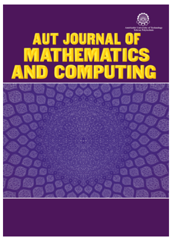

AUT Journal of Mathematics and Computing

AUT J. Math. Comput., 1(1) (2020) 101-112
DOI: 10.22060/ajmc.2019.14873.1013

Statistical and Fuzzy Clustering Methods and their Application to Clustering Provinces of Iraq based on Agricultural Products

I. A. Z. Atiyah[1], S. M. Taheri[2*]

[1] Faculty of Science, University of Al-Qadisiyah, Iraq
[2] School of Engineering Science, College of Engineering, University of Tehran

ABSTRACT: The important approaches to statistical and fuzzy clustering are reviewed and compared, and their applications to an agricultural problem based on a real-world data are investigated. The methods employed in this study includes some hierarchical clustering and non-hierarchical clustering methods and Fuzzy C-Means method. As a case study, these methods are then applied to cluster 15 provinces of Iraq based on some agricultural crops. Finally, a comparative and evaluation study of different statistical and fuzzy clustering methods is performed. The obtained results showed that, based on the Silhouette criterion and Xie-Beni index, fuzzy c-means method is the best one among all reviewed methods.

Review History:

Received: 21 August 2018
Revised: 18 December 2018
Accepted: 26 January 2019
Available Online: 1 February 2020

Keywords:

Hierarchical Clustering
Non-Hierarchical Clustering
Fuzzy C-Means Clustering

1- Introduction

Data clustering is a technique of originating groups of objects to clusters such that the objects in one cluster are quite analogous, and those in various clusters are extremely different. Several techniques of clustering have been extensively utilized in many research areas such as agriculture, sociology, medicine, geology, criminology, and engineering fields. For a review on some basic methods in this topic see, e.g. Gan et al. (2007) and Rencher (2002). During the last decades, there has been a lot of attention to the fuzzy clustering methods, too. To achive the goals of this study, we look at some studies concerning classical and fuzzy clustering particularly related to agricultural studies. Kostov and McErlean (2006) employed a specific technique to classify farms into representative farms. Chang et al. (2011) proposed a Fuzzy C-Means (FCM) clustering method using the cluster center displacement method between reiterate procedures that lessening the computational intricacy of the traditional Fuzzy C-Means clustering method. Chattopadhyay et al. (2011) suggested the Entropy-based fuzzy clustering (EFC) algorithm and compared that with a Fuzzy C-Means method. In EFC case, the cluster centers are real, which were chosen from the data points. However, in the FCM case, the cluster centers are simulated, which are chosen at haphazardly and therefore, could be out of the dataset. Also, algorithms have been compared on four datasets, namely, IRIS, WINES, OLITOS, and psychosis, based on the quality of the clusters obtained, including (discrepancy factor, compactness, distinctness) and their computational time. Volmurgan (2012) probed a two partitions-based clustering performance as K-Means and Fuzzy C-Means. The comparison is achieved through data points clustering that is haphazardly circulated. Panda et al. (2013) employed clustering methods to certain field such as medicine, business, engineering systems, and image processing. A comparison study is done by Grover (2014) in which various fuzzy clustering techniques like Fuzzy C-Means Algorithm, Possibilistic C-Means Method (PCM), Fuzzy Possibilistic C-Means Method (FPCM) and a Possibilistic Fuzzy C- Means Method (PFCM) are compaired by expressing their advantages and disadvantages. A comparative study is presented by Bora et al. (2014) between Fuzzy clustering method and a hard clustering method. Mansour et al. (2015) examined the genetic variety among accessions of pomegranate of South Eastern Tunisia through employing a clustering technique. A study conducted by Aguilar et al. (2015) investigated adopters of cluster analysis of specific drills and technologies for the planters of oil palm in Mexico. Ferraro et al. (2015) studied the R programming language to propose an innovative toolbox for the fuzzy clustering technique. The new toolkit that called fclust comprises a suite of fuzzy clustering methods, conception tools, and fuzzy cluster validity keys for fuzzy clustering outcomes. Lately, Ansari et al. (2016) investigated clusters of genetic diversity in germplasm of cluster bean. Furthermore, Fajardo et al. (2016) studied the fuzzy clustering of definite varieties for the objective recognition of soil morphological prospects of soil outlines. Gomathi and Velusamy (2018) suggested a hybridization of the Fuzzy C-Means and Fuzzy Bee Colony Optimization to solve the local optima problem and achieved a global optimum solution in Fuzzy C-means

*Corresponding author.
E-mail addresses: sm_taheri@ut.ac.ir, israa.zad@aut.ac.ir

algorithms. The experimental results show that the proposed hybridization model created higher performance compared with other clustering methods. For more about fuzzy clustering methods and applications see, e.g. Oliveira and Pedrycz (2007)
In the present research, we review and compare the main approaches to statistical clustering and fuzzy clustering, and guide the readers how to use the methods of clustering (hierarchical clustering and nonhierarchical clustering methods), and fuzzy clustering method (Fuzzy C-Means) through a specific real-world data set on some agricultural crops in the current distribution of Iraq's provinces. The results are calculated by using the software R.

2- Clustering Methods

The hierarchical clustering and partitioning are the two commonly used approaches to clustering.

2- 1- Hierarchical Clustering Methods

Hierarchical clustering starts with n clusters, each cluster contains one object and ends with a single cluster includes all of n objects. There is an alternate method, named the divisive method, which begins with one cluster including all n observations and in each step splits a cluster into two clusters. This approach finishing with n clusters, each one containing a single element (Rencher (2002)).

2- 1- 1- Agglomerative Methods

In such methods, two closest clusters are joined into a single new cluster at each step. Agglomerative methods include different methods: single linkage, complete linkage, centroid method, average linkage, median method, Ward's method, and a flexible beta method. In each method, the distance of every pair of clusters is calculated and at each step, two clusters are joined if they have the smallest distance.
Let A and B be two clusters, the distance between A and B is defined as follows:

1- Single linkage method, $D(A, B) = min\left\{d(y_i, y_j),\ for\ y_i\ in\ A\ and\ y_j\ in\ B\right\}$ where $d\left(y_i, y_j\right)$ is any distance between the vectors $y_i and\ y_j$.

2- Complete linkage method, $D(A,B)=max\left\{d(y_i,y_j),\ for\ y_i\ in\ A\ and\ y_j\ in\ B\right\}$.

3- Centroid method, $D(A,B)=d(\bar{y}_{A,}\ \bar{y}_B)$, $\bar{y}_A = \sum_{i=1}^{n_A} y_{i/n_A}$, $\bar{y}_B = \sum_{i=1}^{n_B} \bar{y}_{i\ /n_B}$

The new centroid of the cluster AB is obtained by calculating the weighted average $\bar{y}_{AB} = \frac{n_A\ \bar{y}_{A+}\ n_B\ \overline{y}_B}{n_A\ +n_B}$.

4- Average linkage method, $D(A,B)=\left(\frac{1}{n_A n_B}\right)\sum_{i=1}^{n_A}\sum_{j=1}^{n_B} d(y_i,y_j)$,

where n_A and n_B are the number of items in A and B, respectively.

5- Ward's method joins the two clusters A and B that minimize the increase in SSE , defined as:

$$I_{AB} = SSE_{AB} - \left(SSE_A + SSE_B\right), \tag{1}$$

Where

$$I_{AB} = \frac{n_A n_B}{n_A + n_B}\left(\bar{y}_A - \bar{y}_B\right)'\left(\bar{y}_A - \bar{y}_B\right), \tag{2}$$

$$SSE_A = \sum_{i=1}^{n_A}\left(y_i - \bar{y}_A\right)'\left(y_i - \bar{y}_A\right)\ ,\ SSE_B = \sum_{i=1}^{n_B}\left(y_i - \bar{y}_B\right)'\left(y_i - \bar{y}_B\right),$$

$$SSE_{AB} = \sum_{i=1}^{n_{AB}}\left(y_i - \bar{y}_{AB}\right)'\left(y_i - \bar{y}_{AB}\right)\ ,\ \bar{y}_{AB} = \frac{n_A\bar{y}_A + n_B\ \bar{y}_B}{n_A\ + n_B}.$$

Additionally, $n_{AB} = n_A + n_B$ is the number of points in AB.
Ward's method has a relationship with the centroid method. If the distance $D(A,B) = d(\bar{y}_A, \bar{y}_B)$ is squared and then compared to (2), the only observed difference is the coefficient $\frac{n_A n_B}{n_A + n_B}$ for Ward's method.

6- Flexible beta method, if the clusters are joined to form the cluster then

$$D(C,AB) = \alpha_A\ D(C,A) + \alpha_B\ D(C,B) + \beta D(A,B) + \gamma\left|D(C,A) - D(C,B)\right., \tag{3}$$

where the distances $D(C,A), D(C,B)and\ D(A,B)in(3)$ are forming the distance matrix before joining $A\ and\ B$. Lance and Williams (1967) proposed the following constraints on the parameter values: $\alpha_A + \alpha_B + \beta = 1$, $\alpha A = \alpha B$, $\gamma = 0 and$ $\beta < 1$. With $\alpha_A = \alpha_B$, we have $2\alpha_A = 1 - \beta$ $or\ \alpha_A = \alpha_B = (1-\beta)/2$, $\beta = -0.5$.

2- 1- 2- Divisive Hierarchical Methods

In this method, our work focuses on a splinter group and what remains is stated below:

1- Calculate the distance between all the items.

2- Calculate the average distances between every item and the other items.

3- Choose the element that has the leading average distance and separates it to be the splinter group.

4- Repeat the point (2) for the items in the remainder group.

5- Determine the divergence between the typical distance of the item in the remains of the other elements in the remain, and its typical distance from the elements in the splinter group.

6- If the major difference in the point (5) becomes positive, the element is relocated to the splinter group. But, if the largest difference is negative, the process ends and the split is complete, (Rencher (2002)).

2- 2- Non-hierarchical Clustering Methods

2- 2- 1- Partitioning Method (K-Means)

In this method, we follow the steps below:

1- Select k items to serve as centers.

2- Calculate the distance between each remaining item and the centers

3- Assigned each item to the cluster with the nearest center.

4- Replaced the centers in step (1) by the centroids (mean vectors) of the clusters.

5- Calculate the distance between each item and the centroid of own cluster, and the centroid of another cluster. If the item is closer to the centroid, of another cluster than to the centroid of its own cluster, the item is moved to the new cluster and the two cluster centroids are updated.

6- This procedure will be continued until no further improvement is possible, (Rencher (2002)).

2- 2- 2- Fuzzy Clustering Methods

In hard clustering methods, each data element belongs to exactly one cluster. In fuzzy clustering, observations can belong to more than one cluster, with possibilities. The Fuzzy C-Means (FCM) Method is one of the famous fuzzy clustering methods. Let $X=\{x_1,x_2,\dots,x_n\}$ be a set of giving data, and $p=\{A_1,A_2,\dots,A_c\}$ be a family of fuzzy subsets of X which satisfies $\sum_{i=1}^{c} A_i(x_k)=1$ for all $k=1,2,\dots,n$ and $0<\sum_{k=1}^{n} A_i(x_k)<n$ for all $i=1,\dots,c$, where $c<n$ is a positive integer and p is called a fuzzy c-partition of X. The c cluster centers $v_1,v_2,\dots,v_c$ are calculated by the formula

$$v_i=\frac{\sum_{k=1}^{n}\left[A_i(x_k)\right]^m x_k}{\sum_{k=1}^{n}\left[A_i(x_k)\right]^m} \quad \text{-}for\ all\ i=1,\dots,c \tag{4}$$

where $x_k=\left[x_{k1},x_{k2},\dots,x_{kp}\right]\in R^p$, $m>1$ is a real number that governs the impact of membership grades. The performance index of a fuzzy c-partition p, $J_m(p)$, is defined by the formula:

$$J_m(p)=\sum_{k=1}^{n}\sum_{i=1}^{c}\left[A_i(x_k)\right]^m\left\|x_k-v_i{}^2\right\|,$$

where $\left\|x_k-v_i{}^2\right\|$ is the distance between x_k $and\, v_i$. The goal of the FCM is to find a fuzzy c-partition that minimizes the performance index $J_m(p)$, Klir and Yuan (1995).

Let T be the maximum number of iterations allowed, $0<\varepsilon<1$, $v_i^{(t)}, t=0$; be the initial centers, $i=1,\dots,c$, and $V^{(t)}$ be the set of centers in the iteration t. The Fuzzy C-means Method has the following steps:

1- Choose the initial centers $\boldsymbol{v}_i$, and fix $\varepsilon>0$.

2- Compute the distance between the elements and the centers $\|x_k-v_i\|_p$.

3- Compute the membership, degree to assign the elements $\boldsymbol{x}_k$ to the clusters, according to:

$$A_i(x_k)=\left(\sum_{j=1}^{c}\left(\frac{\|x_k-v_i\|_p}{\|x_k-v_j\|_p}\right)^{2/(m-1)}\right)^{-1}.$$

4- Compute the new cluster's centers:

$$v_i=\sum_{k=1}^{n}[A_i(x_k)]^m x_k \Big/ \sum_{k=1}^{n}\left[A_i(x_k)\right]^m.$$

5- If $\left\|V^{(t-1)}-V^{(t)}\right\|_p<\varepsilon$; or $T=t$, then stop. Otherwise, go to step 2.

A list of advantages and limitations of FCMare as follows, Yang (1993), and Suganya and Shanthi (2012).

Advantages

1-Unsubstantiated: The FCM technique is an unsubstantiated method that the data may be a considered and class information is unreachable.

2-Converges: The convergence features of algorithms are the significant theoretical matter. The optimal cluster centers are the stable points of FCM clustering algorithms.

Disadvantages

1-The computational time is long.

2- It is sensitive to the initial speculation (velocity, local minima).

3- It is sensitive to noise and outliers.

It is worth mentioning that there are some fuzzy clustering methods which can be compared with the FCM in terms of advantages and disadvantages, as follows:

1-Possibilistic Fuzzy Clustering (PFC): The PFC minimizes an objective function and generally leads to reasonable results, although it suffers from stability problems. The disadvantage of PFC analysis is that the sum of the membership degrees to all clusters must sum up to 1, (Timm et al. (2004)).

2-Fuzzy K-Modes Clustering (FKM): The FKM salgorithm is suited to large categorical data sets. The cost of the FKM algorithm is lower than that of the K-Means algorithm. The FKM algorithm not only split items into clusters, but display how confident an object is assigned to a cluster. The time utilized by the FKM algorithm was less than that used by the hierarchical clustering algorithm. The tricky part of the FKM algorithm is to reduce the cost function and to decrease the computational complexity, Huang and Ng (1999).

3-Fuzzy Shell-Clustering (FSC): The FSC algorithms are new methods for detecting curve boundaries, especially a circular and elliptical. These FSC algorithms are good in the computer storage requirements and computational requirements, converges quickly and keeps computation time. The FSC method is eligible to produce a perfect classification. Large features of the FSC method is in the areas of memory storage requirements and computational requirements. But this method is less strong to noise, Dave (1989).

2- 3- The Employed Methods in This Study

In this research, we used all the above methods, including statistical clustering techniques (hierarchical clustering and non-hierarchical clustering) and Fuzzy C-Means method to cluster 15 provinces of Iraq based on some agricultural productions. Using Silhouette criterion, it is verified to know which of these methods are more suitable for the clustering of these data.

3- Data Set

Among a large number of crops, wheat and barley are of great economic importance and are regarded as the strategic agricultural products owing to their close relationship with daily human and animal nutrition. Wheat is one of the oldest known field crops grown in the world as a primary source of food. Wheat is grown in Iraq in very large areas, especially northern governorates. The barley crop is another important economic crop that is grown in all parts of Iraq and is regarded as one of the most important sources of food for livestock. Wheat is the most important product of cereal crops and rice comes in the second place. Hence, wheat is considered as the most important cereal crop in Iraq in terms of both production and consumption. The winter wheat crop is dependent on either irrigation or rain, which are often available in the northern provinces of Iraq (Nineveh/Dahuk/Erbil/Sulaymaniyah). In the central and southern regions, wheat production relies mainly on irrigation from rivers and its cultivation is concentrated around the basin of the Tigris and Euphrates. However, this is not the case in other provinces located in western Iraq. It is worthy to mention that river irrigation plays a key role in the production of wheat the other crops. Barley and corn are also raised in Iraq; however, they are mostly used as animal feed. The annual productions of wheat, barley and yellow corn by the cities of Iraq are summarized in Table 2.1 for 2010 according to their relative importance, (Al-Fahad and Abbas (2011)).

To illustrate the application of agglomerative techniques, we used the observations listed in Table 3.1 and the distance of Euclidean and Manhattan, respectively. The obtained results are depicted in Table 3.2, Table 3.3 (see Appendix for tables).

By using Euclidean distance, the results of average, centroid and median methods are similar. However, with Manhattan distance, the results of complete linkage and average methods are similar, also centroid and median methods have the same results. In addition, the results of the Ward's method based on Euclidean and Manhattan distances are similar and the results of the flexible beta method based on Euclidean and Manhattan distances are similar.

Figures 1-5 illustrate the results of the agglomerative techniques based on Euclidean and Manhattan distances. For example, in Figure 1, the result of a single linkage method based on Euclidean distance is shown in several stages, at each stage the two closest provinces merge into one cluster as follows:

C1={Karbala, Basra}, C2={Diyala, Anbar}, C3={Salahaddin, C2}, C4={C1, Muthanna}, C5={C3, Najaf}, C6={C5, C4}, C7={C6, Maysan}, C8={Dhi Qar,C7}, C9={C8, Diwaniya}, C10={Baghdad, Babylon}, C11={C10, Kirkuk}, C12={C9, Wasit}, C13={C12, C11}, C14={ Nineveh, C13}.

Observations are used in Table 3.1 to illustrate the divisive hierarchical methods, as in Table 3.4.

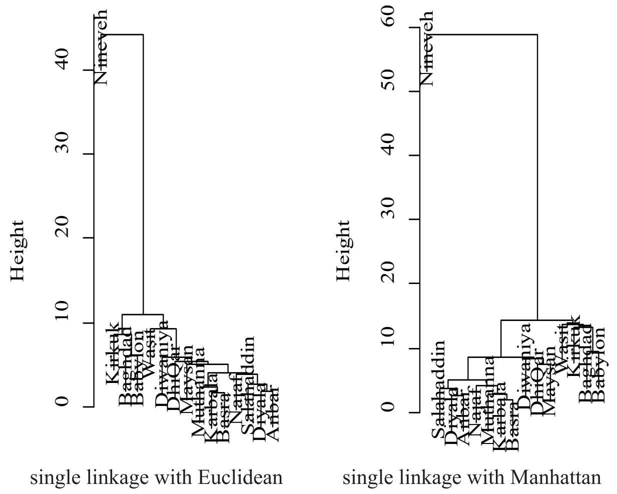

Fig. 1. Cpmparison between the results of the single linkage method based on Euclidean and Manhattan distances for data in Table 3.1

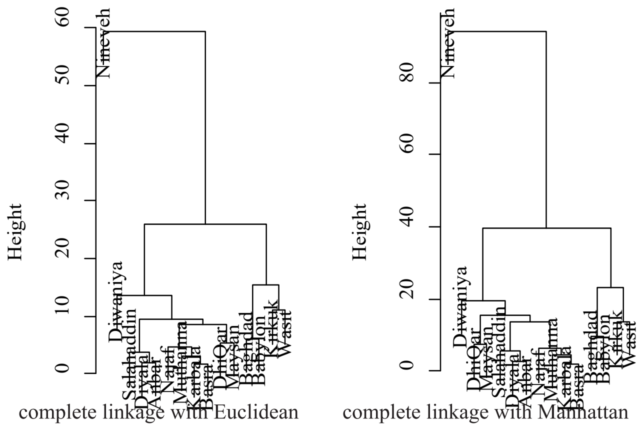

Fig. 2. Cpmparison between the results of the complete linkage method based on Euclidean and Manhattan distances for data in Table 3.1

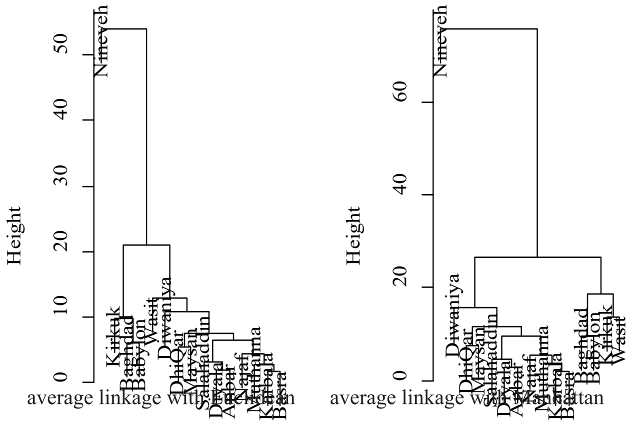

Fig. 3. Cpmparison between the results of the average method based on Euclidean and Manhattan distances for data in Table 3.1

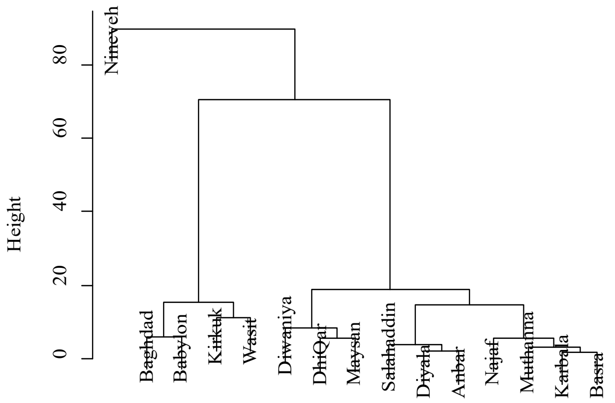

Fig. 4. The results of the Ward's method based on Euclidean and Manhattan distances for data in Table 3.1

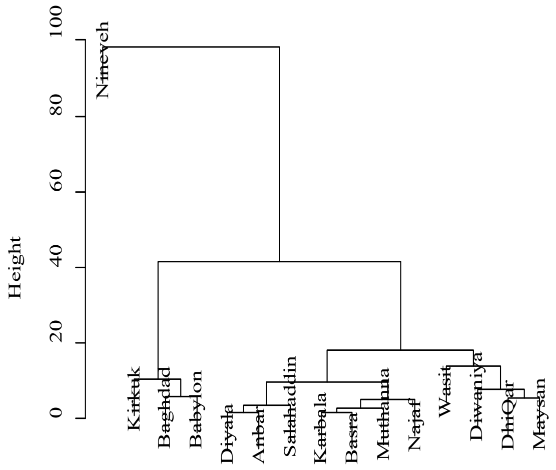

Fig. 5. The results of the flexible beta method based on Euclidean and Manhattan distances for data in Table 3.1

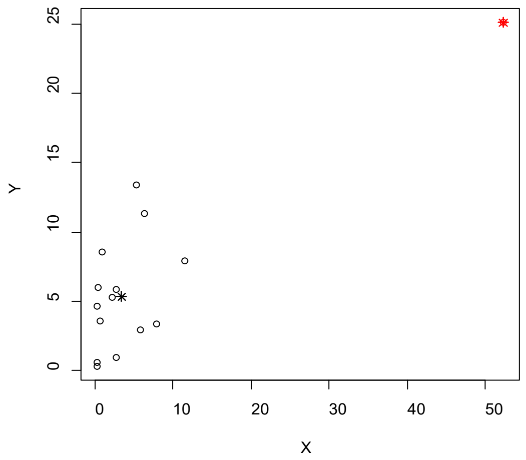

Fig. 6. K-Means clustering (with 2 centers) of the data in Table 3.1

Since Nineveh is the major and has the largest average distance to other cities, it is the initial item in the splinter group. Currently, we have been applying the steps 3-6 above, we obtained Table 3.5, (see Appendix for tables). The second column depicts the average distance from each city (except Nineveh) to the remaining cities. The third column illustrates the average distance from each city of Nineveh, and the fourth column indicates the variance between the second and third columns. Since the largest difference in the third column is not positive, the process halts and the division is complete.
To illustrate the K-Means algorithm, we consider the data in Table 3.1. We choosed 2, by which the data were divided in two clusters of sizes 14 and 1, as follows.

Cluster means:

```
       B      W      Y
1    3.41   5.35   7.14
2   52.27  25.10   0.03
```

Where 1 and 2 are the final centers of clusters 1 and 2, respectively.

Clustering vectors:

```
  Nineveh   Kirkuk   Diyala   Anbar
       2        1        1       1
 Baghdad   Babylon   Karbala   Wasit
     1         1         1         1
  Wasit   Salahaddin  Najaf   Diwaniya
     1         1         1        1
 Muthanna   Dhi Qar   Maysan    Basra
     1          1         1        1
```

Where, Diyala, Anbar, Karbala, Salahaddin, Najaf, Diwaniya, Muthanna, Dhi Qar, Maysan, Kirkuk, Baghdad, Babylon, Wasit, and Basra are located in cluster 1 and Nineveh is located in the cluster 2. Also, once the ratio (SS_between / SS_total= 65.5 %) is decreased, can be concluded that it better results are obtained. It is worthy to mention that this quantity represents the ratio of between-cluster distances and the total distances.
Now, we employ the Fuzzy C-Means algorithm to data set in Table 3.1. Suppose that $c = 6$, $m = 6$,and $\|.\|$ is the Euclidean distance and the initial fuzzy partition is $p^{(0)} = \{A_1, A_2, \ldots, A_6\}$, then we have

$$A_1 = \left\{\frac{0.2}{x_1}, \frac{0.2}{x_2}, \ldots, \frac{0.2}{x_{15}}\right\},\ A_2 = \left\{\frac{0.13}{x_1}, \frac{0.13}{x_2}, \ldots, \frac{0.13}{x_{15}}\right\}$$

$$A_3 = \left\{\frac{0.17}{x_1}, \frac{0.17}{x_2}, \ldots, \frac{0.17}{x_{15}}\right\},.\ A_4 = \left\{\frac{0.15}{{}_1x_1}, \frac{0.15}{{}_2x_2}, \ldots, \frac{0.15}{{}_{15}x_{15}}\right\}$$

$$A_5 = \left\{\frac{0.18}{x_1}, \frac{0.18}{x_2}, \ldots, \frac{0.18}{x_{15}}\right\},\ A_6 = \left\{\frac{0.17}{x_1}, \frac{0.17}{x_2}, \ldots, \frac{0.17}{x_{15}}\right\}$$

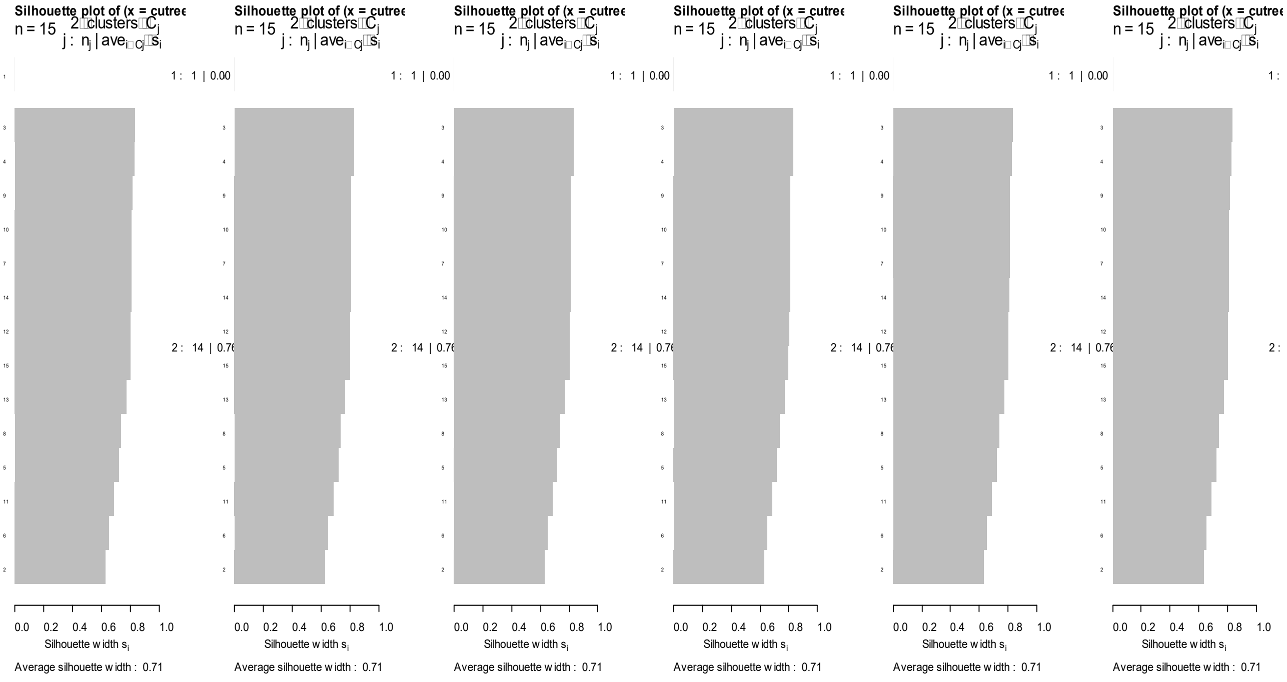

Fig. 7. Evaluation and comparison the clustering algorithms for data of Table 3.1

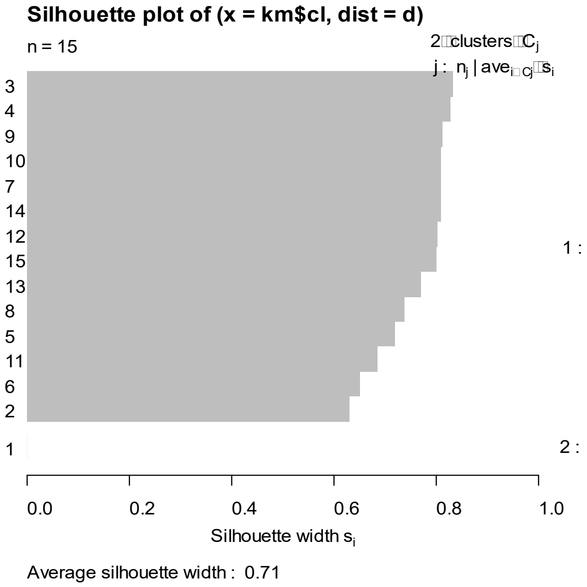

Fig. 8. Evaluation of the K-Means method for data of Table 3.1

Table 3.1. Crop production by relative importance percentage

Province	Barley	Wheat	Yellow corn
Nineveh	52.27	25.1	0.03
Kirkuk	5.3	13.36	22.1
Diyala	2.18	5.3	3.4
Anbar	0.48	5.99	3.7
Baghdad	0.72	3.6	20.1
Babylon	2.71	5.88	25.3
Karbala	0.35	0.34	1.6
Wasit	6.34	11.34	11.4
Salahaddin	0.94	8.58	4.4
Najaf	0.33	4.67	0.1
Diwaniya	11.53	7.95	0.9
Muthanna	2.73	0.96	0
Dhi Qar	7.9	3.35	1
Maysan	5.87	2.96	6
Basra	0.35	0.62	0

Table 3.2. Matrix of Euclidean distance between Provinces

Province	Euclidean distance														
Nineveh	0														
Kirkuk	53.2	0													
Diyala	53.1	20.6	0												
Anbar	55.3	20.4	1.9	0											
Baghdad	59.4	11	16.9	16.6	0										
Babylon	58.9	8.5	21.9	21.7	6.	0									
Karbala	57.5	24.8	5.6	6	18.8	24.5	0								
Wasit	49.3	10.9	10.9	11.1	12.9	15.4	15.9	0							
Salahaddin	54.1	18.8	3.7	2.7	16.5	21.2	8.7	9.3	0						
Najaf	55.8	24.2	3.8	3.8	20	25.3	4.6	14.4	5.8	0					
Diwaniya	44.2	22.8	10	11.6	22.5	26	13.5	12.2	11.2	11.7	0				
Muthanna	55.1	25.5	5.5	6.6	20.4	25.8	2.9	15.8	9	4.4	11.3	0			
Dhi Qar	49.4	23.5	6.5	8.3	20.4	25	8.2	13.2	9.4	7.7	5.9	5.8	0		
Maysan	51.8	19.2	5.1	6.6	15	19.8	7.5	10	7.6	8.3	9.1	7.1	5.4	0	
Basra	57.4	26	6.1	6.5	20.3	26	1.6	16.8	9.1	4.1	13.4	2.4	8.1	8.5	0

Table 3.3. Matrix of Manhattan distance between Provinces

Province	Manhattan distance														
Nineveh	0														
Kirkuk	80.8	0													
Diyala	73.3	29.9	0												
Anbar	74.6	30.6	2.7	0											
Baghdad	93.1	16.3	19.9	19	0										
Babylon	94.1	13.3	23	23.9	9.5	0									
Karbala	78.3	38.5	8.6	7.9	22.1	31.6	0								
Wasit	71.1	13.8	18.2	18.9	22.1	23	26.8	0							
Salahaddin	72.2	26.8	5.5	3.8	20.9	25.4	11.6	15.2	0						
Najaf	72.4	35.7	5.8	5.1	21.5	28.8	5.9	24	8.8	0					
Diwaniya	58.8	32.8	14.5	15.8	34.4	35.3	19.5	19.1	14.7	15.3	0				
Muthanna	73.7	37.1	8.3	11	24.8	30.2	4.6	25.4	13.8	6.2	16.7	0			
Dhi Qar	67.1	33.7	10.1	12.8	26.5	32	11.2	20	15.6	9.8	8.3	8.6	0		
Maysan	74.5	27.1	8.6	10.7	19.9	25.4	12.5	14.3	12.2	13.2	15.8	11.1	7.4	0	
Basra	76.4	39.8	9.9	9.2	23.5	32.9	1.9	28.1	13	4.2	19.4	2.7	11.3	13.9	0

Table 3.4. Illustration the average distance of each city from the remaining cities

Province	Average distance	Province	Average distance
Nineveh	53.95	Salahaddin	13.36
Kirkuk	22.09	Najaf	13.86
Diyala	12.31	Diwaniya	16.09
Anbar	12.76	Muthanna	14.11
Baghdad	19.75	Dhi Qar	14.05
Babylon	23.17	Maysan	12.92
Karbala	14.29	Basra	14.72
Wasit	15.57		

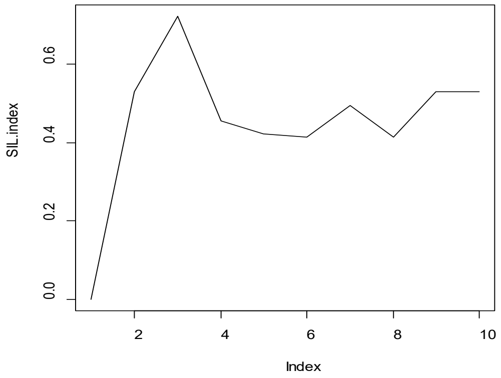

Fig. 9. Evaluation of Fuzzy C-Means method for data of Table 3.1 based on Silhouette index.

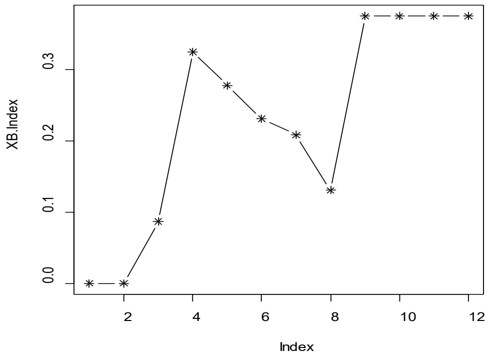

Fig. 10. Evaluation of Fuzzy C-Means method for data of Table 3.1 based on Xie-Beni index.

Table 3.5. Illustration of the average distance of each city from the remaining cities, and splinter groups.

Province	Average distance to remainder(1)	Average distance to splinter group(2)	Difference(1)-(2)
Kirkuk	19.7	53.21	-33.15
Diyala	9.1	53.97	-44.87
Anbar	9.48	55.32	-45.84
Baghdad	16.71	59.35	-42.64
Babylon	20.54	58.86	-38.32
Karbala	10.97	57.54	-46.57
Wasit	12.98	49.28	-36.3
Salahaddin	10.22	54.10	-43.88
Najaf	10.63	55.81	-45.18
Diwaniya	13.93	44.21	-30.28
Muthanna	10.96	55.11	-44.15
Dhi Qar	11.33	49.42	-38.09
Maysan	9.93	51.76	-41.83
Basra	11.44	57.40	-45.96

Table 3.6. Result of Fuzzy C-Means method.

Province	$A_i(x_k), i$					
	$A_1(x_k)$	$A_2(x_k)$	$A_3(x_k)$	$A_4(x_k)$	$A_5(x_k)$	$A_6(x_k)$
Nineveh	0.16	0.17	0.18	0.16	0.17	0.16
Kirkuk	0.15	0.15	0.15	0.16	0.16	0.23
Diyala	0.17	0.16	0.13	0.26	0.17	0.1
Anbar	0.06	0.05	0.05	0.75	0.06	0.04
Baghdad	0.15	0.15	0.14	0.16	0.16	0.24
Babylon	0.03	0.03	0.03	0.03	0.04	0.83
Karbala	0.24	0.16	0.13	0.18	0.17	0.10
Wasit	0.15	0.16	0.17	0.18	0.18	0.15
Salahaddin	0.16	0.16	0.14	0.26	0.17	0.11
Najaf	0.21	0.17	0.14	0.22	0.16	0.10
Diwaniya	0.03	0.04	0.84	0.03	0.03	0.02
Muthanna	0.77	0.05	0.04	0.05	0.05	0.03
Dhi Qar	0.04	0.80	0.04	0.04	0.05	0.03
Maysan	0.04	0.05	0.04	0.04	0.80	0.03
Basra	0.26	0.16	0.13	0.18	0.16	0.10

Fuzzy six partitions $p^{(9)} = \{A_1, A_2, \ldots, A_6\}$ are given in Table 2.6. Then, the algorithm stops at $t = 9$, and we obtain the six fuzzy partitions. The six cluster centers are

$$v_1 = (2.73, 0.96, 0.01),\quad v_2 = (7.9, 3.35, 1),$$
$$v_3 = (11.53, 7.95, 0.9),\quad v_4 = (0.49, 5.99, 3.7),$$
$$v_5 = (5.87, 2.96, 6),\quad v_6 = (2.71, 5.88, 25.29)$$

4- Evaluation and Comparison of Clustering Methods

The Silhouette value is a measure of the similarity of an object to its own cluster compared to its similarity to other clusters. It can be used to choose a number of clusters. For each object, i a certain value $s(i)$ will be introduced, and then these numbers are combined into a plot. Let A be any cluster and B, is a different cluster from A, and i be any object in the cluster A, the Silhouette value is defined as:

$$\mathrm{s}(i) = \begin{Bmatrix} \dfrac{1-a(i)}{b(i)} & \mathit{if}\ a(i) < b(i) \\ 0 & \mathit{if}\ a(i) = b(i) \\ \dfrac{b(i)}{a(i)-1} & \mathit{if}\ a(i) > b(i) \end{Bmatrix}$$

Table 4.1. The Silhouettes of the agglomerative hierarchical methods for the clustering of the data in Table 3.1 into k=2 clusters.

Province	Cluster	Neighbor	
[1]	1	2	0.0000
[2]	2	1	0.6297
[3]	2	1	0.8313
[4]	2	1	0.8278
[5]	2	1	0.7185
[6]	2	1	0.6510
[7]	2	1	0.8093
[8]	2	1	0.7366
[9]	2	1	0.8110
[10]	2	1	0.8095
[11]	2	1	0.6849
[12]	2	1	0.8011
[13]	2	1	0.7707
[14]	2	1	0.8081
[15]	2	1	0.8006

Table 4.2. The Silhouettes of the K-Means method for data in Table 3.1 the clustering into *k*=2 clusters.

Province	Cluster	Neighbor	*s* (*i*)
[1,]	2	1	0.0000
[2,]	1	2	0.6297
[3,]	1	2	0.8313
[4,]	1	2	0.8278
[5,]	1	2	0.7185
[6,]	1	2	0.6510
[7,]	1	2	0.8093
[8,]	1	2	0.7366
[9,]	1	2	0.8110
[10,]	1	2	0.8095
[11,]	1	2	0.6849
[12,]	1	2	0.8011
[13,]	1	2	0.7707
[14,]	1	2	0.8081
[15,]	1	2	0.8006

It can be also written this in the form of

$$s(i)=\frac{b(i)-a(i)}{max\,\{a(i),b(i)\}},$$

where $a(i)$ = Average dissimilarity of i to all other objects of A, $d(i,B)$= Average dissimilarity of i to all objects of B, $b(i)=\min_{B\neq A} d(i,B)$.

From the above definition, we see that $-1\leq s(i)\leq 1$. When $s(i)$ close to 1, this implies that the 'within' dissimilarity $a(i)$ is much smaller than the smallest 'between' dissimilarity $b(i)$, and mean that all the clusters are separate in a partition, Rousseeuw (1987).

Xie-Beni index (XB) is a fuzzy validity criterion depended on the compactness of fuzzy c-partition and separation of the clusters disregard supposition as to the number of substructures found in the data. XB is a measure of compactness divided by a measure of separation. It may be deduced as the ratio of the within-group variance total and the separation of the cluster centers. The optimum cluster number is got when the minimum of XB is established (Xi and Beni (1991)).

We used XB index for data in Table 3.1 to evaluate the FCM methods, and Silhouette value to evaluate the above-mentioned methods. The obtained results are described in the following sections.

4- 1- Evaluation of the Agglomerative Hierarchical Methods

We found the followings:

1-Maximum $s(i)\ =0.71$ when $k\ =2$.

2-$s(i)$ for the all above methods (agglomerative hierarchical methods) = 0.71.

Fig. 7 shows the Silhouettes for the clustering into $k=2$ clusters of the fifteen data mentioned above. The second Silhouette is higher than the first one because the first cluster contains only one object whereas the second contains fourteen. The first column (City) contains the number of cities, the second column (Cluster) shows the index of each cluster (1 and 2), the third column (Neighbor) gives the neighbor of each object, and the fourth column lists the numbers $s(i)$. The plot, as we find scale going from $0.00\,to\,1.00$.

4- 2- Evaluation of the K-Means Method

We found the following results:

Maximum $s(i)=0.71,\ when\ k\ =2$, and the results are as follows table 4.2.

4- 3- Evaluation of the Fuzzy C-Means Method

We found the Maximum $s\left(i\right)$ = 0.72 when k = 3 (see Figure 9), and the Manimim XB = 0.087 when k = 3 (see Figure 10).

5- Conclusion

In this research eight statistical clustering methods (hierarchical and non-hierarchical) as well as fuzzy clustering methods were investigated and compared based on some well-known criteria. Then, such methods were applied on a real-world data set of the distribution of some agricultural productions in the provinces of Iraq. When hierarchical methods were applied based on two types of distances, namely Euclidean and Manhattan, the following results were obtained:

1-The average, centroid, and median methods had the same results based on Euclidean distance.

2-The centroid and median methods had the same results by using Manhattan distance.

3-Ward's method gave the same results by using Euclidean and Manhattan distances.

In addition, a flexible beta method gave the same results as Euclidean and Manhattan distances.
When K-Means method was used, the better results were obtained when the ratio representing the between-cluster distances and the within-cluster distances is decreased. By using Silhouette value and comparison between various clustering methods (both statistical and fuzzy methods), it was found that the Fuzzy C-Means method is the best method for clustering such a data set. It should be mentioned that the clustering methods investigated in this article are general so that the methods and algorithms can be used in other agricultural studies.

References

[1] Y. Al-Fahad and T. Abbas, GIS Center Central Bureau of Statistics (NBS), Iraq, 2011.

[2] N. Aguilar-Gallegos, M. Munoz-Rodriguez, H. Santoyo-Cortes, J. Aguilar-Avila, and L. Klerkx, Information Networks that Generate Economic Value: A Study on Clusters of Adopters of New or Improved Technologies and Practices among Oil Palm Growers in Mexico, *Agricultural Systems*, vol. 135**,** pp. 122-132, 2015.

[3] A. Ansari, P. S. Sikarwar, S. Lade, H. K. Yadav and S. A. Randade, Genetic Diversity in Germplasm of Cluster Bean, an Important Food and an Industrial Legume Crop, *J. Agr. Sci. Tech*., vol. 18**,** pp. 1393-1406, 2016.

[4] D. J. Bora and A. K. Gupta, A Comparative Study between Fuzzy Clustering Algorithm and Hard Clustering Algorithm, *International Journal of Computer Trends and Technology,* vol. 10, pp. 785-790, 2014.

[5] C. T. Chang, J. Z. C. Lai and M. D. Jeng, A Fuzzy K-means Clustering Algorithm Using Cluster Center Displacement, *Journal of Information Science and Engineering,* vol. 27, pp. 995-1009, 2011.

[6] S. Chattopadhyay, D. K. Pratihar, S. C. D Sarkar, A Comparative Study of Fuzzy C-Means Algorithm and Entropy-Based Fuzzy Clustering Algorithms, *Computing and Informatics*, vol.30, pp. 701–720, 2011.

[7] R. Dave, Fuzzy Shell-Clustering and Applications to Circle Detection in Digital Images, *lnt. J. General Systems*, vol. 16,. pp. 343-355, 1989.

[8] J. V. De Oliveira and W. Pedrycz, Advances in Fuzzy Clustering and Its Applications, *John Wiley and Sons*, New York, 2007.

[9] M. Fajardo, A. Mc. Bratney and Whelan, Fuzzy Clustering of Vis–NIR Spectra for the Objective Recognition of Soil Morphological Horizons in Soil Profiles, *Geoderma,* vol. 263, pp. 244–253, 2016.

[10] M. B. Ferraro, and P. Giordani, A Toolbox for Fuzzy Clustering Using the R Programming Language, *Fuzzy Sets and Systems,* vol. 279, pp. 1–16, 2015.

[11] G. Gan, C. Ma and J. Wu, Data Clustering Theory, Algorithms, and Applications, *SIAM*, Virginia, 2007.

[12] C. Gomathi and K. Velusamy, Solving Fuzzy Clustering Problem Using Hybridization of Fuzzy C-Means and Fuzzy Bee Colony Optimization, *International Journal of Computer Engineering and Applications*, vol. 12, pp. 317–324, 2018.

[13] N. Grover, A Study of Various Fuzzy Clustering Algorithms, *International Journal of Engineering Research*, vol. 3, pp. 177–181, 2014.

[14] Z. Huang and M. Ng, A Fuzzy K-Modes Algorithm for Clustering Categorical Data. *IEEE Transactions on Fuzzy Systems*, vol.7, pp.446–452, 1999.

[15] J. G. Klir and B. Yuan, Fuzzy Sets and Fuzzy Logic: Theory and Applications, *Prentice Hall*, New York, 1995.

[16] P. Kostov and S. Mc Erlean, Using the Mixtures- of Distributions Technique for the Classification of Farms into Representative Farms. *Agricultural Systems*, vol. 88, pp. 528-537, 2006.

[17] E. Mansour, A. B. Khaled, T. Triki, M. Abid, K. Bachar, and A. Ferchichi, Evaluation of Genetic Diversity among South Tunisian Pomegranate Accessions Using Fruit Traits and RAPD Markers. *J. Agr. Sci. Tech*., vol. 17**,** pp. 109-119, 2015.

[18] B. Panda, S. Sahoo, and S. K. Patnaik, A Comparative Study of Hard and Soft Clustering Using Swarm Optimization: *International Journal of Scientific & Engineering Research,* vol. 4, pp. 785- 790, 2013.

[19] P. J. Rousseeuw, Silhouettes A Graphical Aid to the Interpretation and Validation of Cluster Analysis, *Journal of Computational and Applied Mathematics,* vol. 20, pp. 53-65, 1987.

[20] A. C. Rencher, Methods of Multivariate Analysis, *John Wiley and Sons*, New York, 2002.

[21] H. Timm, C. Borgelt, C. Doring, and R. Kruse, An Extension to Possibilistic Fuzzy Cluster Analysis, *Fuzzy Sets and Systems*, 147, 3–16, 2004.

[22] T. Volmurgan, Austria Performance Comparison Between K-means and Fuzzy C- means, *Wulfenia Journal Using Arbitrary Data Points*, vol. 19, pp. 1-8, 2012.

[23] R. Suganya, and R. Shanthi, Fuzzy C- Means Algorithm - A Review, *International Journal of Scientific and Research Publications*, vol. 2, pp. 440-442, 2012.

[24] X.L. Xie, and G. Beni, A validity measure for fuzzy clustering, *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 8, pp. 841-847, 1991.

[25] M. S. Yang, Convergence Properties of the Generalized Fuzzy C-Means Clustering Algorithms, *Computer Math. Appl*, vol. 25, pp. 3-11, 1993.

Please cite this article using:

I. A. Z. Atiyah, S. M. Taheri, Statistical and Fuzzy Clustering Methods and their Application to Clustering

Provinces of Iraq based on Agricultural Products , *AUT J. Math. Comput.*, 1(1) (2020) 101-112.

DOI: 10.22060/ajmc.2019.14873.1013