^th^ data sample and wqi_{ri} is the water quality index value of every data sample within the cluster.

5. Repeat the computations until the gap between target functions is below the specified critical value in two phases, between 10^{-5} and 10^{-3}.

5. RESULTS AND ANALYSIS

The identification of homogeneous clusters by using K-means, Fuzzy C Means and proposed method was started by setting the number of clusters to 2 and then increasing the number of cluster center to 7. The optimum number of clusters was then attempted to be found by increasing the number of cluster center to 7. Further, to find the optimal number of clusters we used the average Silhouette score. The optimal number of clusters in the groundwater data is identified by using the clustering scores as shown in Table 2. The Silhouette score values produced by the improved Fuzzy C Means model are presented in Figure 3. It was observed that the average silhouette score 0.857 is for number of cluster are 4. The silhouette scores of the K-means and Fuzzy C Means clustering model with different number of clusters are shown in Figure 1 and Figure 2. From Figure 1, Figure 2 and Figure 3 the silhouette scores of all the clustering models used in this work are greater than 8 which clearly indicates that identification of homogeneous clusters in the groundwater is considerable.

The clustering quality is being assessed for variable number of cluster ranging from 2 to 7 over multiple iterations. The Figure 1 presents the subplots for K-means clustering. The optimal number of clusters are determined based on the Silhouettes score associated the number of clusters, it is desired to better Silhouettes Score. The maximum silhouette score produced by the K Means model is 0.837 for 4 clusters. The DBI is computed using equations 3-5. The clustering model performs well if the DBI value is relatively close to 0. For 4 cluster the computed DBI value 0.508.

The performance of the Fuzzy C Means model is being evaluated through Silhouette Coefficient, Davies-Bouldin Index approaches, and it is observed that the model has exhibited a better performance for 4 clusters concerning to Silhouette Coefficient value is 0.837. For 4 cluster the computed DBI value 0.504. The evaluation of Fuzzy C Means for variable number of clusters are being presented in Table 2 and the corresponding graph is presented in Figure 2. To evaluate the validity and efficiency of the suggested methodology, the results of the proposed method were also compared with the previously investigated groundwater quality assessment approaches using clustering methods such as: Mohammadrezapour et al. [8], Obeidat [27], Devi [30], Güler et al. [33] and Lee et al. [40].

Table 2. Statistics calculated for the clusters determined by FCM and K-means clustering

No. of clusters	Name of the clustering model	Scores of different clustering metrics	
		Average silhouette coefficient	Davies-Bouldin index
2	Improved FCM	0.868	0.320
	FCM	0.849	0.322
	K-means	0.848	0.327
3	Improved FCM	0.832	0.638
	FCM	0.812	0.639
	K-means	0.812	0.638
4	Improved FCM	0.857	0.502
	FCM	0.837	0.502
	K-means	0.837	0.502
5	Improved FCM	0.805	0.473
	FCM	0.802	0.485
	K-means	0.823	0.567
6	Improved FCM	0.826	0.521
	FCM	0.808	0.526
	K-means	0.806	0.529
7	Improved FCM	0.834	0.500
	FCM	0.815	0.501
	K-means	0.814	0.505

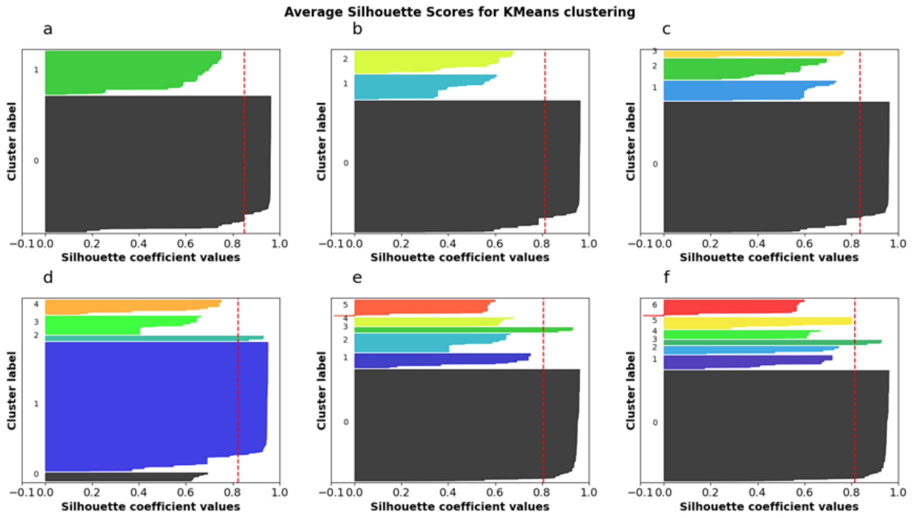

Figure 1. The average silhouette scores computed for different number of clusters for K Means

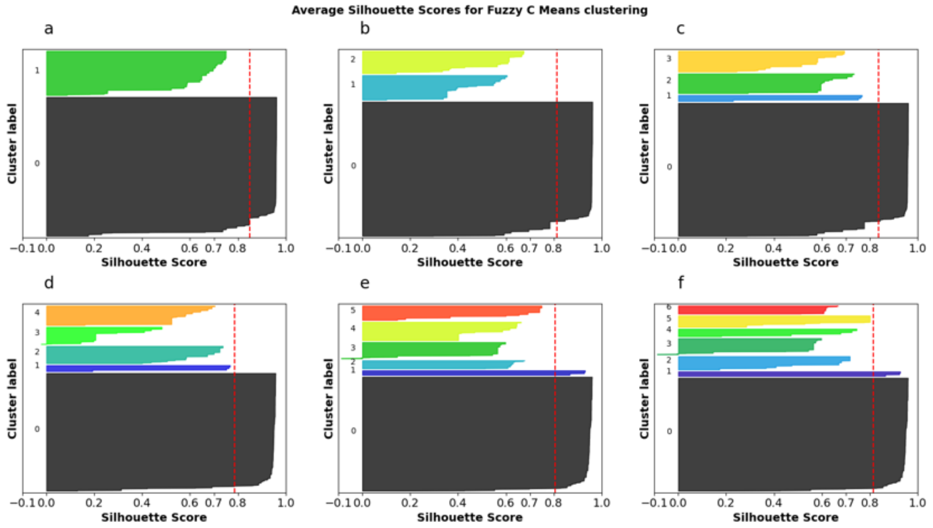

Figure 2. The average Silhouette scores computed for different number of clusters for Fuzzy C Means

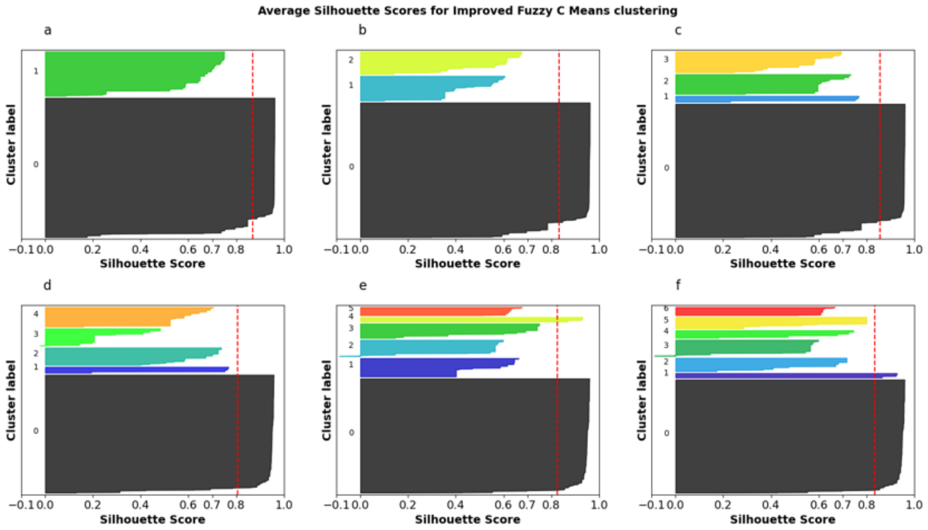

Figure 3. The average Silhouette scores computed for different number of clusters for improved Fuzzy C Means

The statistics obtained for the improved Fuzzy C Means clustering models is also presented in the Table 2. The average silhouette scores for the proposed clustering are plotted in Figure 3. From Figure 3 it is noticed that the maximum silhouette score for number of clusters are 2 and 4. Furthermore, the average silhouette scores is found to be maximum as 0.868 and 0.857 respectively. The average silhouette scores produced by the Fuzzy C Means and k-means clustering models are plotted in Figure 1 and Figure 2. These two clustering models also have maximum silhouette scores for number of clusters are 2 and 4 respectively. Based on the silhouette scores it is observed that the optimal number of clusters in the considered clustering models are 4. It can be seen that the maximum average silhouette score is obtained for number of clusters are 4 and it indicates that the clusters are well formed because when the silhouette score is near to 0.9 and if the silhouette score is less means between 0 and 0.5 indicates that the clustering model performance is low. For 4 clusters the computed DBI values is 0.502. Finally we conclude that the proposed fuzzy clustering model out performs the other clustering models considered for identification of potential regions of groundwater.

After forming the homogeneous clusters by the clustering methods we applied the Weighted Arithmetic Water Quality Index method to check the suitability of the groundwater for potable purpose. We analysed results of the to the K-Means clustering method based on the maximum Silhouette score (number of clusters are 4). The average Silhouette score is 0.837 (Figure 1.c). According to Weighted Arithmetic Water Quality Index method the samples in the cluster 0 belongs to good category, sample from the cluster 1 belongs to fair category, samples from cluster 2 and cluster 3 belongs poor category. Hence samples from both cluster 0 and cluster 1 are suitable for potable purpose. Further, the assessment of the groundwater for potable purpose is done by applying the conventional groundwater quality assessment method it indicates that cluster 0 and cluster 1 are suitable for potable purpose.

6. CONCLUSION

In this research work, an improved Fuzzy C Means clustering method is proposed to identify the homogeneous clusters from the groundwater data. The objective functions of the Fuzzy C Means clustering method is improved based on intra class and inter cluster distance. The results of the improved Fuzzy C Means clustering method are better when compared with the K Means and Fuzzy C Means clustering methods. The quality of the groundwater must be determined prior to its use for aquatic purposes. In this research work, we first identify the homogeneous clusters of the groundwater data after that we assess the quality of the groundwater for potable purpose. The K-Means, Fuzzy C Means and improved Fuzzy C Means clustering methods are used to identify the homogeneous clusters from the groundwater data. The results produced by the clustering models are compared and validated using Silhouette, Davies-Bouldin Index. These validation metrics tell us that the improved Fuzzy C Means clustering method identifies the homogeneous clusters more accurately than the Fuzzy C Means and K-Means clustering methods. As a future extension, build a framework based on machine learning and deep learning models to assess the groundwater quality of large data sets.

ACKNOWLEDGMENT

I have collected the groundwater samples data from Water Quality Monitoring Lab, RWS&S Sub-Division, Narsapuram, West Godavari, Andhra Pradesh, India and I express my thanks to them for their timely support and for according permission for utilizing the groundwater data samples in this research work.

REFERENCES

[1] Hasan, M.A., Ahmad, S., Mohammed, T. (2021). Groundwater contamination by hazardous wastes. Arabian Journal for Science and Engineering, 46(5): 4191-4212. https://doi.org/10.1007/s13369-021-05452-7

[2] Wagh, V.M., Mukate, S.V., Panaskar, D.B., Muley, A.A., Sahu, U.L. (2019). Study of groundwater hydrochemistry and drinking suitability through Water Quality Index (WQI) modelling in Kadava river basin, India. SN Applied Sciences, 1(10): 1-16. https://doi.org/10.1007/s42452-019-1268-8

[3] Deepa, S., Venkateswaran, S. (2018). Appraisal of groundwater quality in upper Manimuktha sub basin, Vellar river, Tamil Nadu, India by using Water Quality Index (WQI) and multivariate statistical techniques. Modeling Earth Systems and Environment, 4(3): 1165-1180. https://doi.org/10.1007/s40808-018-0468-3

[4] Sakizadeh, M. (2016). Artificial intelligence for the prediction of water quality index in groundwater systems. Modeling Earth Systems and Environment, 2(1): 1-9. https://doi.org/10.1007/s40808-015-0063-9

[5] Zaqoot, H.A., Hamada, M., Miqdad, S. (2018). A comparative study of Ann for predicting nitrate concentration in groundwater wells in the southern area of Gaza Strip. Applied Artificial Intelligence, 32(7-8): 727-744. https://doi.org/10.1080/08839514.2018.1506970

[6] Gupta, R., Singh, A.N., Singhal, A. (2019). Application of ANN for water quality index. International Journal of Machine Learning and Computing, 9(5): 688-693. https://doi.org/10.18178/ijmlc.2019.9.5.859

[7] Hosseini-Moghari, S.M., Ebrahimi, K., Azarnivand, A. (2015). Groundwater quality assessment with respect to fuzzy water quality index (FWQI): An application of expert systems in environmental monitoring. Environmental Earth Sciences, 74(10): 7229-7238. https://doi.org/10.1007/s12665-015-4703-1

[8] Mohammadrezapour, O., Kisi, O., Pourahmad, F. (2020). Fuzzy c-means and K-means clustering with genetic algorithm for identification of homogeneous regions of groundwater quality. Neural Computing and Applications, 32(8): 3763-3775. https://doi.org/10.1007/s00521-018-3768-7

[9] Bui, D.T., Khosravi, K., Tiefenbacher, J., Nguyen, H., Kazakis, N. (2020). Improving prediction of water quality indices using novel hybrid machine-learning algorithms. Science of the Total Environment, 721: 137612. https://doi.org/10.1016/j.scitotenv.2020.1376

[10] Sutadian, A.D., Muttil, N., Yilmaz, A.G., Perera, B.J.C. (2016). Development of river water quality indices—a review. Environmental Monitoring and Assessment, 188(1): 1-29. https://doi.org/10.1007/s10661-015-5050-0

[11] Vidyalakshmi, R., Brindha, B., Benedict Roosvelt, P.S., Rajakumar, S., Prashanthi Devi, M. (2013). determination of land use stress on drinking water quality in Tiruchirappalli, India using derived indices. Water Quality, Exposure and Health, 5(1): 11-29. https://doi.org/10.1007/s12403-012-0083-x

[12] Lumb, A., Sharma, T.C., Bibeault, J.F. (2011). A review of genesis and evolution of Water Quality Index (WQI) and some future directions. Water Quality, Exposure and Health, 3(1): 11-24. https://doi.org/10.1007/s12403-011-0040-0

[13] Zhang, Q., Xu, P., Qian, H. (2020). Groundwater quality assessment using improved water quality index (WQI) and human health risk (HHR) evaluation in a semi-arid region of northwest China. Exposure and Health, 12(3): 487-500. https://doi.org/10.1007/s12403-020-00345-w

[14] Aouiti, S., HamzaouiAzaza, F., El Melki, F., Hamdi, M., Celico, F., Zammouri, M. (2020). Groundwater quality assessment for different uses using various water quality indices in semi-arid region of central Tunisia. Environmental Science and Pollution Research, 28(34): 46669-46691. https://doi.org/10.1007/s11356-020-11149-5

[15] El-Zeiny, A.M., Elbeih, S.F. (2019). GIS-based evaluation of groundwater quality and suitability in Dakhla Oases, Egypt. Earth Systems and Environment, 3(3): 507-523. https://doi.org/10.1007/s41748-019-00112-1

[16] Poonam, T., Tanushree, B., Sukalyan, C. (2013). Water quality indices-important tools for water quality assessment: A review. International Journal of Advances in Chemistry, 1(1): 15-28.

[17] Balamurugan, P., Kumar, P.S., Shankar, K. (2020). Dataset on the suitability of groundwater for drinking and irrigation purposes in the Sarabanga River region, Tamil Nadu, India. Data in Brief, 29: 105255. https://doi.org/10.1016/j.dib.2020.105255

[18] Udeshani, W.A.C., Dissanayake, H.M.K.P., Gunatilake, S.K., Chandrajith, R. (2020). Assessment of groundwater quality using water quality index (WQI): A case study of a hard rock terrain in Sri Lanka. Groundwater for Sustainable Development, 11: 100421. https://doi.org/10.1016/j.gsd.2020.100421

[19] Asadi, E., Isazadeh, M., Samadianfard, S., Ramli, M.F., Mosavi, A., Nabipour, N., Shamshirband, S., Hajnal, E., Chau, K.W. (2019). Groundwater quality assessment for sustainable drinking and irrigation. Sustainability, 12(1): 177. https://doi.org/10.3390/su12010177

[20] Verma, P., Singh, P.K., Sinha, R.R., Tiwari, A.K. (2020). Assessment of groundwater quality status by using water quality index (WQI) and geographic information system (GIS) approaches: A case study of the Bokaro district, India. Applied Water Science, 10(1): 1-16. https://doi.org/10.1007/s13201-019-1088-4

[21] Adimalla, N., Qian, H. (2019). Groundwater quality evaluation using water quality index (WQI) for drinking purposes and human health risk (HHR) assessment in an agricultural region of Nanganur, south India. Ecotoxicology and Environmental Safety, 176: 153-161. https://doi.org/10.1016/j.ecoenv.2019.03.066

[22] Subba Rao, N., Srihari, C., Deepthi Spandana, B., Sravanthi, M., Kamalesh, T., Abraham Jayadeep, V. (2019). Comprehensive understanding of groundwater quality and hydrogeochemistry for the sustainable development of suburban area of Visakhapatnam, Andhra Pradesh, India. Human and Ecological Risk Assessment: An International Journal, 25(1-2): 52-80. https://doi.org/10.1080/10807039.2019.1571403

[23] Saleem, M., Hussain, A., Mahmood, G. (2016). Analysis of groundwater quality using water quality index: A case study of greater Noida (Region), Uttar Pradesh (UP), India. Cogent Engineering, 3(1): 1237927. https://doi.org/10.1080/23311916.2016.1237927

[24] Qureshimatva, U.M., Maurya, R.R., Gamit, S.B., Patel, R.D., Solanki, H.A. (2015). Determination of physico-chemical parameters and water quality index (WQI) of Chandlodia Lake, Ahmedabad, Gujarat, India. Journal Environ Anal Toxicol, 5(288): 2161-0525. https://doi.org/10.4172/2161-0525.1000288

[25] Wu, J., Li, P., Wang, D., Ren, X., Wei, M. (2019). Statistical and multivariate statistical techniques to trace the sources and affecting factors of groundwater pollution in a rapidly growing city on the Chinese Loess Plateau. Human and Ecological Risk Assessment: An

International Journal, 1-19. https://doi.org/10.1080/10807039.2019.1594156

[26] Wunsch, A., Liesch, T., Broda, S. (2022). Feature-based groundwater hydrograph clustering using unsupervised self-organizing map-ensembles. Water Resources Management, 36(1): 39-54. https://doi.org/10.1007/s11269-021-03006-y

[27] Obeidat, M., Awawdeh, M. (2021). Assessment of groundwater quality in the area surrounding Al-Zaatari Camp, Jordan, using cluster analysis and water quality index (WQI). Jordan Journal of Earth and Environmental Sciences, 12(3): 187-197

[28] Suleiman, A.A., Ibrahim, A., Abdullahi, U.A. (2020). Statistical explanatory assessment of groundwater quality in Gwale LGA, Kano state, northwest Nigeria. Hydrospatial Analysis, 4(1): 1-13. https://doi.org/10.21523/gcj3.2020040101

[29] BS, S., Raman, S. (2020). A novel approach for the formulation of Modified Water Quality Index and its application for groundwater quality appraisal and grading. Human and Ecological Risk Assessment: An International Journal, 26(10): 2812-2823. https://doi.org/10.1080/10807039.2019.1688638

[30] Devi, S.G. (2021). Cluster and outlier analysis for ground water quality data in the regions of Kadapa district in Andhra Pradesh. Recent Patents on Engineering, 15(2): 121-129. https://doi.org/10.2174/1872212113666190211144935

[31] Mohamed, A., Dan, L., Kai, S., Mohamed, M., Aldaw, E., Elubid, B. (2019). Hydrochemical Analysis and Fuzzy Logic Method for Evaluation of Groundwater Quality in the North Chengdu Plain, China. International Journal of Environmental Research and Public Health, 16(3): 302. https://doi.org/10.3390/ijerph16030302

[32] Oorkavalan, G., Chidambaram, S. M., Mariappan, V., Kandaswamy, G., Natarajan, S. (2016). RETRACTED: Cluster analysis to assess groundwater quality in Erode District, Tamil Nadu, India. Circuits and Systems, 7(6): 877-890. https://doi.org/ 10.4236/cs.2016.76075

[33] Güler, C., Kurt, M.A., Alpaslan, M., Akbulut, C. (2012). Assessment of the impact of anthropogenic activities on the groundwater hydrology and chemistry in Tarsus coastal plain (Mersin, SE Turkey) using fuzzy clustering, multivariate statistics and GIS techniques. Journal of Hydrology, 414: 435-451. https://doi.org/10.1016/j.jhydrol.2011.11.021

[34] Goyal, M.K., Gupta, V. (2014). Identification of homogeneous rainfall regimes in northeast region of India using fuzzy cluster analysis. Water Resources Management, 28(13): 4491-4511. https://doi.org/10.1007/s11269-014-0699-7

[35] Caniani, D., Lioi, D.S., Mancini, I.M., Masi, S. (2015). Hierarchical classification of groundwater pollution risk of contaminated sites using fuzzy logic: A case study in the Basilicata Region (Italy). Water, 7(5): 2013-2036. https://doi.org/10.3390/w7052013

[36] Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., Liu, J. (2021). A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. EURASIP Journal on Wireless Communications and Networking, 2021(1): 1-16. https://doi.org/10.1186/s13638-021-01910-w

[37] Surono, S., Putri, R.D.A. (2021). Optimization of fuzzy c-means clustering algorithm with combination of Minkowski and Chebyshev distance using principal component analysis. International Journal of Fuzzy Systems, 23(1): 139-144. https://doi.org/10.1007/s40815-020-00997-5

[38] Kiran, R.N.D.S.S., Aparna, C., Radhika, S. (2021). An enhanced weight update method for simplified ARTMAP to classify groundwater data. International Journal of Design & Nature and Ecodynamics, 16(5): 517-524. https://doi.org/10.18280/ijdne.160505

[39] Anwar, K.M., Aggarwal, V. (2014). Analysis of groundwater quality of Aligarh city, (India): Using water quality index. Current World Environment, 9(3): 851-857, https://doi.org/10.12944/CWE.9.3.36

[40] Lee, K.J., Yun, S.T., Yu, S., Kim, K.H., Lee, J.H., Lee, S.H. (2019). The combined use of self-organizing map technique and fuzzy c-means clustering to evaluate urban groundwater quality in Seoul metropolitan city, South Korea. Journal of Hydrology, 569: 685-697. https://doi.org/10.1016/j.jhydrol.2018.12.031