*International Journal of Mathematics and Statistics Invention (IJMSI)*
*E-ISSN: 2321 – 4767 P-ISSN: 2321 - 4759*
*www.ijmsi.org Volume 7 Issue 1 || January, 2019 || PP-12-33*

# Comparison between Cluster Techniques for Clinical Data

Ahmed Mohamed Mohamed Elsayed
*Al-Obour High Institute For Management & Informatics Department of Basic Science*
*Kilo 21 Cairo-Belbies Road, P.O. Box 27 Obour City, Egypt*

***ABSTRACT:*** *Data clustering has a widely used in many practical fields. The clustering process, as important method of data mining, is similar to classification process for data input. It forms groups based on object similarities. There are various techniques for the data clustering. The most popular ones are Kmeans, Kmedoids (PAM), Hierarchical and Model based. In this paper, all of these techniques are devoted and explained in details. Some packages of R program and hence some functions related to these packages are applied on the practical clinical data. There are various methods for selecting an appropriate number of clusters. For each technique, if it is possible, the optimal numbers of clusters are determined graphically depending on various measures. Also, in this research, we will study various measures of cluster validation, whether these measures are external or internal measures. The obtained results are comparable between all techniques to specify the best technique.*

***KEYWORDS****: Data mining; Hierarchical; Kmeans; Model based; Kmedoids(PAM); Clustering Validity; Silhouette Measure; Sum of Squared Errors.*

---

Date of Submission: 08-03-2019 Date Of Acceptance: 28-03-2019

---

## I. INTRODUCTION

Data clustering has a widely used in different applications. Used data must be standardized (scaled) to make variables comparable. The nominal variables, if exist, must be eliminated from the original data. To compare between clustering techniques, we need some information based on distances [15]. In R program, the Euclidean distance is used by default to measure the dissimilarity between each pair of observations. There are many methods to compute dissimilarities between two clusters such as: Complete method that considers the largest value of dissimilarities as a distance. Single method that considers the smallest of these dissimilarities as a distance. Average method that considers the average of these dissimilarities as a distance. Centroid method that computes the dissimilarity between the centroids of two clusters. Finally, Ward's method that minimizes the total within cluster variance. A pair of clusters with minimum between cluster distance are combined. It identifies the strongest clustering structure of the four methods [24].

Many references have devoted the classification and cluster analysis as important methods of data mining such as Gordon [5], and Kaufman and Rousseeuw [11]. Many researches devoted clustering technique such as [8, 10, 12, 18, 20]. An external clustering validation consists in comparing the results of a cluster analysis to external known results. An internal clustering validation uses the internal information of the clustering process to evaluate the goodness of a clustering structure without any references. We aim to make the mean distance within cluster be small, and the mean distance between clusters to be large as possible. Calinski and Harabasz [1] and Everittet et al. [4] presented methods for clustering analysis. There are various researches presented some methods for selecting an appropriate number of clusters such as [4, 16, 23]. Tippaya et al. [22] have studied the clustering validity techniques to quantify the appropriate number of clusters for Kmeans technique. Rousseeuw [19] presented graphs for interpretation and validation of cluster analysis. Then clustering validation measures are used to evaluate the results of a clustering technique [6,7,14,21,22,23]. The commonly used cluster validation indices are Silhouette width and Dunn index.

Silhouette width measures how well an observation is clustered, and it estimates the average distance between clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters. A large silhouette suggests the observations very well clustered, a small silhouette means that the observation lies between two clusters, and observations with a negative silhouette are placed in the wrong cluster. We can find the name of these samples and determine a neighbor cluster. If the data (well) separated, the diameter of the clusters expected to be (small) and the distance between the clusters expected to be (large).

Many techniques of clustering process are studied in this paper. As shown below in the next sections, Kmeans, Kmedoids (PAM), Hierarchical, Model based clustering techniques are explained. This research shows comparative results for all clustering techniques.

The aim of this article is to compute the validation clustering measures for different clustering techniques using some packages of R program. In each technique, we want to analyze the obtained results, display the plots of clustering process and construct the validation clustering measures, and of course if it is possible, determine the optimal number of clusters in each technique graphically using different measures.

The rest of this paper is organized as follows. Section II explains the material and methods. Section III presents the calculations. Section IV presents the cluster validation measures. Section V presents the discussion and conclusions.

## II. MATERIAL AND METHODS

Before presenting the types of clustering techniques, we will present some notations and some packages of R program and their related functions that are used in this paper.

| | |
|---|---|
| PAM | Partitioning Around Medoids. |
| CLARA | Clustering for Large Applications. |
| AGNES | Agglomerative Nesting. |
| DIANA | Divise Analysis. |
| HC | Hierarchical Clustering. |
| FPC | Flexible Procedures for Clustering. |
| BIC | Bayesian Information Criterion. |
| Rand Index<br>Meila's VI | Two indices to assess the similarity of two clustering[17],<br>VI: Variation of information [15] . |
| Pearson gamma | Correlation between distances, (0) means same cluster, (1) means different clusters. |
| Dunn | Min. separation / Max. diameter. Dunn index should be (maximized). |
| Dunn2 | Min. average dissimilarity between two cluster / Max. average within cluster dissimilarity. Another version of the family of Dunn index. |
| Entropy | Entropy measures the purity of the clusters class labels. However, as the class labels of objects in a cluster become more varied, the entropy increases. This is an external validation measure. |
| Wb. ratio | Average within/Average between. Should be (minimized). |
| C wide gap | Vector of widest within cluster gaps. |
| Widest gap | Widest within cluster gap. |
| S. index | Separation index. |

**Packages**

| | |
|---|---|
| Cluster | Computing PAM clustering, and for analyzing cluster silhouettes. |
| Factoextra | Simplifying clustering workflows. |
| ggplot2 | Visualizing clusters. |
| NbClust | Determining the optimal number of clusters. |
| Fpc | Computing clustering validation statistics. |

**Functions**

| **Function** | **Package** | |
|---|---|---|
| eclust | cluster | Stands for Enhancing Clustering. |
| pam<br>pamk<br>clara | fpc | Perform a partitioning around medoids clustering (with the number of clusters) estimated by optimum average silhouette width. |
| cluster.stats | fpc | Provides a mechanism for comparing the similarity of two clusters solution using a variety of validation criteria. |
| hclust | stats | Applies hierarchical clustering. |
| Mclust | Mclust | Select the optimal model according to BIC. Choose the model and number of clusters with the largest BIC. |
| agnes | cluster | For agglomerative hierarchical clustering. |
| diana | cluster | For divisive hierarchical clustering. |
| silhouette | cluster | Computes the silhouette coefficient of observations. |
| fviz_silhouette | factoextra | Draws silhouette plot, also print a summary of the silhouette analysis output. |
| NbClust | NbClust | Can be used to determine the numbers of clusters. |

The types of clustering techniques that used in this paper are: Kmeans, Kmedoids (PAM), Hierarchical, and Model based clustering.

**II.1 Kmeans Clustering (KC)**

In Kmeans technique, we want observations in the same group to be similar and observations in different groups to be dissimilar. It is commonly used clustering method for splitting a dataset into a set of *k*-groups. Each cluster represented by its center. The basic idea is defining clusters so that the total within cluster variation (minimized). The within groups sum of squares can help us to determine the appropriate number of clusters [3].

**II.2 Kmedoids Clustering (or PAM)**

The difference between Kmeans and Kmedoids is: Kmedoids represented with the object closest to the median of the cluster. PAM is a classic method for Kmedoids clustering. While the PAM technique is not suitable for clustering huge data, the CLARA is good [13]. The function pamk() in package fpc [9] does not require to specify *k*-clusters, it is not necessarily produce the best result.

**II.3 Hierarchical Clustering (HC)**

It is an alternative to Kmeans clustering method. It has an attractive tree, called a dendrogram. Hierarchical clustering divided into two main types:

**Agglomerative Clustering (Nesting):** It's also known as AGNES. Each observation is considered as a single element cluster. These combined clusters continue until having one big cluster. It is good for small clusters. If coefficient of AGNES near to (1), this leads to a strong clustering.

**Divisive Clustering (Divise Analysis):** It's also known as DIANA. It begins with the root cluster. The process of separation clusters continue until each observation become cluster. Diana is good for large clusters.

**II.4 Model Based Clustering (MBC)**

It applies maximum likelihood estimation and Bayesian criteria to identify the most likely model and number of clusters.

## III. CALCULATIONS

A respiratory clinical data is containing (555) observations and (7) variables [2], that can be presented as:

| | |
|---|---|
| **Center** | Two centers |
| **Treatment** | Placebo and Active treatment. |
| **Gender** | Female and Male. |
| **Age** | Age of the patient. |
| **Status** | Respiratory status (Poor and Good). |
| **Month** | Each patient was examined at months ( 1, 2, 3, 4 and 5). |
| **Subject** | Patient ID, From1 to 111. |

In each center, the patients were randomly selected. The experiment contains 111 patients (54 Active, 57 Placebo). During the treatment, the respiratory status (Poor or Good) was determined at each monthly visit. The question is: Is the treatment is effective or not?

Data are standardized (scaled) to make variables comparable. The categorical variables are coded as: Treatment (Active=1, Placebo=0), Gender (Male=1, Female=0), Status: (Good=1, Poor=0). The Euclidian method is used to calculate the distance between each pair of observations.

In the next subsections, Kmeans, Kmedoids (PAM), Hierarchical, and Model based techniques are used respectively applying on these data.

**III.1 Kmeans**

Using the (kmeans) function on the scaled data, we have the next results:

| No. Clusters | Cluster Size | Within sum of squares | Total within sum of squares | Total sum of squares | Between sum of squares | Ratio=Between sum of squares ÷ Total sum of squares |
|---|---|---|---|---|---|---|
| *k*=2 | 280<br>275 | 1291.152<br>1539.838 | 2830.989 | 3878 | 1047.011 | 26.99873 % |
| *k*=3 | 250<br>195<br>110 | 1028.7748<br>839.7801<br>528.7063 | 2397.261 | 3878 | 1480.739 | 38.18305 % |
| *k*=4 | 80<br>160 | 283.6247<br>594.1343 | 2123.863 | 3878 | 1754.137 | 45.23304 \|% |

| | | | | | | |
|---|---|---|---|---|---|---|
| | 120<br>195 | 406.3238<br>839.7801 | | | | |
| *k*=5 | 80<br>80<br>135<br>145<br>115 | 276.4022<br>283.6247<br>448.1071<br>553.4367<br>370.0696 | 1931.64 | 3878 | 1946.36 | 50.18978 % |

The ratio, (Between sum of squares ÷ Total sum of squares), is increased as number of clusters (*k*) increase. This returns to increase between sums of squares. The total within sum of squares decrease as (*k*) increases.

Figure 1 displays the data points according to the first two principal components for *k*=2, *k*=3, *k*=4 and *k*=5 respectively:

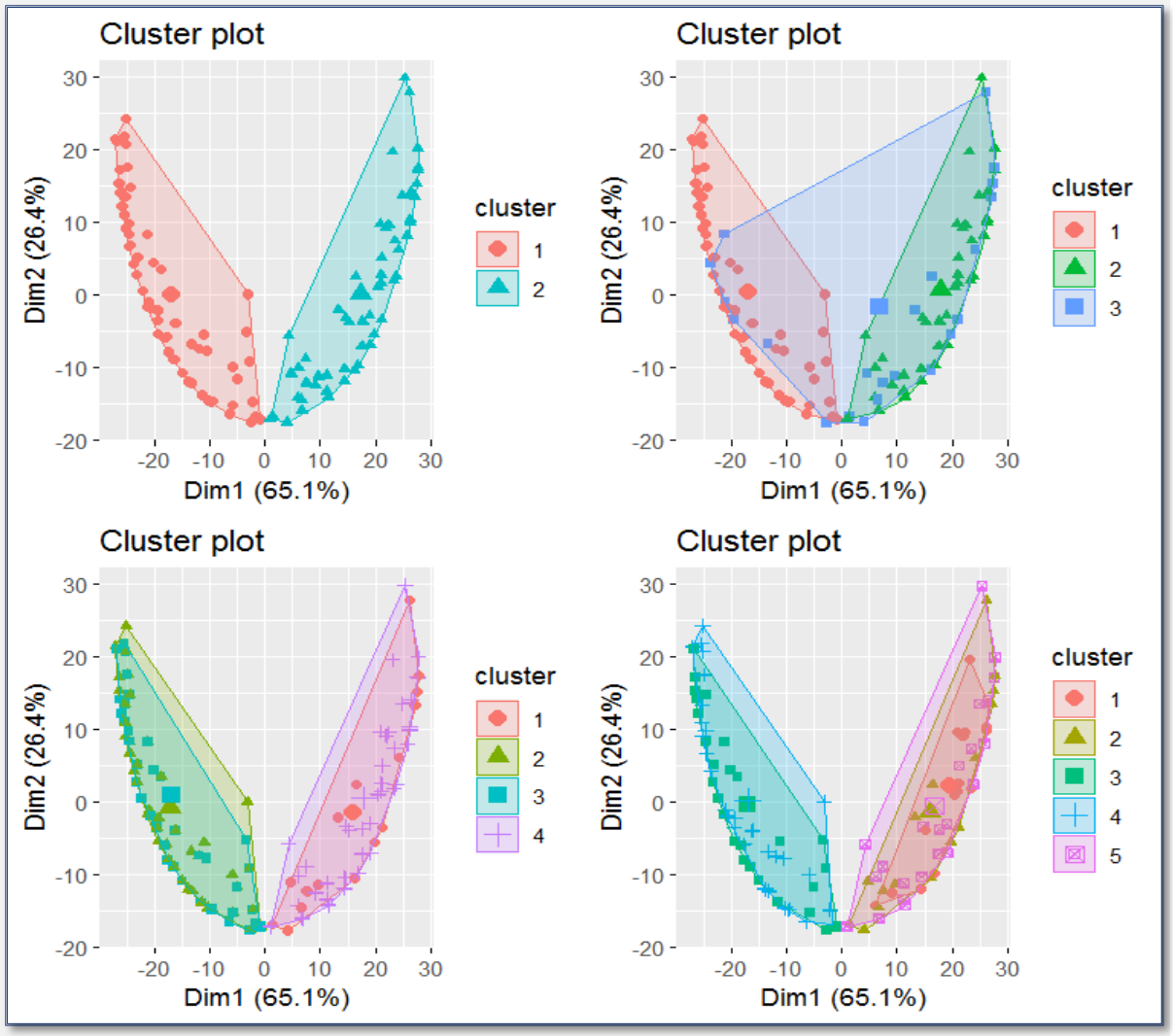


Figure 1: Data points according to the first two principal components.

A cross-tabulation can be computed as:

| **Treatment** | **1** | **2** | **3** | **Gender** | **1** | **2** | **3** | **Status** | **1** | **2** | **3** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Placebo | 125 | 81 | 80 | Female | 245 | 195 | 0 | Poor | 136 | 60 | 60 |
| Active | 125 | 114 | 30 | Male | 5 | 0 | 110 | Good | 114 | 135 | 50 |

When we clustered for treatment, gender, status variables used *k*=3, the cluster1 contains 250 objects, cluster2 contains 195 objects, and cluster3 contains 110 objects but in different details.

The optimal numbers of clusters, using Dindex values, can be determined as shown in Figure 2.

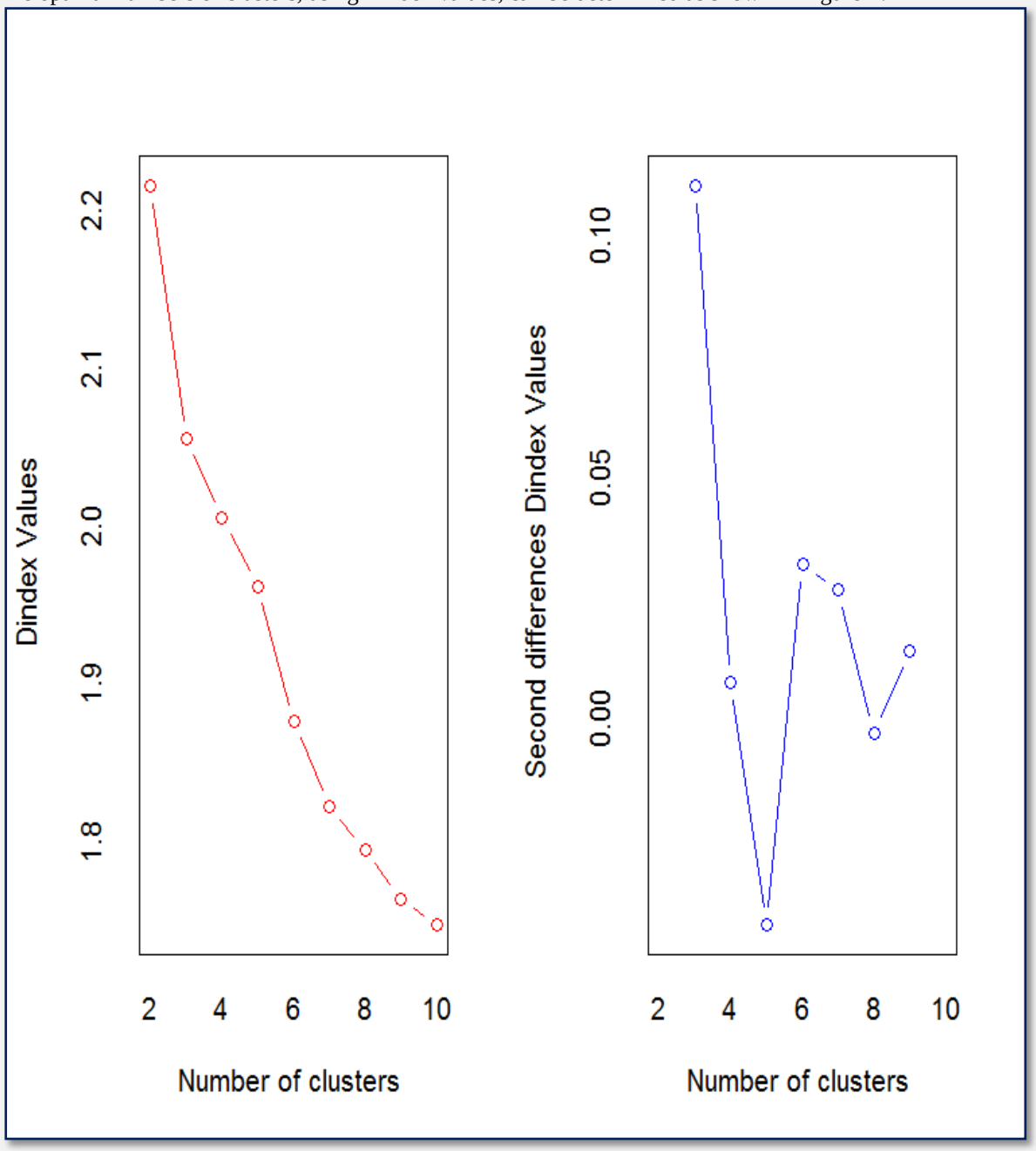


Figure 2: Optimal number of clusters - Kmeans

Figure 2 explains that the Dindex values decrease as number of clusters increase to $k$=10.

There are many methods to determine the optimal numbers of clusters such as: Elbow method, Average silhouette method and Gap statistic method.

Figure 3 explains the Elbow method:

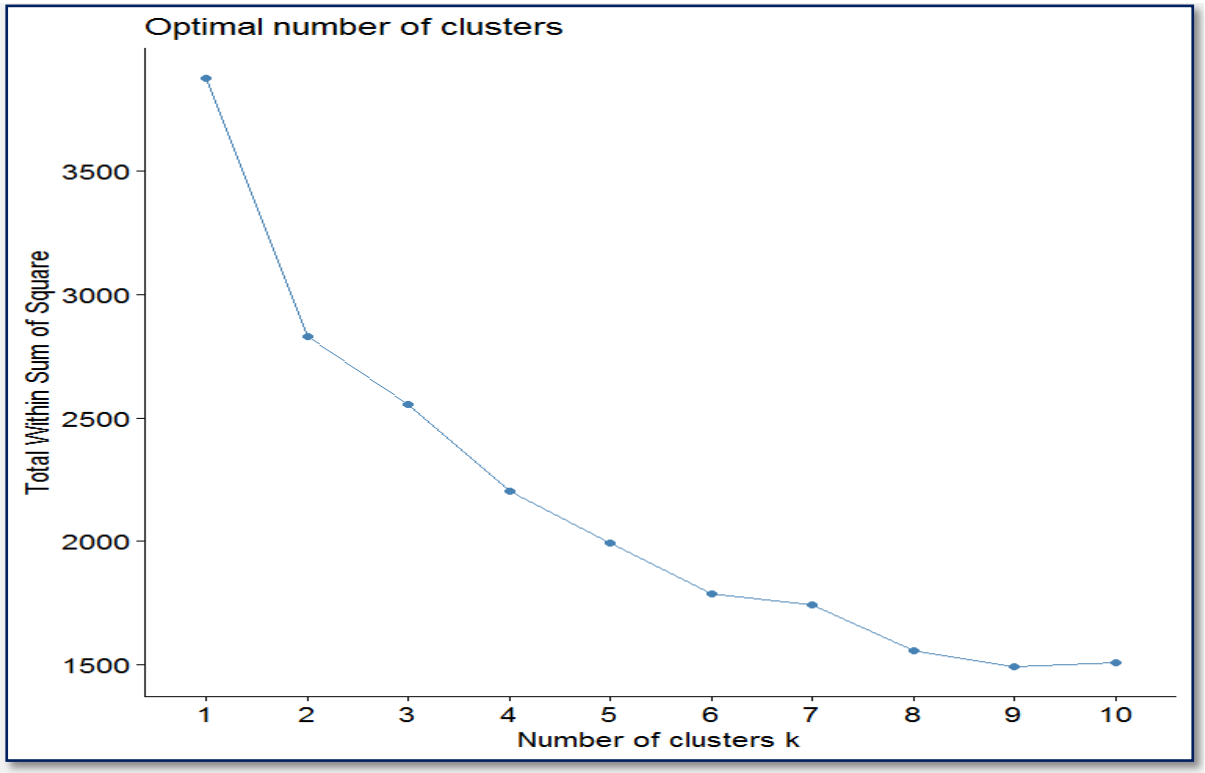


Figure 3: Elbow method - Kmeans

From Figure 3, the total within sum of squares is minimized at $k$=10 clusters.

Figure 4 explains the Average silhouette method:

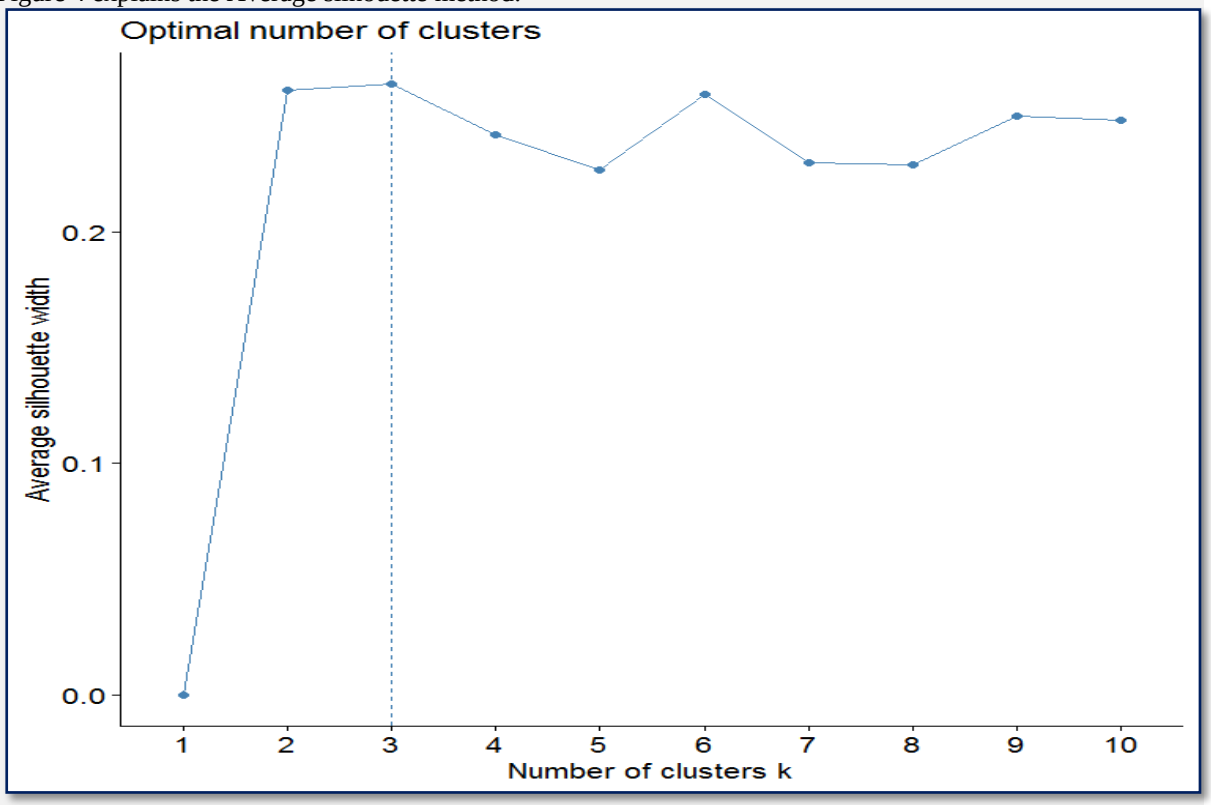


Figure 4: Silhouette method - Kmeans

From Figure 4 the average silhouette width is maximized at *k*=3 clusters.

Figure 5 explains the Gap statistic method:

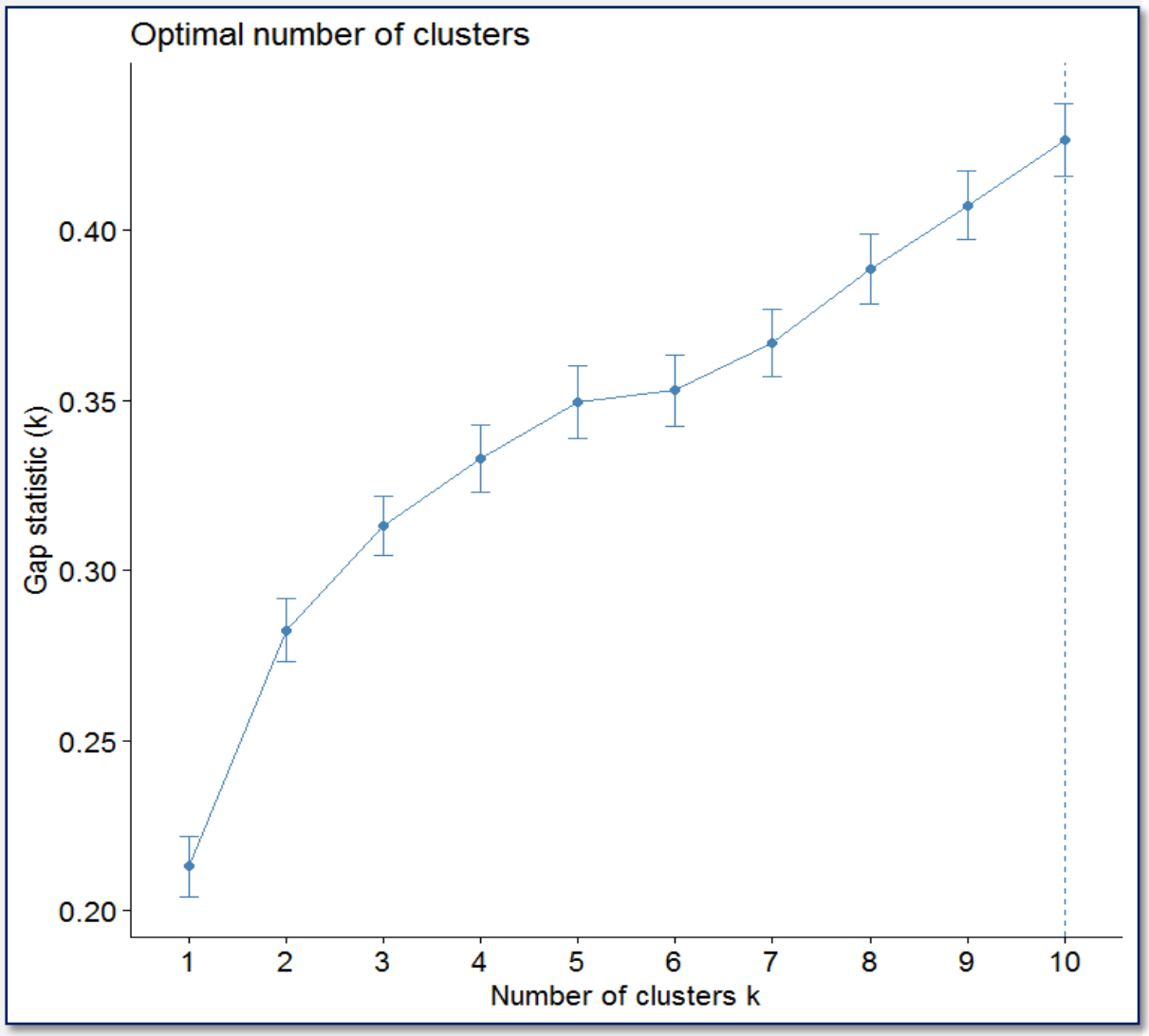


Figure 5: Gap statistic method - Kmeans

From Figure 5, Gap statistic factor is maximized at *k*=10 clusters

**III.2 Kmedoids (PAM-Pamk)**

In this section, we used (pamk) function to cluster data. The hierarchical clustering is done without determine the number of clusters. The used data is scaled data. Figure 6 displays the Kmedoids clustering explaining the average silhouette for *k*=10 clusters, without specify the *k* clusters before.

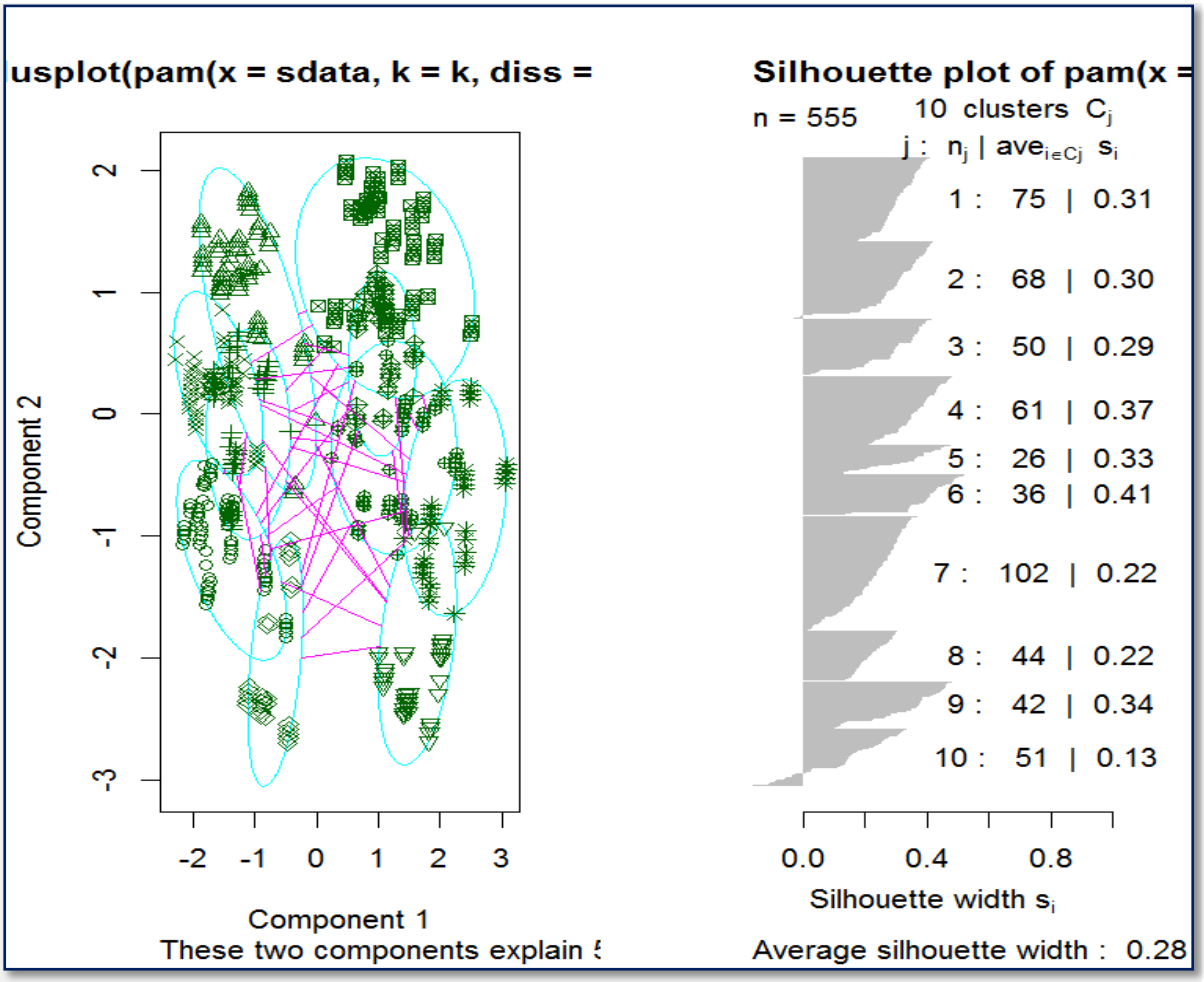


Figure 6: Pamk clustering

Figure 6 explains that: an average silhouette width is 0.28, and demonstrates the number of objects and average silhouette for each cluster. The same results must be similar if *k*=10, and the (pam) function is used.

A cross-tabulation for the treatment, gender and status variables can be computed as follow:

| **Treatment** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
|---|---|---|---|---|---|---|---|---|---|---|
| Placebo | 75 | 0 | 50 | 0 | 20 | 34 | 0 | 26 | 42 | 38 |
| Active | 0 | 68 | 0 | 61 | 6 | 2 | 102 | 18 | 0 | 13 |
| **Gender** | | | | | | | | | | |
| Female | 75 | 64 | 45 | 61 | 0 | 0 | 102 | 0 | 42 | 51 |
| Male | 0 | 4 | 5 | 0 | 26 | 36 | 0 | 44 | 0 | 0 |
| **Status** | | | | | | | | | | |
| Poor | 75 | 0 | 0 | 61 | 24 | 36 | 16 | 0 | | |
| Good | 0 | 68 | 50 | 0 | 2 | 0 | 86 | 44 | 38 | 11 |

The previous table ensures the clusters size as shown in figure 6.

**III.3 Hierarchical**

If we choose 20 objects (Active=8, Placebo=12) represented as one sample, we can plot the dendrogram for the treatment variable as shown below in Figure 7 with *k*=3 :

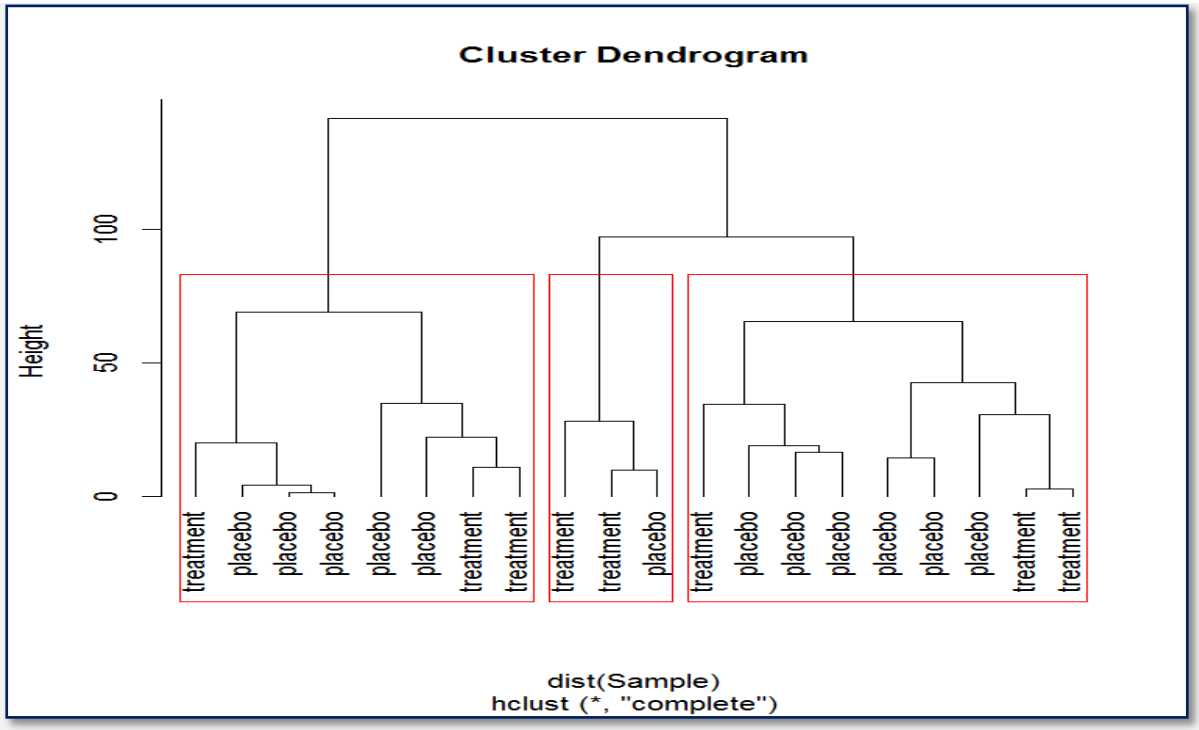


Figure 7: Dendrogram with *k*=3 , Hierarchical clustering

As we mentioned above, there are two methods for hierarchical clustering.

Agglomerative hierarchical clustering for 20 objects (Active=12 , Placebo=8) has AGNES coefficient = 0.84 as shown in Figure 8.

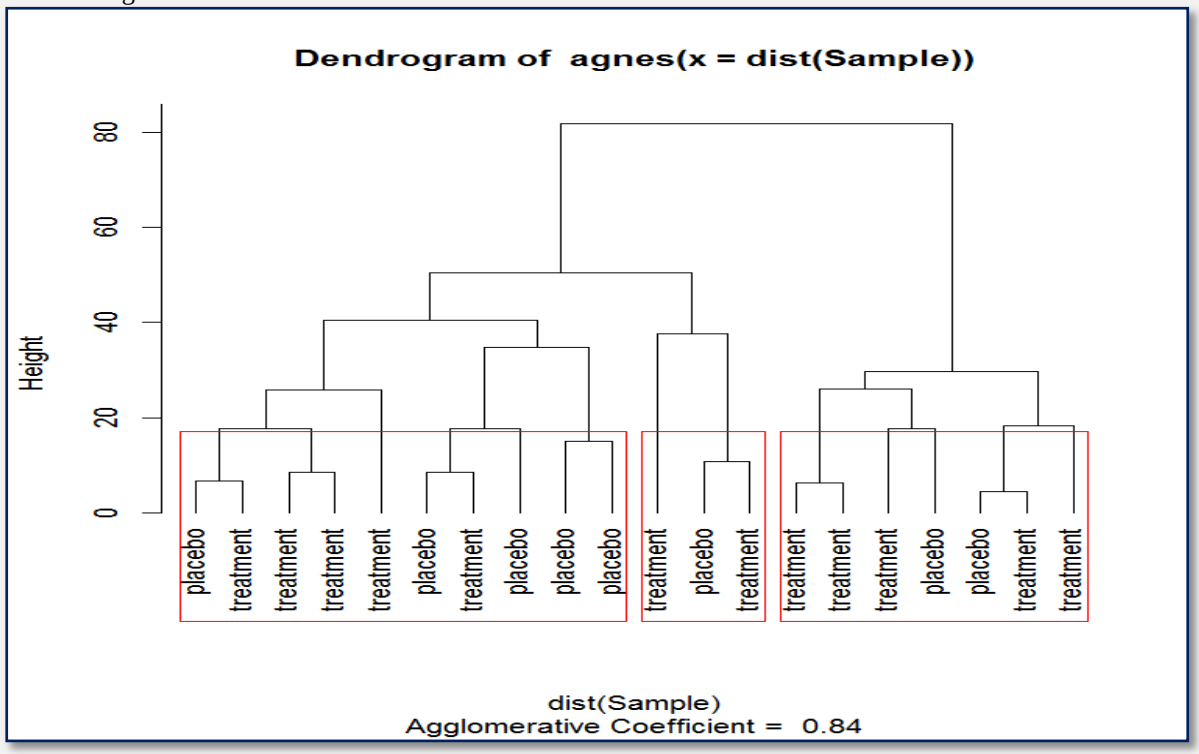


Figure 8: AGNES method *k*=3, Hierarchical clustering

AGNES coefficient (AC) for different linkages can be obtained as shown below:

| Single | Average | Complete | Ward |
|---|---|---|---|
| 0.929 | 0.980 | 0.989 | 0.999 |

Divisive hierarchical clustering for 20 objects (Active=7, Placebo=13) has DIANA coefficient (AC) = 0.87 as shown in Figure 9.

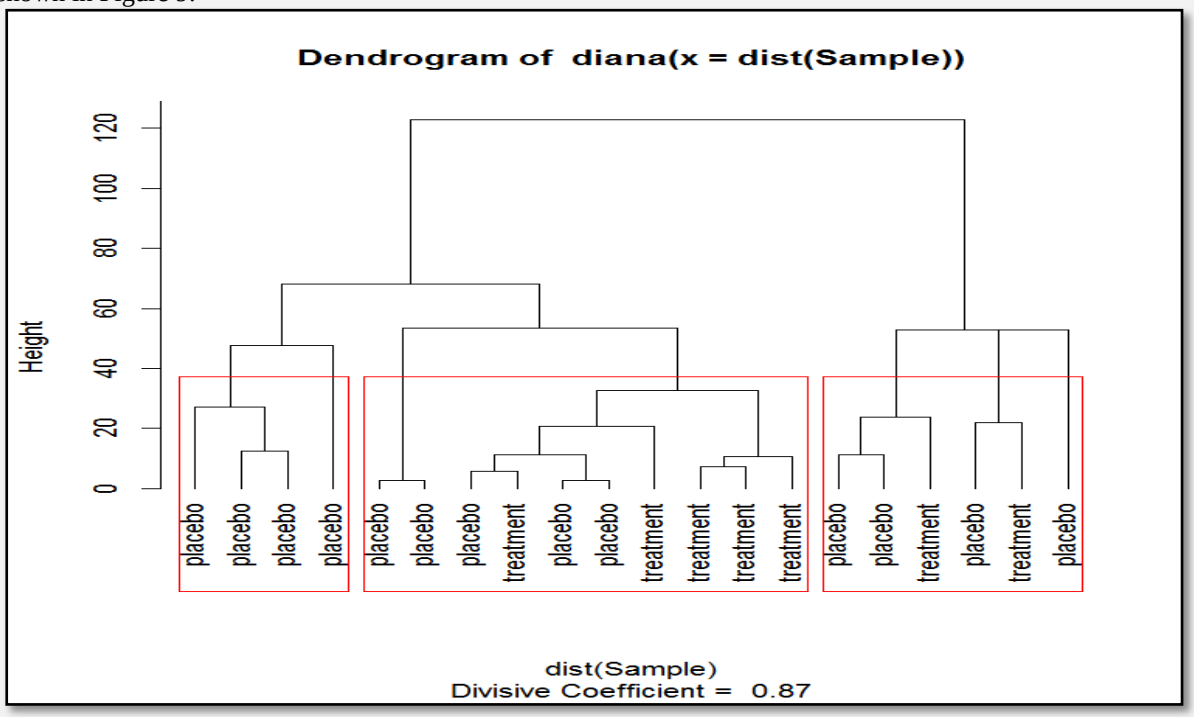


Figure 9: DIANA method, Hierarchical clustering

We can determine the optimal numbers of clusters for hierarchical clustering.

Figure 10 explains Elbow method:

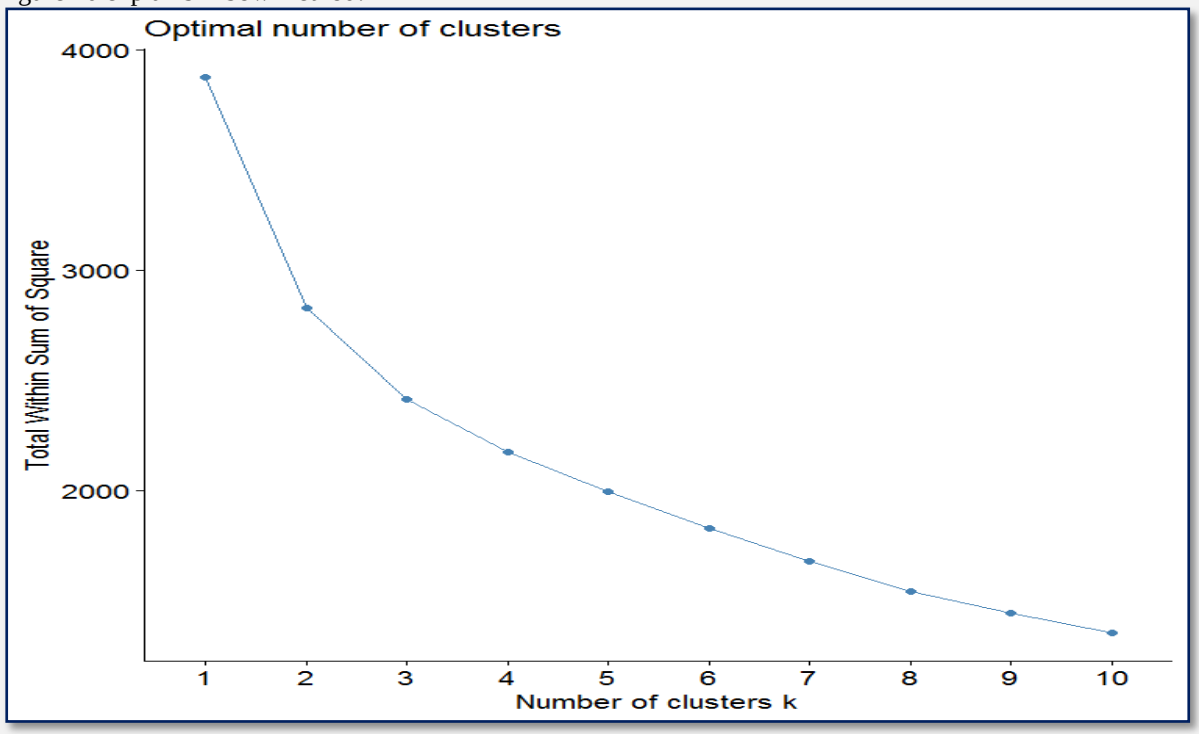


Figure 10 : Elbow method – Hierarchical clustering

From Figure 10, the total within sum of squares is minimized at $k$=10 clusters.

Figure 11 explains the Average silhouette method:

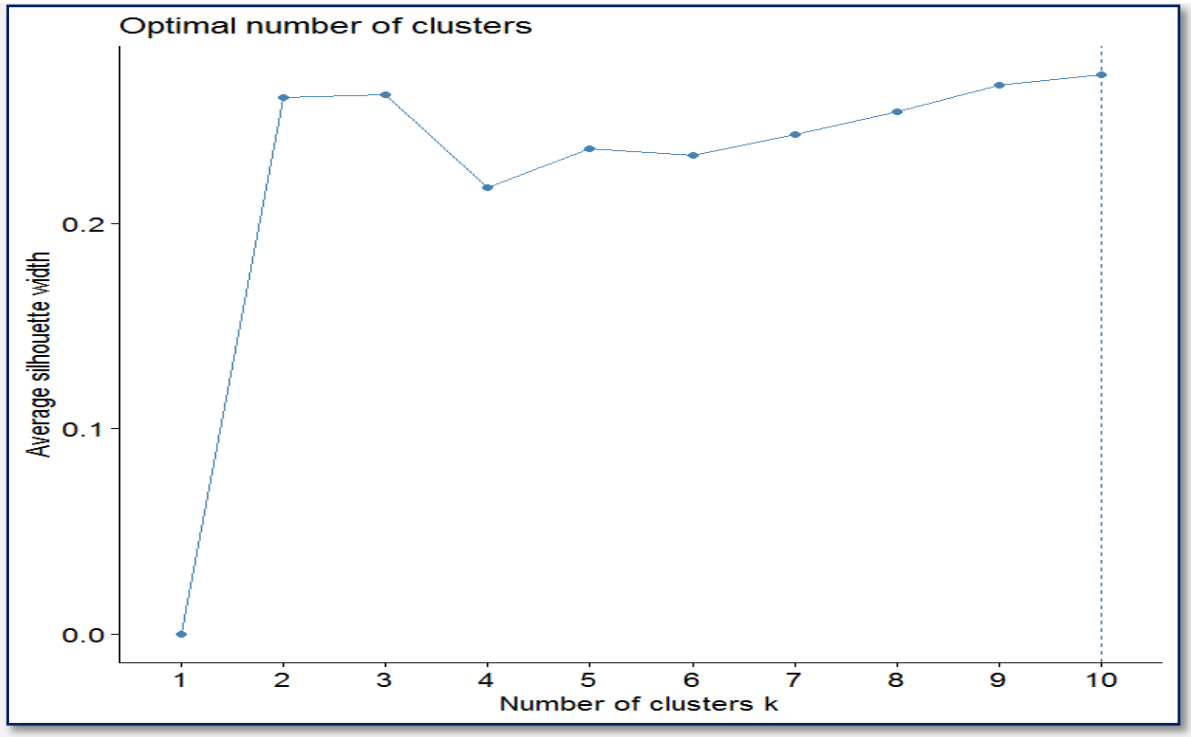


Figure 11: Silhouette method – Hierarchical clustering

From Figure 11, the average silhouette width is maximized at *k*=10 clusters.

Figure 12 explains the Gap statistic method for determining the clusters' number.

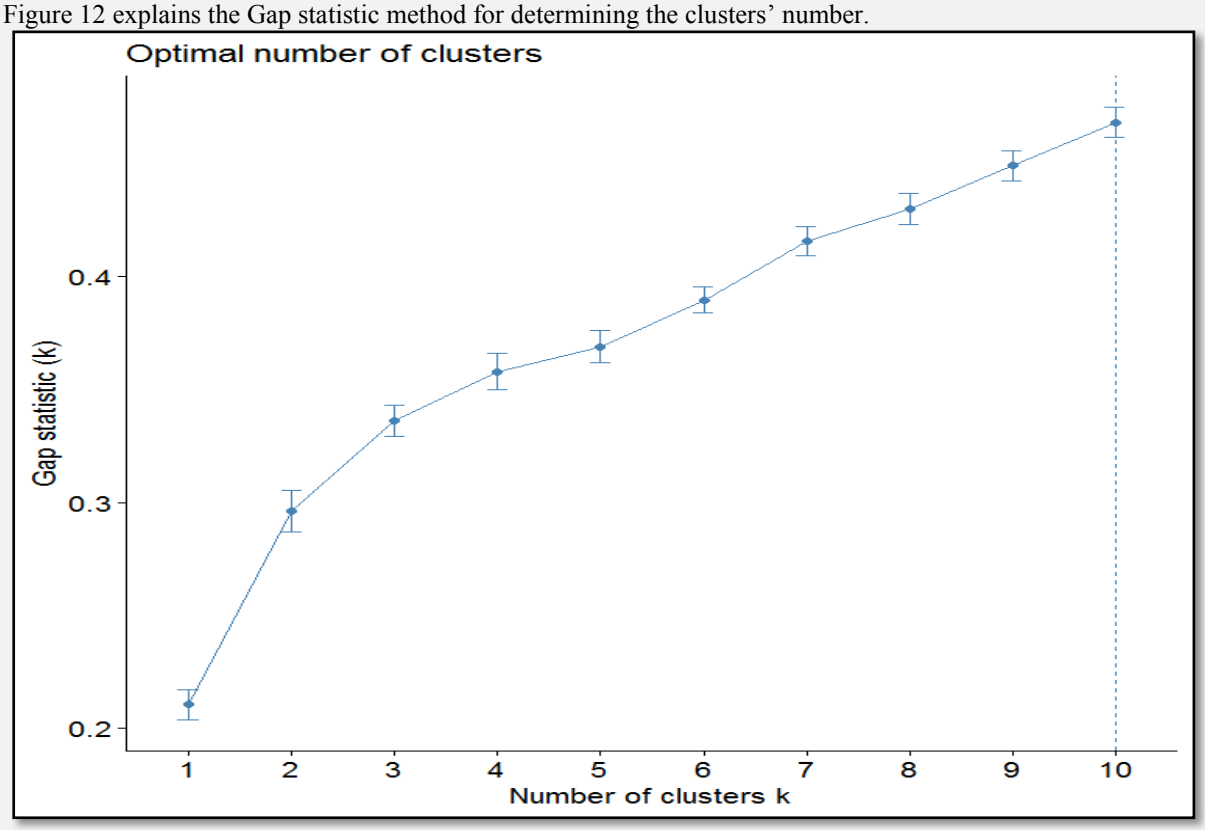


Figure 12: Gap statistic method – Hierarchical clustering

From Figure 12, Gap statistic is maximized at *k*=10 clusters.

Note that the average silhouette method has determined the optimal number of clusters at k=3, in Kmeans technique, but it is achieved at k=10, in Hierarchical technique.

**III.4 Model Based**

In this technique, the respiratory data are used after coding the categorical variables, such as treatment, gender and status variables as mentioned before.

Using the (Mclust) function, we get the next results: BIC: -14098.67, Loglik : -6938.75.

The mean of all variables are:

| Variable | Center | Age | Month | Subject | Treatment | Gender | Status |
|---|---|---|---|---|---|---|---|
| **Mean** | 1.5 | 33.28 | 3 | 56 | 1.5 | 1.21 | 1.5 |

Figures from 13 to 16 display the BIC, Classification, Uncertainty, and Density respectively:

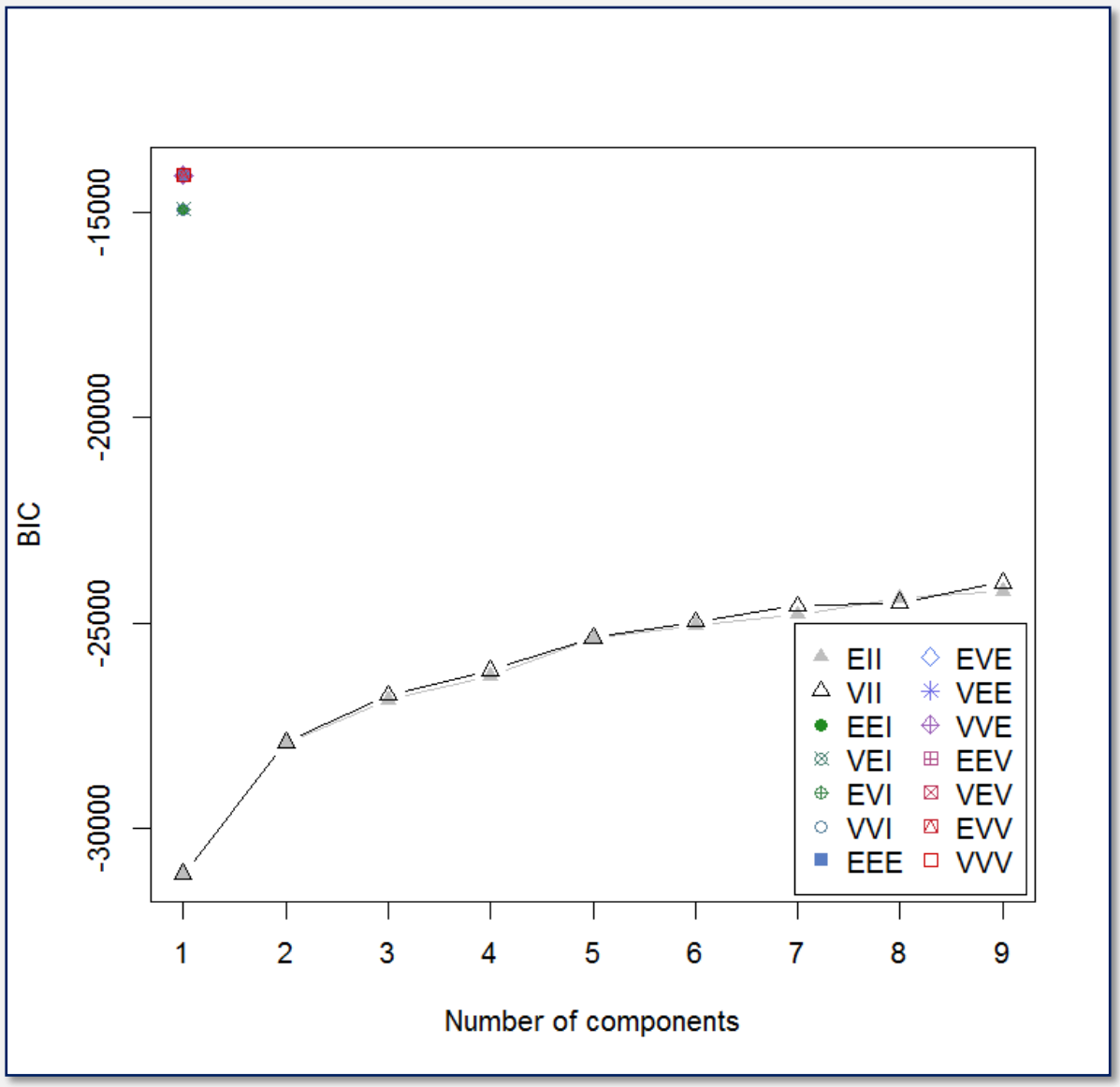


Figure 13: Mdel Based : BIC

Figure 13 explains all models with BIC measure. The best models are concentrated on the left corner with different colors at maximum value of BIC (-14098.67).

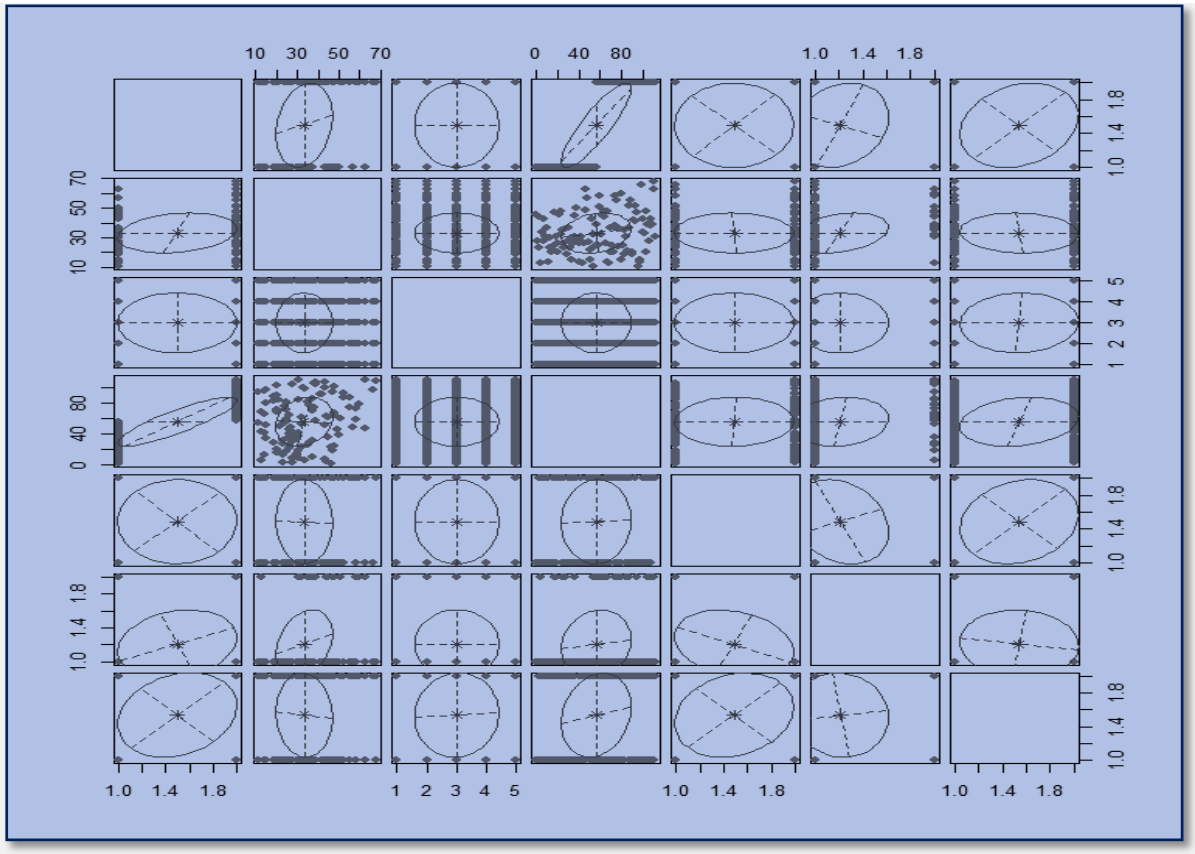


Figure 14: Model Based: Classification

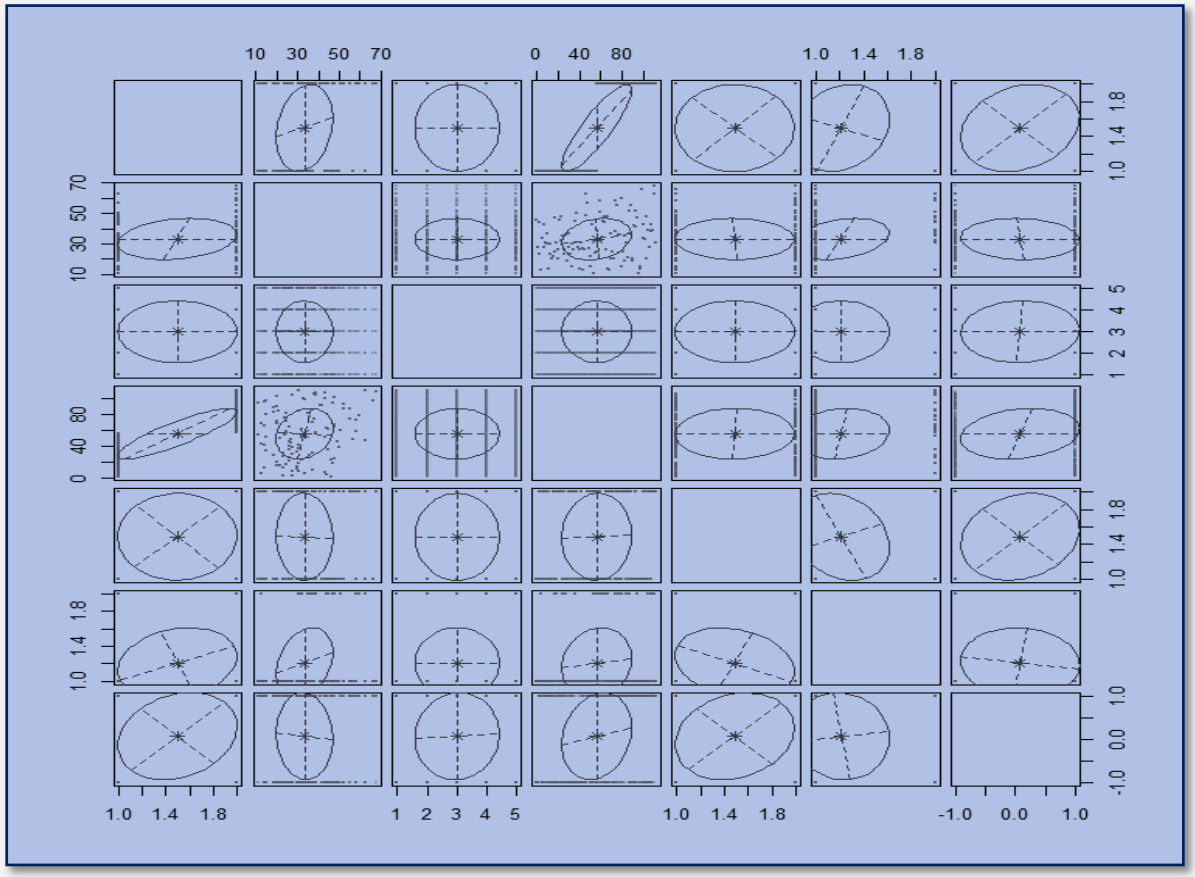


Figure 15: Model based clustering: Uncertainty

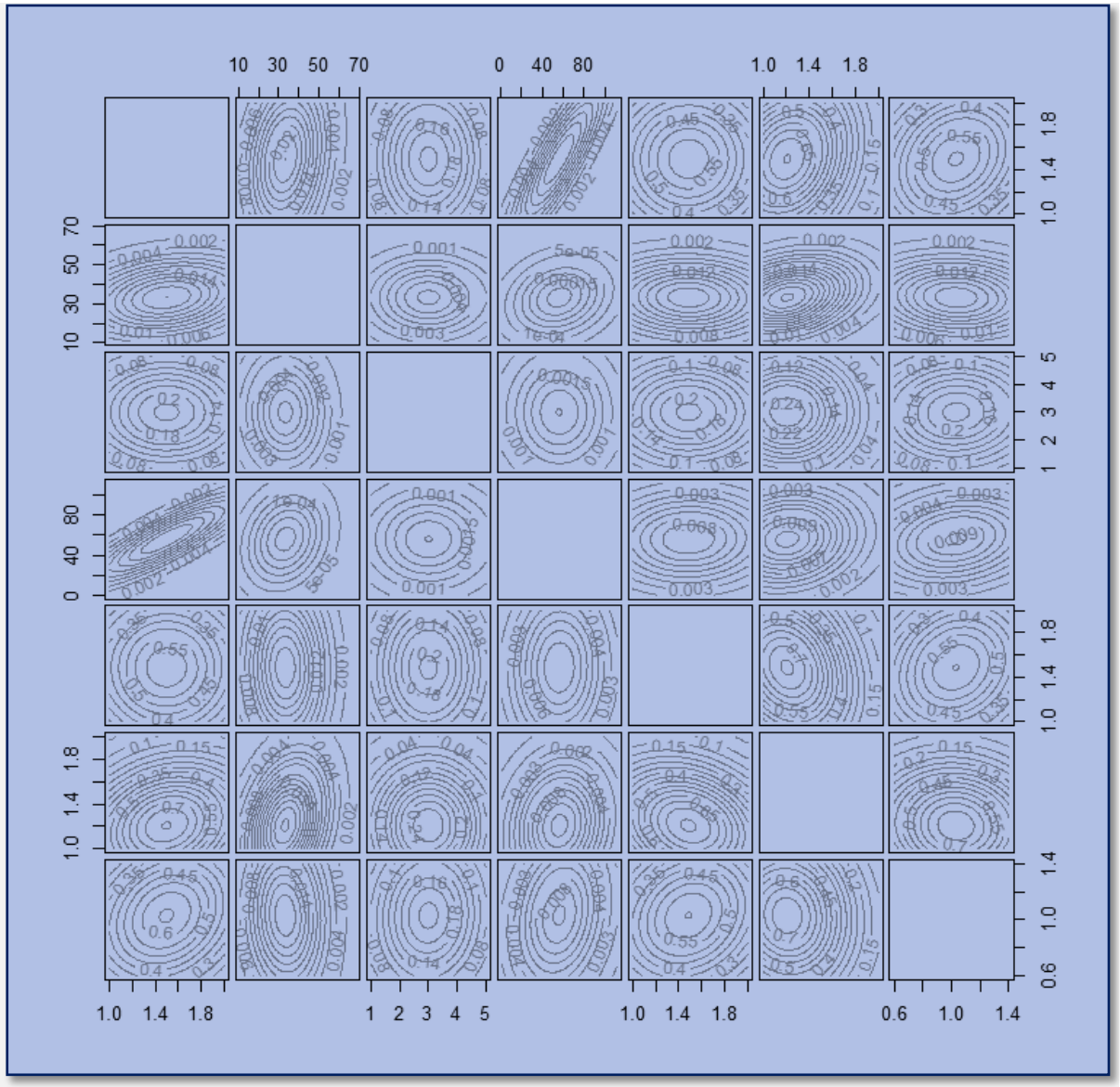


Figure 16: Model based clustering: Density

We will choose the model achieves large BIC with negative values. The 3 top models based on the BIC are:

| EEE | EEV | EVE |
|---|---|---|
| -14098.67 | -14098.67 | -14098.67 |

These models are the best model for BIC criterion.

## III.5 Enhancing Clustering

In this section, we'll use the function (eclust), also the function (fviz_cluster), for techniques: Kmeans, PAM, Hierarchical as follow:

Applying these functions on the scaled data for (Kmeans) technique we have for *k*=3 clusters:

Within cluster sum of squares for each cluster: [1028.77, 839.78, 528.71],
and the ratio (between sum of squares ÷ total sum of squares) = 38.2 %

We saw that the results of using (eclust) similar to using the function (kmeans).

Figure 17 display the enhancing clustering with *k*=3 clusters.


Figure 17: Enhancing Clustering - Kmeans

In figure 17, there are some objects did not belong to any cluster.

Applying these functions on the scaled data for (PAM) technique we have for k=3 clusters with Figure 18, an o bjective function is

| build | swap |
|---|---|
| 2.259024 | 2.242149 |


Figure 18: Enhancing Clustering - PAM

In figure 18, there are some objects don't belong to cluster3.

Applying these functions on the scaled respiratory data for (PAM) technique we have for *k*=3 clusters with Figure 18 and Figure 19:

| | |
|---|---|
| Merge | 1108 |
| Height | 554 |
| Order | 555 |

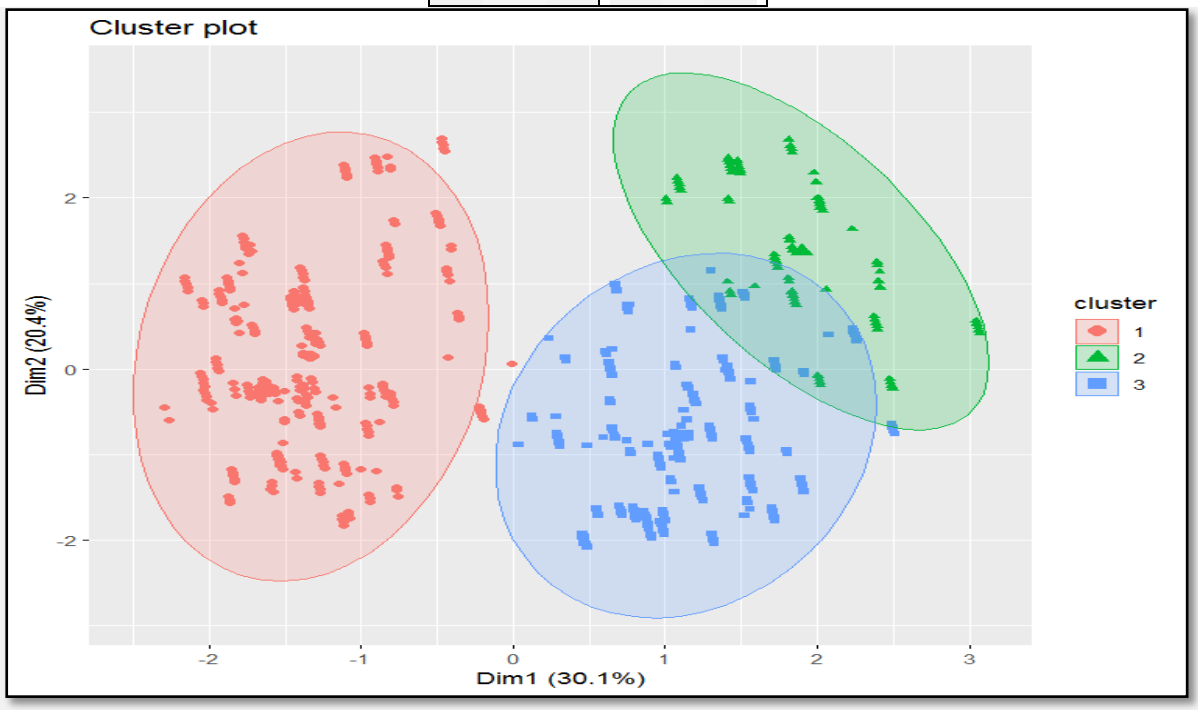


Figure 19: Enhancing Clustering – Hierarchical

In figure 19, there are some objects did not belong to cluster1.

Figure 20 displays the dendrogram with *k*=3 clusters and 555 objects.

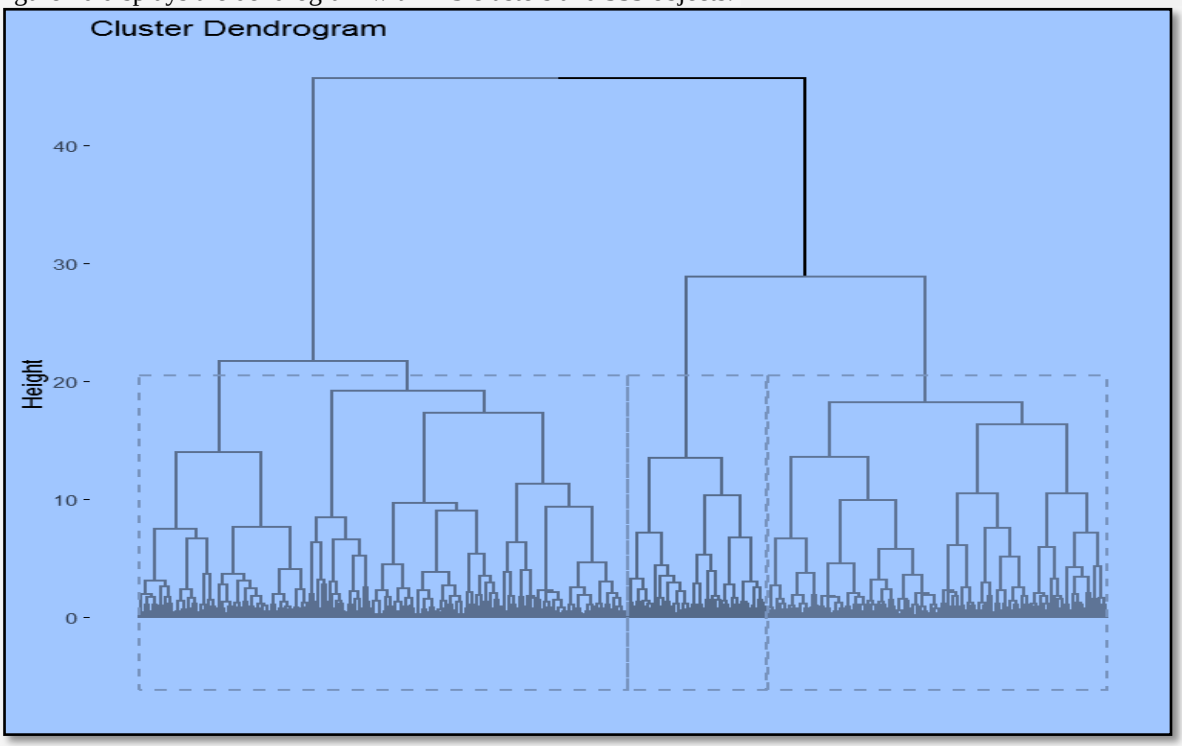


Figure 20: Enhancing Clustering - Hierarchical Dendrogram

## IV. CLUSTER VALIDATION MEASURES

In this section, we'll compute the quality of clustering for Kmeans, PAM and Hierarchical clustering. There are two indices to assess the similarity of two clustering: Rand index and Meila's VI. Rand Index has range (0) indicating that when no pair of points appear either in the same cluster or in different clusters in both clustering, (1) indicating that the two clustering are the same. VI is a non-negative measure.

Question is that: Does the clustering results similar to structure of true data? To answer this question we need to compute a cross-tabulation for Kmeans, PAM, and Hierarchical clustering techniques as follow:

| **Treatment** | **Kmeans** | | | **PAM** | | | **Hierarchical** | | |
|---|---|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **1** | **2** | **3** | **1** | **2** | **3** |
| **Placebo** | 125 | 80 | 80 | 150 | 39 | 96 | 145 | 60 | 80 |
| **Active** | 125 | 115 | 30 | 29 | 106 | 135 | 135 | 20 | 115 |

Total objects in all techniques are (555) objects, but the objects in each cluster for each technique are different. So, to compare between these techniques, we must present the cluster validation for each technique using (cluster.stats) function applying on the distance matrix for scaled data as following:

### IV.1 Kmeans

Cluster validation measures using Kmeans technique can be summarized as following:

| **Description** | **Cluster 1** | **Cluster 2** | **Cluster 3** |
|---|---|---|---|
| Cluster size | 250 | 195 | 110 |
| Diameter | 65.43699 | 69.15201 | 96.56086 |
| Average distance | 25.51138 | 27.01346 | 35.86551 |
| Median distance | 24.24871 | 26.21068 | 32.68027 |
| Separation | 2.236068 | 1.414214 | 1.414214 |
| Average to other | 56.16049 | 53.48915 | 43.39921 |
| Separation matrix | 0.000000 | 2.236068 | 3.316625 |
| | 2.236068 | 0.000000 | 1.414214 |
| | 3.316625 | 1.414214 | 0.000000 |
| Average between matrix | 0.00000 | 60.64394 | 48.21256 |
| | 60.64394 | 0.00000 | 37.22826 |
| | 48.21256 | 37.22826 | 0.00000 |
| Average between clusters | 52.00392 | | |
| Average within clusters | 27.12617 | | |
| N between | 97700 | | |
| N within | 56035 | | |
| Max diameter | 96.56086 | | |
| Min separation | 1.414214 | | |
| Within cluster sum of squares | 288729.5 | | |
| Cluster average silhouette widths | 0.4389172 | 0.2610857 | -0.1842263 |
| Average silhouette width | 0.25293 | | |
| Pearson gamma | 0.4931404 | | |
| Dunn | 0.01464583 | | |
| Dunn2 | 1.037996 | | |
| Entropy | 1.047521 | | |
| WB. ratio | 0.5216178 | | |
| Cluster high | 368.2351 | | |
| Cluster wide gap | 14.14214 | 14.89966 | 22.64950 |
| Widest gap | 22.6495 | | |
| Separation index | 2.679306 | | |

### IV.2 PAM

Cluster validation measures using PAM technique can be summarized as following:

| **Description** | **Cluster 1** | **Cluster 2** | **Cluster 3** |
|---|---|---|---|
| Cluster size | 179 | 145 | 231 |
| Diameter | 94.03723 | 61.66847 | 69.15201 |

| | | | |
|---|---|---|---|
| Average distance | 33.00715 | 23.92833 | 27.89795 |
| Median distance | 30.21589 | 22.49444 | 26.88866 |
| Separation | 1.414214 | 1.414214 | 1.414214 |
| Average to other | 45.20008 | 48.19383 | 56.92924 |
| Separation matrix | 0.000000<br>1.414214<br>1.414214 | 1.414214<br>0.000000<br>2.236068 | 1.414214<br>2.236068<br>0.000000 |
| Average between matrix | 0.00000<br>31.71754<br>53.66314 | 31.71754<br>0.00000<br>60.96118 | 53.66314<br>60.96118<br>0.00000 |
| Average between clusters | 50.43741 | | |
| Average within clusters | 28.65267 | | |
| N between | 100799 | | |
| N within | 52936 | | |
| Max diameter | 94.03723 | | |
| Min separation | 1.414214 | | |
| Within cluster sum of squares | 294270 | | |
| Cluster average silhouette widths | -0.1727330 | 0.2444332 | 0.4432958 |
| Average silhouette width | 0.1926575 | | |
| Pearson gamma | 0.4263229 | | |
| Dunn | 0.01503887 | | |
| Dunn2 | 0.9609295 | | |
| Entropy | 1.080469 | | |
| WB. ratio | 0.5680836 | | |
| Cluster high | 356.1056 | | |
| Cluster wide gap | 14.49138 | 25.51470 | 14.89966 |
| Widest gap | 25.5147 | | |
| Separation index | 1.414214 | | |

**IV.3 Hierarchical**

Cluster validation measures using Hierarchical technique can be summarized as following:

| **Description** | **Cluster 1** | **Cluster 2** | **Cluster 3** |
|---|---|---|---|
| Cluster size | 280 | 80 | 195 |
| Diameter | 65.43699 | 58.33524 | 69.15201 |
| Average distance | 25.04655 | 25.49409 | 27.01346 |
| Median distance | 23.55844 | 24.21776 | 26.21068 |
| Separation | 2.236068 | 1.414214 | 1.414214 |
| Average to other | 59.40137 | 45.60722 | 53.48915 |
| Separation matrix | 0.000000<br>6.557439<br>2.236068 | 6.557439<br>0.000000<br>1.414214 | 2.236068<br>1.414214<br>0.000000 |
| Average between matrix | 0.00000<br>56.96522<br>60.40082 | 56.96522<br>0.00000<br>29.29830 | 60.40082<br>29.29830<br>0.00000 |
| Average between clusters | 54.33001 | | |
| Average within clusters | 25.67824 | | |
| N between | 92600 | | |
| N within | 61135 | | |
| Max diameter | 69.15201 | | |
| Min separation | 1.414214 | | |
| Within cluster sum of squares | 232921 | | |
| Cluster average silhouette widths | 0.51793579 | 0.11290988 | 0.06615864 |
| Average silhouette width | 0.3008212 | | |
| Pearson gamma | 0.5775431 | | |
| Dunn | 0.0204508 | | |
| Dunn2 | 1.084581 | | |
| Entropy | 0.9918724 | | |

| | | | |
|---|---|---|---|
| WB. ratio | 0.4726346 | | |
| Cluster high | 522.5955 | | |
| Cluster wide gap | 11.53256 | 22.64950 | 14.89966 |
| Widest gap | 22.6495 | | |
| Separation index | 2.788212 | | |

An agreement between the treatment effects (Active, Placebo) and clustering solution is:

| **Measure** | **PAM** | **Hierarchical** | **Kmeans** |
|---|---|---|---|
| Rand Index | 0.8636433 | 0.4423883 | 0.4249209 |
| VI | 0.2879184 | 1.008264 | 1.127055 |

The best technique is PAM follow Hierarchical and Kmeans.
Also, for Entropy (near to 1) and WB. ratio (must be minimized), we can order them as:

| **Measure** | **Hierarchical** | **Kmeans** | **PAM** |
|---|---|---|---|
| Entropy | 0.9918724 | 1.047521 | 1.080469 |
| WB. ratio | 0.4726346 | 0.5216178 | 0.5680836 |

So, the best technique is Hierarchical, Kmeans and PAM.
Also, for Dunn (must be minimized) and Dunn2 (near to 1) measures, we can order them as:

| **Measure** | **Kmeans** | **PAM** | **Hierarchical** |
|---|---|---|---|
| Dunn | 0.01464583 | 0.01503887 | 0.0204508 |
| Dunn2 | 1.037996 | 0.9609295 | 1.084581 |

So, the best technique is Kmeans, PAM and Hierarchical.

**IV.4 Silhouette Measure**

Silhouette analysis measures a well clustered and estimates the average distance between clusters**.** The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters. In this section, we will use clustering validation silhouette technique for Kmeans, PAM and Hierarchical clustering methods respectively.

For Kmeans, the average silhouette width is (0.27) and no-negative silhouettes as shown in Figure 21.

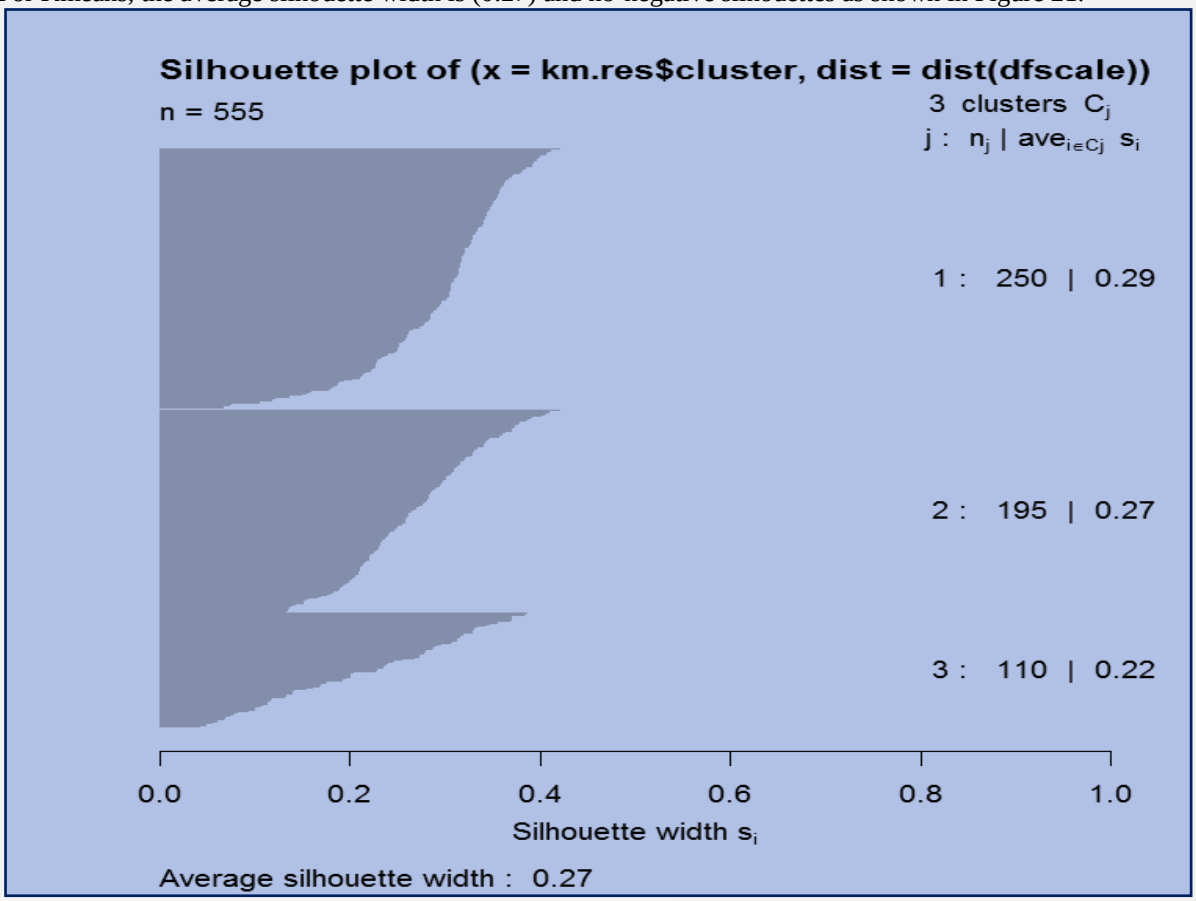


Figure 21: Silhouette measure - Kmeans

For PAM clustering technique, the average silhouette width is (0.17) and there are large negative silhouettes as shown in Figure 22.


Figure 22: Silhouette measure - PAM

3For Hierarchical clustering technique, the average silhouette width is (0.26) and there are little negative silhouettes as shown in figure 23.


Figure 23: Silhouette measure - Hierarchical

Hence, there are not negative silhouettes in Kmeans clustering technique, so it is the best technique for silhouette measure. Follow it Hierarchical technique and the PAM technique. Also, the average silhouette width ensures this conclusion (Kmeans = 0.27, Hierarchical = 0.26, PAM = 0.17 and) respectively.

**IV.5 Cluster Plot Against 1st and 2nd Principal Components**

In this section, we'll use (plotcluster) function applying on the scaled data. Figures 24, 25 and 26 display the clusters against 1st and 2nd principal components for pairs of (Kmeans, PAM), (Kmeans and Hierarchical) and (PAM and Hierarchical) clusters respectively as shown below:

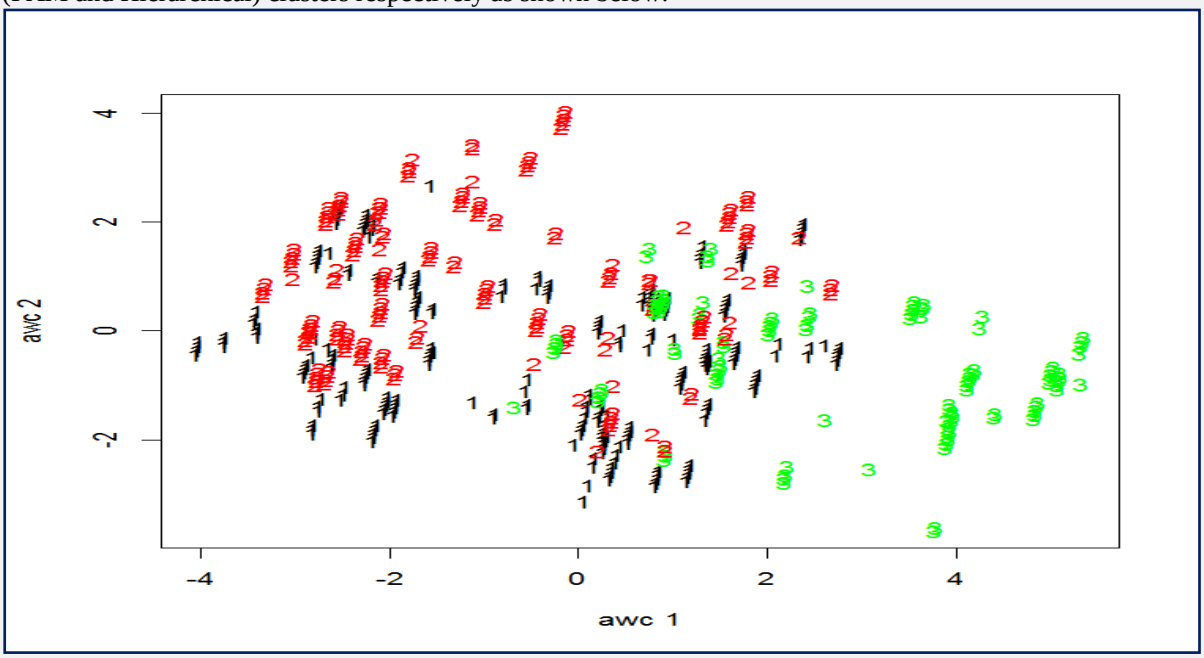


Figure 24: Plot cluster, (Kmeans and PAM)

We see that in Figure 24 overlap between all clusters

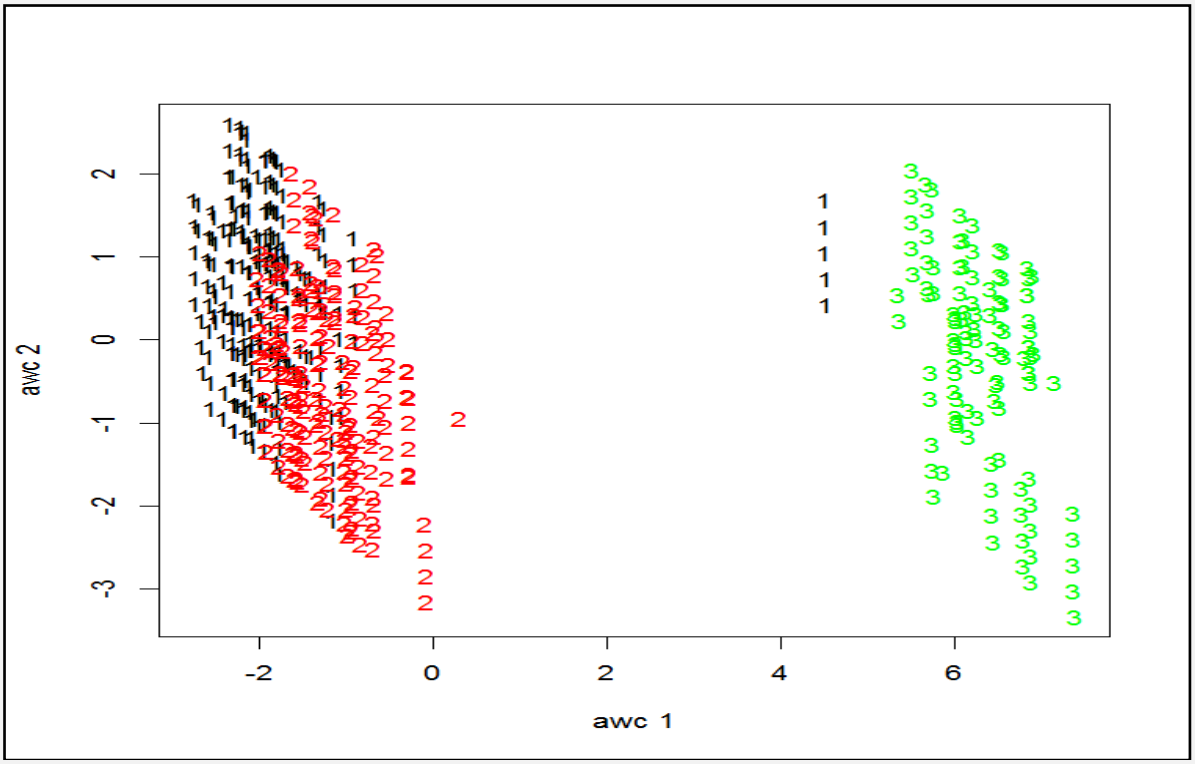


Figure 25: Plot cluster – (Kmeans and Hierarchical)

In Figure 25, there are overlap between objects in clusters1, cluster2 and stay in the left corner but cluster3 is separated in the right corner.

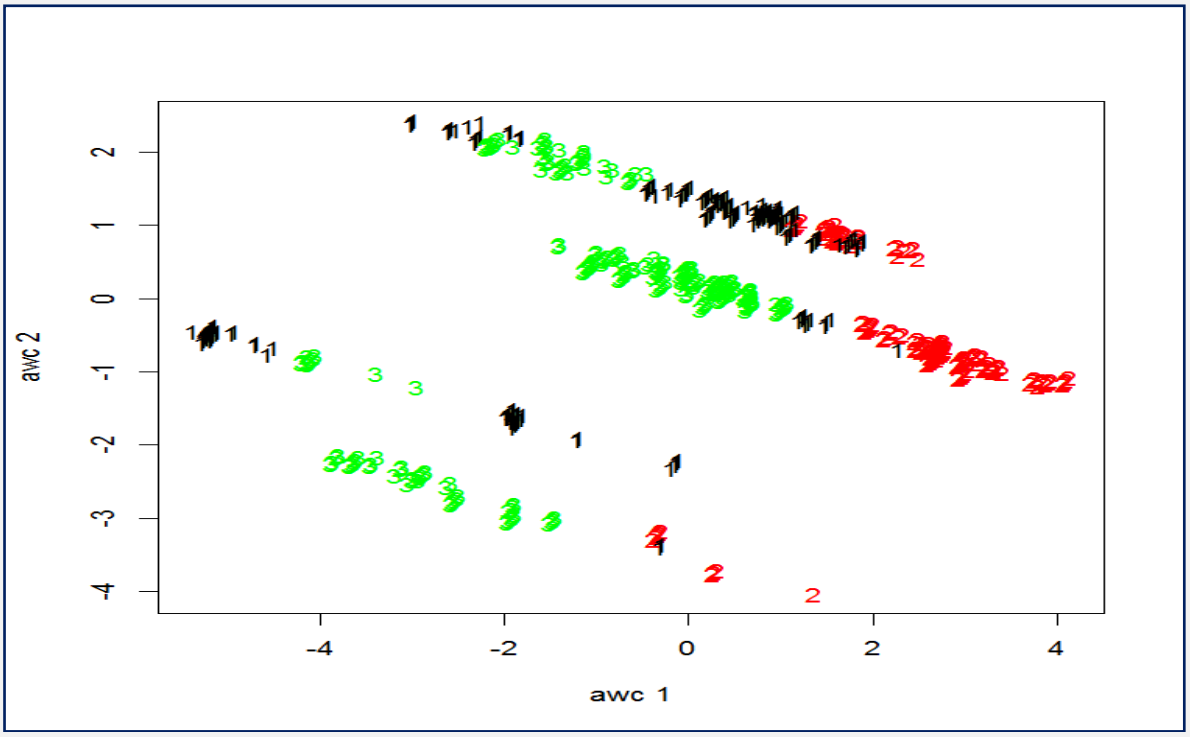


Figure 26: Plot cluster – (PAM and Hierarchical)

Finally, in Figure 26, there are overlap between objects in all clusters but all clusters are separated randomly.

So, the plot clusters between (Kmeans and Hierarchical) techniques is the best one to cluster data hence indicates a well separated cluster.

## V. DISCUSSION AND CONCLUSIONS

In this paper, we have explained the differences between different techniques that are used in clustering process. These techniques are Kmeans, Kmedoids (PAM), Hierarchical and Model based. For purposes of comparison, we explained, for each technique, the cluster size, the optimal number of clusters, if it was possible, graphically using some measures, some of cluster validation measures to assess the quality of clustering process. Finally, we have displayed plot clusters for each pair of these techniques against the first and the second principal components. There are some differences between these techniques:

In Kmeans technique, the ratio (between sum of squares / total sum of squares) is increased as number of clusters *k* increase. This due to increase between sum of squares. The total within sum of squares decrease as *k* increases.

In Kmedoids (PAM) technique, there is no difference between (pamk) and (pam) functions except in specify the number of *k* clusters in (pam) function, and the obtained results are similar.

For different linkages, they can be arranged, from large value to small value of (AC), as follow: Ward , Complete, Average and Single. This indicates the best linkage is Ward method.

The best technique for the average silhouette method is Kmeans technique. Hence, there are no-negative silhouette. Also, the average silhouette width in Kmeans is the largest. The obtained results of using (eclust) function similar to the (kmeans) function.

Rand Index and VI measures, indicate the agreement between the treatment effects (Active, Placebo) and clust ering solution, and the best technique is PAM follow Hierarchical and Kmeans.

Entropy (near to 1) and WB. ratio (must be minimized) measures, indicate that the best technique is Hierarchical follow Kmeans and PAM.

Dunn (must be minimized) and Dunn2 (near to1) measures, indicate that the best technique is Kmeans follow PAM and Hierarchical.

We conclude there are no technique is better than another one for all measures. Since Kmeans is better for one measure, and PAM is better for another measure and so on.

## REFERENCES


[1]. Calinski, T., and Harabasz, J. (1974). A Dendrite method for cluster analysis. Communications in Statistics 3, pp.1-27.

[2]. Davis, C. S. (1991). Semi-parametric and non-parametric methods for the analysis of repeated measurements with applications to clinical trials. Statistics in Medicine 10, pp.1959–1980.

[3]. Everitt, B.S. and Hothorn, T. A. Handbook of statistical analyses using R (2nd ed.).

[4]. Everitt, B. S., Landau, S., and Leese, M. (2001). Cluster analysis (4th ed.). Arnold. (1002) Jun, S., and Uhm, D. (2010). Patent and statistics, What's the connection? Communications of the Korean Statistical Society 17, 2, pp.205–222.

[5]. Gordon, A. D. (1999). Classification, 2nd ed. Chapman and Hall.

[6]. Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2001). On clustering validation techniques. Journal of Intelligent Information Systems 17, pp.107-145.

[7]. Halkidi, M., Vazirgiannis, M. and Hennig, C. (2015). Method-independent indices for cluster validation. In C. Hennig, M. Meila, F. Murtagh, R. Rocci (eds.) Handbook of Cluster Analysis, CRC Press/Taylor & Francis, Boca Raton.

[8]. Han, J., and Kamber, M. (2006). Data mining concepts and techniques (2nd ed.). United States of America: Morgan Kaufman Publishers.

[9]. Hennig, C. (2015). fpc: Flexible procedures for clustering. R package version 2.1.

[10]. Jain, A., Murty, M. N., Flynn, P. J. (1999). Data clustering: a review. ACM Comput. Surv. 31,3, pp.264–323.

[11]. Kaufman, L. and Rousseeuw, P.J. (1990). Finding groups in data: An introduction to cluster analysis. Wiley, New York.

[12]. Kwedlo, W. (2011). A clustering method combining differential evolution with the k-means algorithm. Pattern Recognition Letters 32, pp.1613–1621.

[13]. Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K (2015). Cluster: Cluster analysis basics and extensions. R package version 2.0.

[14]. Malika C., Nadia G., Veronique B. and Azam N. (2014). NbClust: An R package for determining the relevant number of clusters in a data Set. Journal of Statistical Software 61, 6, pp.1-36.

[15]. Meila, M. (2007). Comparing clusterings: an information based distance. Journal of Multivariate Analysis 98, pp.873-895.

[16]. Milligan, G. W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters. Psychometrika 50, pp.159-179.

[17]. Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association. American Statistical Association 66 ,336,pp. 846–850.

[18]. Roiger, R. J., Geatz, M. W. (2003). Data mining a tutorial – based primer. Pearson Education, Inc. Addison Wesley. pp. 11-12.

[19]. Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. Vol.20, pp.53-65.

[20]. Shahbaba, M., Beheshti, S. (2014). MACE-means clustering. Singnal Processing. Vol.105, pp.216-225.

[21]. Theodoridis, S., Koutroubas, K. (2008). Pattern recognition. 4th edition. Academic Press.

[22]. Tippaya T., Nuntawut K., Pongsakorn D., Kittisak K. and Nittaya K. (2015). The clustering validity with silhouette and sum of squared errors. Proceedings of the 3rd International Conference on Industrial Application Engineering. Institute of Industrial Applications Engineers, Japan.

[23]. Wang, L., Leckie, C., Ramamohanarao, K., and Bezdek, J. (2009). Automatically determining the number of clusters in unlabeled data sets. IEEE Transactions on Knowledge and Data Engineering 21,3, pp.335–350.

[24]. Hierarchical Cluster Analysis· UC Business Analytics R Programming Guide. https://uc-r.github.io/hc_clustering


Ahmed Mohamed Mohamed Elsayed" Comparison between Cluster Techniques for Clinical Data" International Journal of Mathematics and Statistics Invention (IJMSI), vol. 07, no. 01, 2019, pp. 12-33