Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Classification of Cancers Diseases using Multivariate Analysis

Asia Hammood Hussein, Nabaa Naeem Mahdi, Haifa Taha Abd
University of Mustansiriyha / collage of Management and Economics
asiaalsaleem@uomustansiriyah.edu.iq
nabaanaeemmahdi@uomustansiriyah.edu.iq
haefaa_adm@uomustansiriyah.edu.iq

Article Info
Page Number: 4888 - 4898
Publication Issue:
Vol 71 No. 4 (2022)

Abstract

Cancer is a serious disease in Iraq due to the pollution of the environment as a result of the wars that Iraq fought. It turns out that there is no interest in studying cancer tumors as the mother of the disease, while we find that there is little interest in studying the causes of environmental pollution.

Due to the importance of the topic, some types of cancerous diseases spread in each Iraqi governorate were studied through an analytical study using hierarchical cluster analysis methods. Data were collected on some cancerous diseases, including 901 cases infection with one of the cancerous tumors, and from the results of the analysis of hierarchical methods (The averages method between the groups, Single Linkage , The Complete Linkage , Average Linkage , Centroid Linkage, Midean linkage and the Ward method), which were similar in terms of convergence between the provinces for the similarity of the infection conditions with the difference in the numerical amount of coalescence only. It was found that the provinces of Babil, DhiQarudiala, and Najaf were close in terms of tumor incidence with the province of Erbil, while the province of Anbar converged with the province of Salah al-Din As for the governorates of Muthanna and Dohuk, Karbala, Wasit, Kirkuk, Sulaymaniyah, Maysan and Diwaniyah, they were associated with each other and contracted with the previous groups, then they contracted with the Nineveh Governorate and Basra. As for the governorate of Baghdad, we find it coalescing in the last stage with all the governorates.

Key Words: cluster analysis, Hierarchical Method, Clustering

Article History
Article Received*: 25 March 2022*
Revised*: 30 April 2022*
Accepted*: 15 June 2022*
Publication*: 19 August 2022*

1- Introduction:

Good for man to seek reassurance that leave himself he does not know that there is a disease that is eating away at his body, and if prevention is better than cure, the early detection of cancer is the best way to get rid of this chronic problem if you do not get away and prevent possible scenarios.

The research aims at an analytical study of the reality of cancer in Iraq, based on the methods of cluster analysis of the hierarchical clustering in order to find out the similarities between the provinces in terms of incidence of species cancer is (breast cancer, brain,

bronchitis, lung, bladder, blood, Alkaline and rectum, lymphatic, skin, stomach, thyroid, liver, pancreas, prostate, ovarian and Hodgkin's lymphoma and Hodgkin's non-lymphocytic) Attempting to know the governorates most affected by types of cancer by classifying them into homogeneous groups**.**

2-Previous studies:

There are many studies dealing with the use of clustering analysis including in 2008, (Fazzo& et al) to study through the application of cluster analysis on the analysis of mortality and distortions in the province of Naples, Italy[4].

In 2011 Veronicarosa study, Robert Vida, Franzescuvalun at the University of SPIE Lanza, Rome, Italy, the title of the study "A cluster analysis of High School Students Styles of" Living - Together "in the classroom" The study aims to identify homogeneous groups of students who have perceptions different about living together in the classroom[8].

In 2012 both (Muntauer& et al) study used a cluster analysis to classify sequential middle-income countries and low-income, according to the regulations and the dependence of the labor market discrimination in health indicators of population[7].

In the same year he Franescustudy "Poor mental health symptoms among Romanian employees. Acluster analysis" The study in Rome, Italy, the University of Bucharest and aims to identify and classify mental health problems suffered by the Romans staff at work[3].

3- The theoretical side

3-1 The concept cluster analysis:

Cluster analysis is one of the multivariate methods and is one of the methods of classification, this analysis contributes to the classification of data into clusters and clusters characterized by this being a great degree of similarity or internal homogeneity among them but they are not similar among them. Cluster analysis is also considered a useful and effective tool to analyze data in different ways [1].

There are several methods used in cluster analysis including hierarchical cluster analysis and non-hierarchical method such as the method averages (K-means).

3-2 Clustering Steps:

1- Calculate the distance matrix, the Correlation matrix, or the similarity matrix.
2- Linking the two elements between which the shortest distance is within the matrix calculated in the previous step, and if there are equal distances, the binding process can be performed for more than two elements at one stage (for two elements together).
3- The new distance matrix is calculated, taking into account the changes that occurred in the second step.

4- The binding process is continued until the cluster tree is reached **[2]**.

3-3 Hierarchical Clustering Methods:

There are many methods of cluster analysis and each method has certain characteristics that are available from other methods some adopt the method of aggregation and the other adopts the method of fragmentation, and we will deal with the most commonly used methods which are methods of cluster -sequential analysis, these methods do not require prior knowledge of the number of clusters on which cases will be classified and are suitable for relatively small samples[5], and in cluster analysis there are different methods and we will focus on methods of chain integration, especially the so-called These methods are suitable for compiling cases as well as compiling variables, which is not achieved for other assembly methods:

1-Single Linkage Clustering:

It is considered one of the simplest methods of clustering and called (the nearest neighborhood), this method depends mainly on the consideration that the two elements most similar between the elements form the nucleus of the cluster, and then the rest of the units are added to this nucleus in sequence and according to the degree of similarity with the elements of the nucleus of the cluster where the most semi- and then theleast Gradually, if a group of clusters are linked together, this is based on the lowest distances or coefficients of symmetry between the pairs of elements and according to the following formula[9]:

$$D = (A,B) = Min(d(y_i y_j) \quad for\, y_i in\ A \& y_j in\ B \qquad \text{KK}(1)$$

Where $d(y_i, y_j)$ represents the Euclidean distance calculated according to the formula:

$$d(x,y) = \sum_{i=1}^{p} |x_i - y_i|\, r \qquad \ldots(2)$$

2- Complete Linkage Clustering:

It called as well as the farthest neighbor method is based this mainly on the grounds that most elements similarity between elements form the nucleus of the cluster, and then working this way is completely counter to the principle of the work method single linkage clustering, and in this method you know the distance between cluster A and B as a greater distance between point A and B according to another in the following[6]:

$$D = (A,B) = Max\,(d(y_i y_j) \quad for\, y_i in\ A \& y_j in\ B \qquad \ldots(3)$$

3- Average Linkage:

In this way the distance between the clusters such as A and B shall be the rate for the distance to nA, nB between the points nA in points A and nB in B accordance with the following formula [9]:

$$D(A,B) = \frac{1}{nA\,nB}\sum_{i=1}^{nA}\sum_{i=1}^{nB} d(y_i, y_j) \qquad \cdots(4)$$

Where the total taken for each y_i in A each y_i in B at every step the clusters is linked to the adoption of the smaller distance.

4- Centroid:

Known as the central link method as Euclidean distance between the vector mean of clusters A and B as follows:

$$D(A,B) = d(\bar{y}_A, \bar{y}_B) \qquad \ldots(5)$$

Where $\bar{y}_A$ is a vector mean of observations in vector A and $\bar{y}_B$ is a vector mean of observations in the vector B

then clustering by integrating all clusters of two smaller distance between the centers in each step, and after connecting clusters such as A and B , which reflects the new cluster center is center-aligned AB weighted according to the following formula[5]:

$$\bar{y}_{AB} = \frac{n_A\bar{y}_A + n_B\bar{y}_B}{n_A + n_B} \qquad \ldots(6)$$

5- Median:

If the number of vocabulary of one cluster is larger than the other we use this method and in this case when using the centroid method, the new cluster center tends to cluster with the largest vocabulary and to avoid this problem we need to use the medium instead of the mean weighted to calculate the center of cluster new according to the following formula:

$$m_{AB} = \frac{1}{2}(\bar{y}_A + \bar{y}_B) \qquad \ldots(7)$$

According to this formula, any two clusters with the smallest distance between their two mediums are linked at each stage [6].

6- Ward's Method:

The hierarchical method is based on the least loss of information for the work of the cluster, and is also called the sum of the squares added method and depends on the use of the space square within each cluster and the square of spaces between clusters, which are expressed in the following formulas on the assumption AB that is the cluster resulting from linking clusters $A\&B$ as follows[2]:

$$SEE_A = \sum_{i=1}^{nA} (y_i - \bar{y}_A)' (y_i - \bar{y}_A) \quad \ldots(8)$$

$$SEE_B = \sum_{i=1}^{nB} (y_i - \bar{y}_B)' (y_i - \bar{y}_B) \quad \ldots(9)$$

$$SEE_{AB} = \sum_{i=}^{nAB} (y_i - \bar{y}_{AB}) (y_i - \bar{y}_{AB}) \quad \ldots(10)$$

Any two clusters are linked so that it reduces the increase in the distances square (SSE) and expresses the amount of that increase as it comes:

$$I_{AB} = SEE_{AB} - (SEE_A - SSB_B) \quad \ldots(11)$$

3-4 Measuring the similarities and differences:

There are several methods used to measure the similarity between each pair of There are several methods used to measure the similarity between each pair of observations and that the appropriate measure of convergence is the distance between two observations and this distance is a measure of the spacing and the difference is calculated Euclidean distance between the two vectors in accordance with the following formula [9]:

$$D(x, y) = \sqrt{(x-y)'(x-y)} \quad \ldots(12)$$

To regulate the difference between common variation and contrast to p of the variables are according to the following relationship:

$$D(x, y) = \sqrt{(x-y)' s^{-1} (x-y)} \quad \ldots(13)$$

4- Practical side

This research includes a study of the reality of cancer in the Iraqi provinces and knowledge of the most Iraqi provinces affected by these diseases, the areas of concentration of these types in each Iraqi province based on methods of analysis of pyramidal and non-hierarchical cluster

4-1 Description of the research sample

The data was based on 901 cases of cancer types for the year 2011 from the Ministry of Health / Cancer Council scattered in all Iraqi provinces and to be able to compare among

them was based on the standard values of these data as these data are not homogeneous, where variables represent x's most types of cancer diseases Deployed in all Iraqi provinces.

Variable	Descriptive	Variable	descriptive
X_1	Breast Cancer	X_{10}	thyroid
X_2	Brain	X_{11}	kidney
X_3	bronchus and lung	X_{12}	Liver
X_4	Bladder	X_{13}	prostate
X_5	Blood (leukemia)	X_{14}	Hodgkin lymphatic
X_6	Colorectal	X_{15}	non-lymphatic Hodgkin
X_7	Ovary	X_{16}	mucous membranes
X_8	Skin	X_{17}	Pancreas
X9	Stomach		

The variables Y's represent (18) Iraqi provinces are:

1- Baghdad, 2- Nineveh, 3- Kirkuk, 4- Anbar, 5- Salah al-Din, 6- Diyala, 7- Karbala, 8- Najaf, 9- Babylon, 10- Wasit, 11- Maysan, 12- Diwaniya, 13 - Dhiqar, 14 - Basra, 15 - Muthanna, 16 - Sulaymaniyah, 17 - Dohuk, 18 - Erbil

4-2The averages method between the groups:

The results of the analysis represented in table (1) and planned for this method showed that the levels of fusion between (3.606) and (72.525) From these levels shows that the provinces of Babil, , Dhiqar, Diyala and Najaf are close to the governorate of Arbil (8.022), and we note that the province of Anbar and Salah al-Din were linked together with the level of fusion (7.483), while the provinces of Muthanna, Dohuk, Karbala, Wasit, Kirkuk, Sulaymaniyah, MissanDiwaniya was connected with each other with the level of fusion (11.496) and contracted with the previous groupsWe find the convergence of Nineveh and Basra governoratesat the level of fusion (27.586) and then contracted with the province of Nineveh level of fusion (35.469) The Baghdad province we find it at the last stage with all provinces with the level of fusion (72.525), Figure 1 shows the docking graphically.

Table (1) shows the levels of docking averages between groups

Stage	Cluster Combined		Coefficients
	Cluster 1	Cluster 2	
1	**9**	**13**	**3.606**
2	**3**	**16**	**4.796**
3	**11**	**1 2**	**5.385**
4	**7**	**10**	**5.657**
5	**6**	**8**	**5.745**
6	**3**	**11**	**6.410**
7	**6**	**9**	**6.981**
8	**4**	**5**	**7.483**

9	**15**	**17**	**7.746**
10	**6**	**18**	**8.022**
11	**3**	**7**	**9.169**
12	**3**	**15**	**11.496**
13	**4**	**6**	**13.503**
14	**3**	**4**	**14.256**
15	**2**	**14**	**27.586**
16	**2**	**3**	**36.469**
17	**1**	**2**	**72.525**

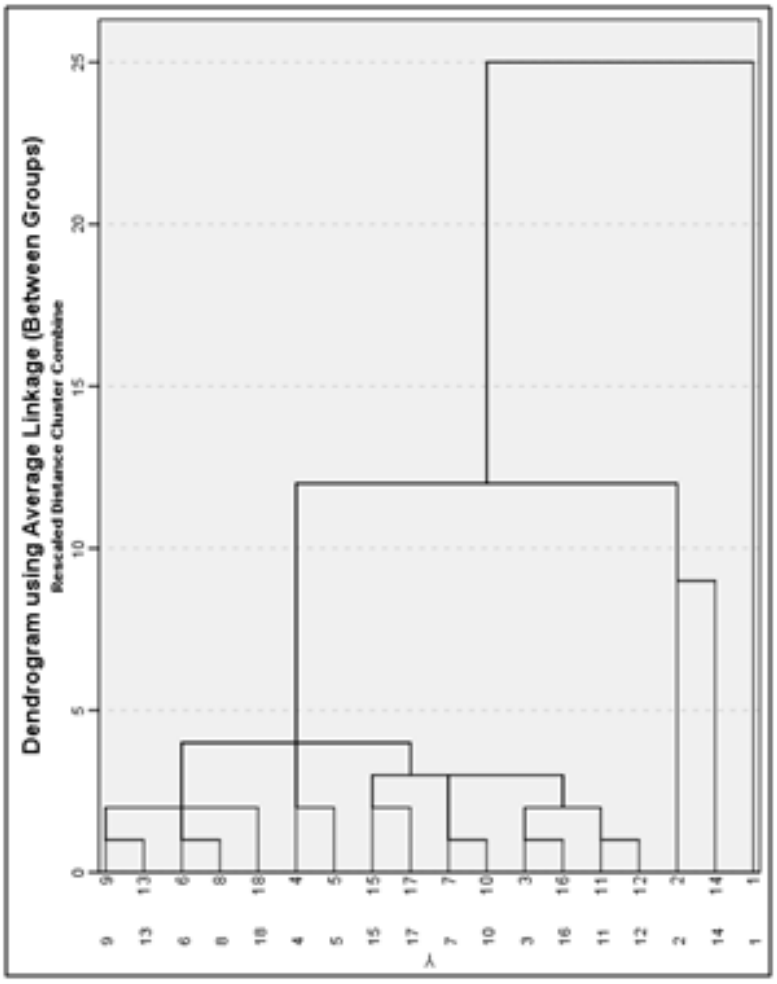

Figure 1: show the level of fusion

From the results of the analysis of the hierarchical methods (single linking, universal linking, linkage to the central mediator and Ward method) we find that despite the different levels of adhesion of the methods mentioned, but they lead to the same results reached in the method of averages between groups as shown in tables (2), (3),(4), (5), (6) and Figure 2,3,4,5,6 respectively.

Table (2) shows the levels of docking single linking between groups

Stage	Cluster Combined		Coefficients
	Cluster 1	Cluster 2	
1	9	13	3.606
2	3	16	4.796
3	11	12	5.385
4	7	10	5.657
5	3	11	5.745
6	6	8	5.745
7	3	15	6.083
8	6	9	6.245
9	3	7	6.245
10	3	5	6.481
11	6	18	6.557
12	3	6	6.633
13	3	4	7.483
14	3	17	7.746
15	3	14	19.053
16	2	3	27.586
17	1	2	35.185

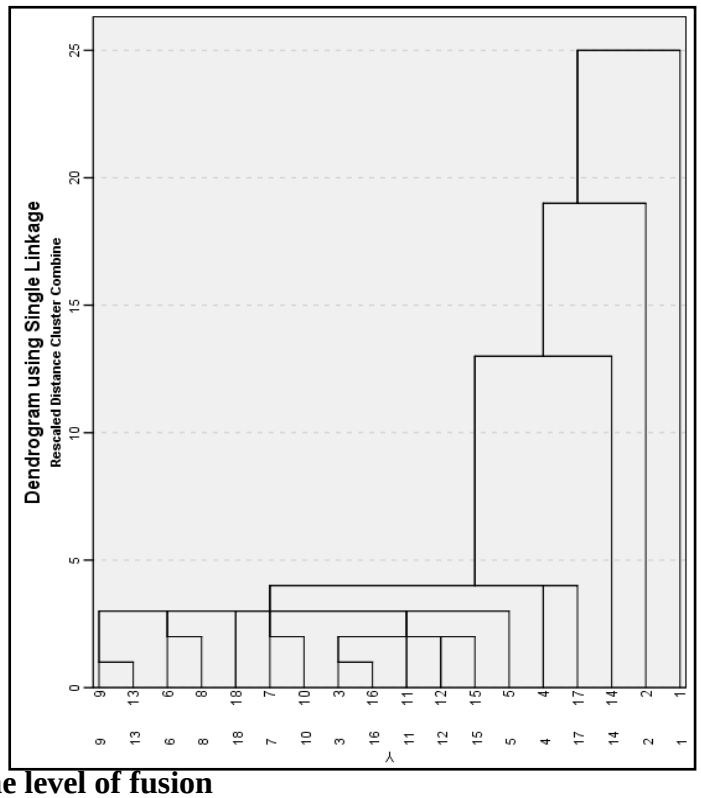

Figure2: show the level of fusion

Table (3) shows the levels of docking universal linking between groups

Stage	Cluster Combined		Coefficients
	Cluster 1	Cluster 2	
1	**9**	**13**	**3.606**
2	**3**	**16**	**4.796**
3	**11**	**12**	**5.385**
4	**7**	**10**	**5.657**
5	**6**	**8**	**5.745**
6	**3**	**11**	**7.416**
7	**4**	**5**	**7.483**
8	**6**	**9**	**7.616**
9	**15**	**17**	**7.746**
10	**6**	**18**	**9.592**
11	**3**	**7**	**11.790**
12	**3**	**15**	**15.264**
13	**4**	**6**	**18.138**
14	**3**	**4**	**23.749**
15	**2**	**14**	**27.586**
16	**1**	**2**	**57.018**
17	**1**	**3**	**88.944**

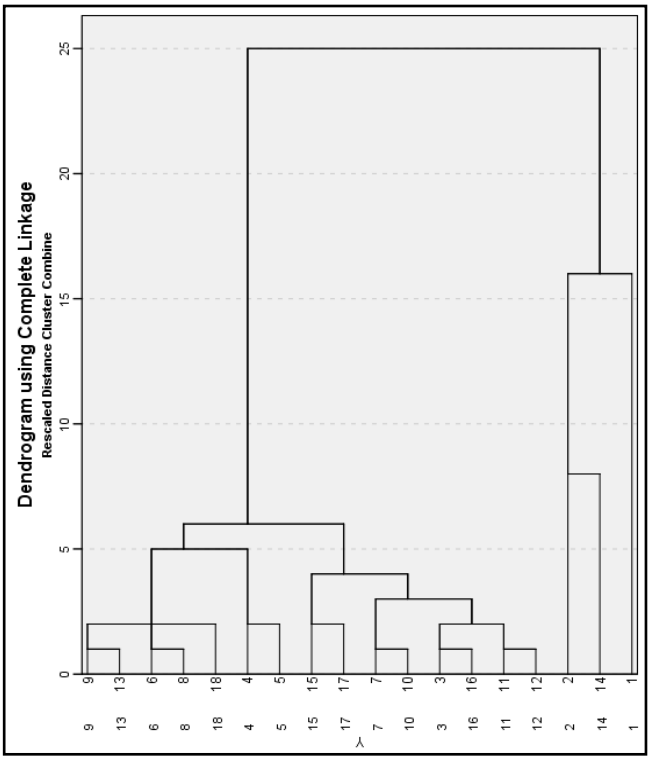

Figure3: show the level of fusion

Table (4) shows the levels of dockinglinkage to the central between groups

Stage	Cluster Combined		Coefficients
	Cluster 1	Cluster 2	
1	9	13	3.606
2	3	16	4.796
3	3	12	4.673
4	3	11	4.590
5	6	9	5.612
6	6	8	5.032
7	7	10	5.657
8	3	7	5.516
9	6	18	5.692
10	3	6	6.246
11	3	5	6.986
12	15	17	7.746
13	3	15	8.734
14	3	4	10.097
15	3	14	22.870
16	1	2	35.185
17	1	3	43.682

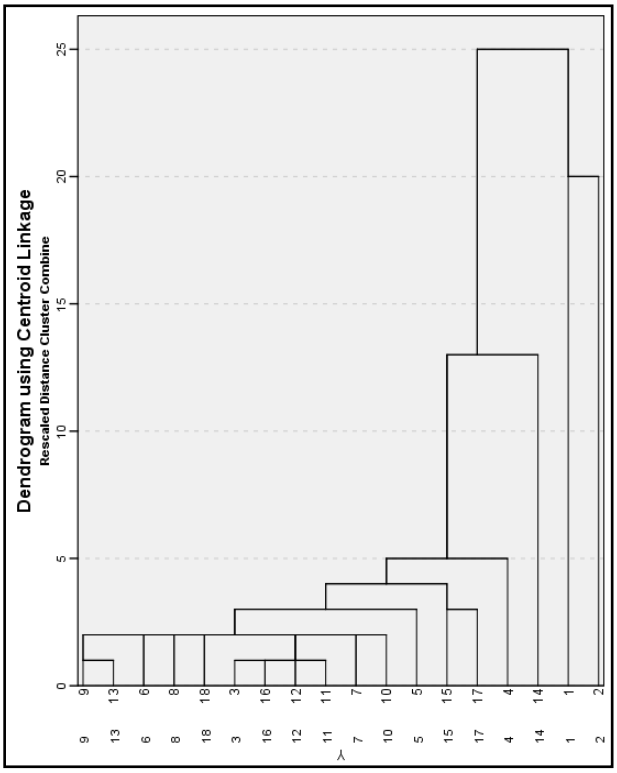

Figure 4: show the level of fusion

Table (5) shows the levels of docking mediator between groups

Stage	Cluster Combined		Coefficients
	Cluster 1	Cluster 2	
1	9	13	3.606
2	3	16	4.796
3	3	12	4.673
4	3	11	4.399
5	6	9	5.612
6	6	8	4.742
7	6	18	5.357
8	3	10	5.625
9	3	7	4.924
10	4	5	7.483
11	4	6	7.706
12	15	17	7.746
13	3	4	8.253
14	3	15	8.517
15	2	14	27.586
16	2	3	28.754
17	1	2	49.833

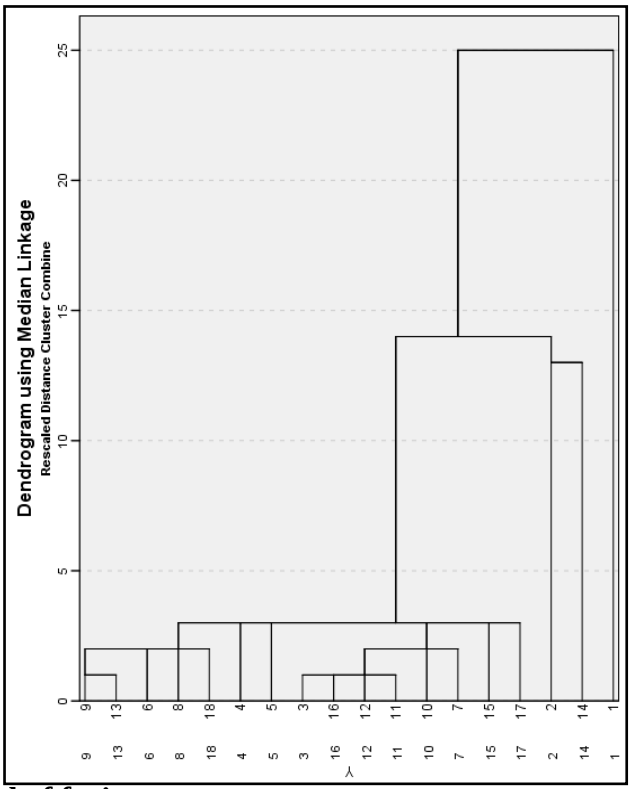

Figure 5: show the level of fusion

Table (6) shows the levels of docking Ward method between groups

Stage	Cluster Combined		Coefficients
	Cluster 1	Cluster 2	
1	9	13	1.803
2	3	16	4.201
3	11	12	6.893
4	7	10	9.722
5	6	8	12.594
6	4	5	16.336
7	3	11	20.201
8	15	17	24.074
9	6	18	27.969
10	6	9	33.271
11	3	7	40.625
12	3	15	50.179
13	4	6	62.833
14	2	14	76.627
15	3	4	98.509
16	1	2	124.645
17	1	3	201.406

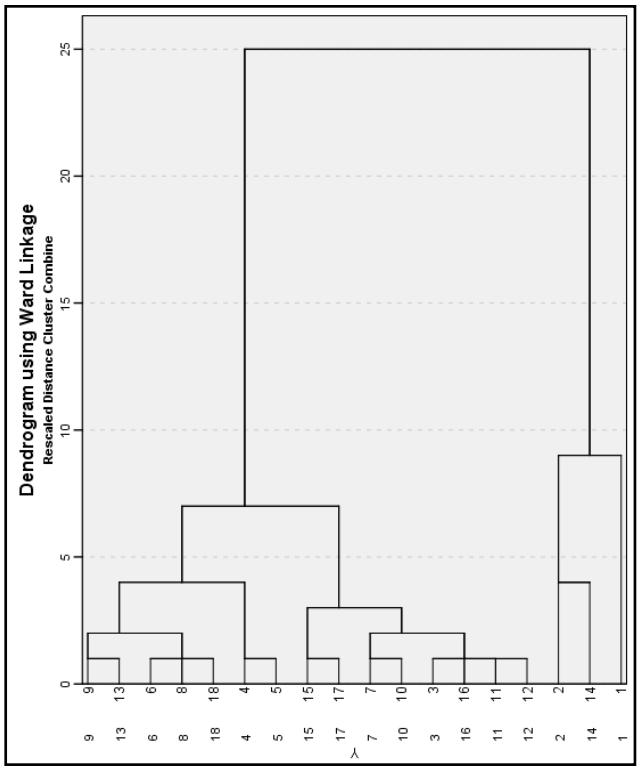

Figure 6: show the level of fusion

5- Conclusions

The results of the hierarchical methods have shown that the means of averages between groups and single linking and the method of universal linking and linking mediator and central and Ward that have the same results but were different in the levels of fusion.

Dohuk, Karbala, Wasit, Kirkuk, Sulaymaniyah, Maysan, Diwaniyah are close in terms of the incidence of cancer and that the province of Nineveh, Basra and Baghdad, the most affected by these diseases of cancer of high levels of docking.

Perhaps the reason is that the Baghdad governorate is the capital and that the population rate in it is high and that people go to its hospitals for the provision of health services more than the rest of the governorates, in addition to that it is exposed to environmental pollution due to overcrowding and the large number of cars, factories and refineries, which leads to the registration of cases of this dangerous disease, followed by Nineveh And Basra for the same reasons.

References

1- Alvin C. Rencher(2002) “Methods of Multivariate Analysis”, Second Edition, Brigham Young University.

2- Brian S. Everitt , Sabine Landau, MorvenLeese ,Daniel Stahl(2011) "Cluster Analysis" 5th Edition, John Wiley &Sons,P77
3- Diana, F(2011);"Poor mental health symptoms among Romanian employees. A Two-Step Cluster analysis", Romania, Procedia- Social and Behavioral Sciences.
4- Fazzo,L.et al " Cluster Analysis of mortality and mal formations in the provinces of Naples and Caserta (compania Region)",Ann 1st super Sanita, Rome, Italy,Vol.(44),NO.(1),PP(99-111)2008
5- Hardle ,W. and Simer, L (2003) "Applied Multivariate StatisticalAnalysis", Springer , Berli
6- Johnson R. A.,&Wichern, D. W. (2002);"Applied Multivariate statistical analysis, Uppre Saddle River(NJ);prentice-Hall.
7- Muntaner, C.; H. Chung; and J.Benach (2012).Hierarchical cluster analysis of labor market regulations and population health: a taxonomy of low-and middle-income countries" BMC Public Health, Department of Healthcare Management, Korea University, Seoul, Korea.
8- Rosa, V. Fida, R. Avallone ,F (2011) ;"A Cluster Analysis of High School Students' Styles of Living – Together in the Classroom" , Italy, Procedia – Social and Behavioral Sciences.
9- Timm, N. H (2002) " Applied Multivariate Analysis" , Springer-Verlag New York, Inc.USA.