ISSN 2303-4521
Periodicals of Engineering and Natural Sciences *Original Research*
Vol. 11, No. 1, February 2023, pp.174-185

A comparison between the hierarchical clustering methods for postgraduate students in Iraqi universities for the year 2019-2020 using the cophenetic and delta correlation coefficients

Ameena Kareem Essa[1], Laith Fadhil S. H. [1], Dhamyaa Hamid Shihab[2]
[1]Statistics Department, Mustansiriyah University, Baghdad, Iraq
[2]Statistics Department, Baghdad University, Baghdad-Iraq.

ABSTRACT

The educational sector is one of the important sectors in the world, and it is considered one of the means of community development. In addition, it is one of the means of making the country's renaissance and development because it represents the factory of thinking minds that make change. There is no doubt that this sector is the same as any other sector. The deficit in the studied scientific planning has been prolonged, which led to its deterioration, and the problems of education remain diverse and inherited from previous time periods, where the hierarchical cluster analysis was used on postgraduate students in universities in Iraq, except for Kurdistan region, and the number of universities that were included in the study was (30) universities. In the whole of Iraq for the year 2020, when using the comparison measures the Cophenetic Correlation Coefficient (CCC) and the Coefficient Delta (DC), it was found that the Complete Linkage Method is the best among the hierarchical methods, as the value of (CCC) is 0.952061, and the value of (DC(0.1)) it is 0.288973, and in the case (DC(0.5)) it is 0.26877, then followed by Median method, Ward's method and finally Single Linage Method.

Keywords: Cophenetic Correlation Coefficient, Delta Coefficient, Cluster Analysis

Corresponding Author:

Collage of Management and Economic
Mustansiriyah University, Baghdad, Iraq
E-mail: ameena@uomustansiriyah.edu.iq

1. Introduction

Postgraduate studies are among the advanced stages of education, as the beginning of postgraduate studies goes back to the Middle Ages, the beginning of the Renaissance [1]. Postgraduate studies include the higher diploma, master's and doctoral stages. Postgraduate studies require the student to obtain a bachelor's degree, and despite the technological development, the Education in Iraq still suffers from a lack of laboratory equipment, modern teaching tools and modern methods of teaching.

The most important obstacles to education in Iraqi universities are the absence of strategic planning for the educational system, the small number of universities distributed throughout the country and the increase in the number of students applying for postgraduate studies in recent years, the absence of libraries, the small number of halls and their overcrowding with students. The graduates are faced with acceptance or rejection of certificate holders, so the educational reality was studied by one of the statistical analysis methods, which is Cluster Analysis and the classification of postgraduate students (high diploma, master's, and doctorate) on Iraqi universities and their arrangement in certain ways within clusters [2-4].

2. Method

2.1. Cluster analysis

It is one of the important analysis methods common in statistics used to classify the groups into the states of the variables in certain ways and then they are arranged into clusters where the cases classified within the cluster

are homogeneous with each other. There are many researchers who have used cluster analyzes to refer to the methods that search for groups of variables and then suggest them as an alternative to Principal Component Analysis, and researchers who used this term (Kendall and Sturat) and also used the word Classification for the purpose of collecting vocabulary [5-8].

2.2. Uses of cluster analysis

Cluster Analysis is used to describe and make spatial and temporal comparisons of communities (clusters) of organisms in heterogeneous environments, and the most important uses of cluster analysis are the following [9]:

a) Data exploration: Cluster analysis is usually used in the exploratory phase of research when the researcher does not have any pre-designed hypotheses.
b) Classification: It is the specialization and summarization of the study data and placing them in fewer clusters.
c) Identification: It is important in the analysis process, where the observations or the study population are divided into clusters.
d) Generating special hypotheses for each study, as the cluster analysis can provide us with the assumptions about the study community and it can also prepare a hypothesis about the structure of the community from which the data were taken.
e) Prediction: We get the results of the study from the cluster analysis after categorizing them into forecast groups for later studies.

2.3. Distance coefficients

They are coefficients that measure the distance between variables, as the distance between any two variables increases, the difference increases. These are the criteria for assessing the lack of similarity [10, 11, 12]:

- Euclidean distance.
- The Minkowfski method for measuring distance.
- Distance for mixed data.
- Distance for binary variables.

2.4. Euclidian distance measure

This scale is the most widely used to calculate the Euclidian distance between the different elements to determine the degree of convergence between the elements and is calculated according to the following formula:

$$D_{k,l} = \sum_{j=1}^{p} \left(X_{kj} - X_{lj}\right)^2 \quad , \ldots (1)$$

As $D_{k,l}$ represents the Euclidian distance between the two points (D_k), (D_l), (X_i) and $\left(X_j\right)$ which are the (ith) and (jth) elements in the dimension.

2.5. Steps of clustering methods

The clustering process employs variables in groups in two ways, the first being the Agglomerative Method and the second the Divisive Method. For the purpose of performing the clustering process, the following steps must be followed:

1) Calculating the Distance Matrix or Correlation Matrix [12].
2) Inside the matrix, the shortest distance between two elements is searched to be linked. The linking process can be done for more than two elements if they have an equal distance in one stage to form the beginning of the initial clusters.
3) We calculate again the calculation of the new distance matrix and take into account the changes that occurred in the previous stage of the process of linking the matrix in the previous stage, so the new matrix will be of lower dimensions.
4) Continuing the linking process and depending on the shortest distance until the cluster tree is reached.

2.6. Hierarchical cluster analysis

This method is one of the preferred methods by which we do not mean dividing the study data into a number of clusters in one step, but rather it is a hierarchy of interconnected clusters, which shows the process of linking clusters with each other through an overlapping series to give a hierarchical shape is the Dendrogram, which It frequently merges small clusters into larger clusters, which is the Agglomerative method, or by separating large clusters into small clusters, which is the Divisive method, and this process can be illustrated in the diagram below[1,5, 13, 14]:

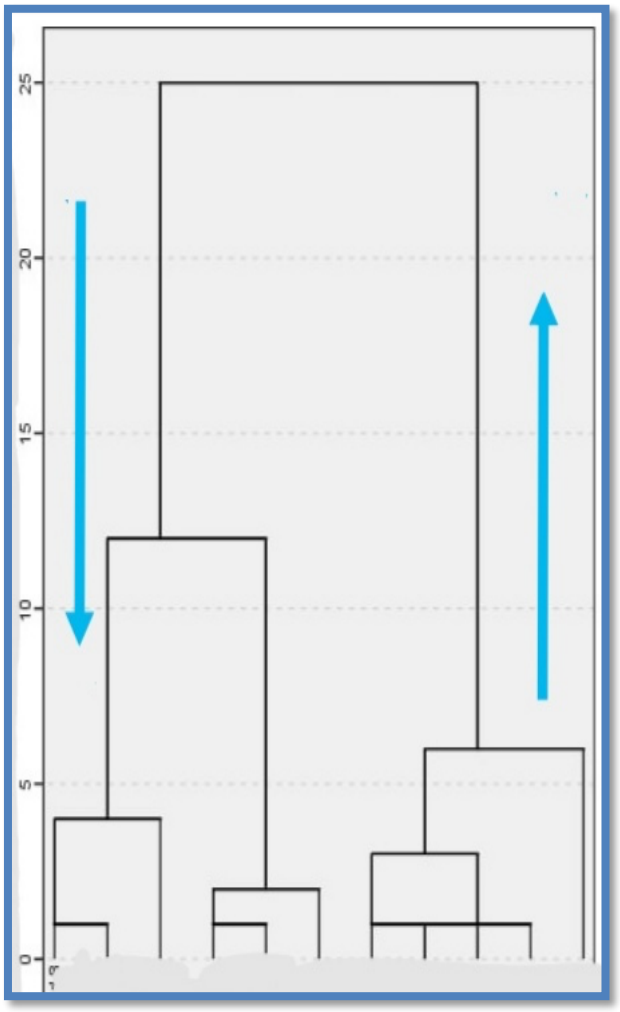

Figure 1. Hierarchical clustering

2.7. Methods of typing hierarchical clustering

There are many types of cluster analysis methods, the most important of which are as follows:

2.7.1. Single linkage method (nearest neighbor)

This method is the simplest and oldest method, and it is also considered the most widespread, as it collects the closest in the distance to the elements to form the nucleus of the clusters, then add the rest of the elements that are more similar and close in the distance, resulting in a long series of interconnections. To determine the distance between the clusters, it is calculated according to the following formula[4]:

$$D(I,J) = \min(D_{ij}) \quad , i \in I \, , j \in J \qquad \ldots (2)$$

Where D(I,J) represents the Euclidian distance for the variables in the clusters I,J

2.7.2. Complete linkage method (furthest neighbor)

In this method, the cluster is formed in a way that reflects the first method, that is, it depends on at least similarity between the variables, which is the farthest distance, according to the following formula [1]:

$$D(I,J) = \max(D_{ij}) \quad , i \in I \, , j \in J \qquad \ldots (3)$$

2.7.3. Median method

This method is used if one of the clusters contains a number of items that is greater than the other cluster, then the cluster is grouped by the middle point in the central distance that is closer to the contracts with the most part than the other contracts that are closer to the central distance, so the clusters with the smallest distance as in the following formula[7]:

$$m_{AB} = \frac{1}{2}(\bar{X}_A + \bar{X}_B) \quad \ldots (4)$$

That is, any two clusters that have the smallest distance between their mean in each stage are linked.

2.7.4. Ward's method

The hierarchical method is based on the least loss of information for the purpose of clustering. This method depends on the use of the square of the distances within each cluster, the square of the distances between the clusters, which is expressed by the following formulas [2,9]:

$$SSE_A = \sum_{i=1}^{n_A} (X_i - \bar{X}_A)'(X_i - \bar{X}_A) \quad \ldots (5)$$

$$SSE_B = \sum_{i=1}^{n_B} (X_i - \bar{X}_B)'(X_i - \bar{X}_B) \quad \ldots (6)$$

$$SSE_{AB} = \sum_{i=1}^{n_{AB}} (X_i - \bar{X}_{AB})'(X_i - \bar{X}_{AB}) \quad \ldots (7)$$

Whereas, AB is the cluster resulting from linking clusters A and B so that they are linked to reduce the increase in the square of distances (SSE), and the amount of that increase is expressed as follows:

$$I_{AB} = SSE_{AB} - (SSE_A + SSE_B) \quad \ldots (8)$$

2.8. Standard of comparison

Among the criteria that are commonly used today for comparisons in statistical analysis, especially in cluster analysis, are[10]:

2.8.1. Cophenetic correlation coefficient (CCC)

It is the difference between the value of the correlation matrix for the original and shrinked distances, and the high value of this criterion is considered good as in the following formula:

$$\mathrm{CCC} = \left[\frac{\sum_{i<j}^{n}(d_{ij} - d_{ij}^{*})}{\sum_{i<j}^{n} d_{ij}^{2}}\right]^{\frac{1}{2}} \quad \ldots (9)$$

whereas:

$\boldsymbol{d_{ij}}$ values of the original distance matrix.

$\boldsymbol{d_{ij}^{*}}$ values of the matrix of shrinking distances.

2.8.3. Coefficient delta (DC)

This standard measures the degree of distortion rather than the degree of similarity. The formula of the standard is:

$$\Delta_{\mathrm{A}} = \left[\frac{\sum_{\mathrm{j<k}}^{\mathrm{n}} |\mathrm{d_{jk}} - \mathrm{d_{jk}^{*}}|^{1/\mathrm{A}}}{\sum_{\mathrm{j<k}}^{\mathrm{n}} (\mathrm{d_{jk}^{*}})^{1/\mathrm{A}}}\right]^{\mathrm{A}} \quad \ldots (10)$$

Whereas A takes the values (0.5) or (0.1) and the value of d_{jk}^{*} is the formation distance when forming clusters, and the closer the value of the delta coefficient is to zero, the analysis is good.

1. Results and discussion

In this study, postgraduate students were included in all Iraqi universities except for the Kurdistan region, and the number of universities was (30) universities. The higher diploma stage was encoded by the variable x1, the master's stage by the variable x2, and the doctoral stage by the variable x3, Table 1 shows the data used in this research and the preparation of postgraduate students distributed among Iraqi universities as follows:

Table 1. The number of postgraduate students in Iraqi universities

ID	University Name	x1	x2	x3
1	Baghdad	718	5878	3237
2	Al-Mustansiriya University	113	2226	948
3	University of Technology	10	993	420
4	Al-Nahrain	63	989	484
5	Iraqi University	14	1183	228
6	Iraqi commission for computers and informatics	0	0	2594
7	Iraqi Board for Medical Specializations	30	73	19
8	Mosul	380	2401	772
9	Nineveh	0	34	0
10	Basrah	97	1644	991
11	Kufa	265	1842	666
12	Tikrit	168	1811	705
13	Samarra	0	315	127
14	Al Qadissiya	65	1214	322
15	Anbar	3	1299	413
16	Faluja	6	108	12
17	Babylon	302	2089	1007
18	Al Qasim Green University	0	244	3
19	Diyala	45	1038	337
20	Karbala	100	1100	257
21	Thi - Qar	18	739	218
22	sumer	0	28	0
23	Wasit	26	598	118
24	Kirkuk	9	452	89
25	Al Muthana	0	343	0
26	Maysan	0	226	21
27	Technical Education/ Northern area	0	121	0
28	Technical Education/ Middle area	0	455	29
29	Al- furat Al-Awsat Technical University	2	180	12
30	Technical Education/ Southern area	11	135	0
Sum		2445	29758	14029

Table 2. Agglomeration schedule single linkage method

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	9	22	6.000	0	0	8
2	27	30	17.804	0	0	3
3	16	27	18.682	0	2	5
4	18	26	25.456	0	0	6
5	7	16	43.012	0	3	7
6	18	29	46.915	4	0	7
7	7	18	47.434	5	6	8
8	7	9	52.745	7	1	11
9	24	28	60.745	0	0	16
10	3	4	83.193	0	0	12
11	7	25	99.045	8	0	16
12	3	19	100.693	10	0	15
13	11	12	109.046	0	0	24
14	5	14	111.346	0	0	17
15	3	20	115.191	12	0	17
16	7	24	115.694	11	9	18
17	3	5	122.988	15	14	19
18	7	13	130.050	16	0	20
19	3	15	139.104	17	0	23
20	7	23	149.820	18	0	21
21	7	21	173.046	20	0	23
22	2	17	240.772	0	0	25
23	3	7	322.941	19	21	27
24	10	11	338.712	0	13	26

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
25	2	8	364.541	22	0	26
26	2	10	422.681	25	24	27
27	2	3	612.073	26	23	28
28	2	6	2298.207	27	0	29
29	1	2	4275.511	0	28	0

In the table above, the single linking method was used based on linking the Euclidian distance variables, i.e. the shortest distance represents the first stage in the clustering process. In the table, we note the first stage, which brings together the Nineveh and Sumer Universities for codes (9,22) with a Euclidean distance (6.000), which is the shortest, then the second stage connects the Technical University in the northern and southern region of codes (27,30) and the distance (17.804) and so on for the rest of the stages.

In the above figure, we have three clusters, the University of Baghdad in the first cluster and the Iraqi commission for computers and informatics University in the third cluster, while the second cluster included the rest of the universities.

Table 3. Agglomeration schedule complete linkage method

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	9	22	6.000	0	0	5
2	27	30	17.804	0	0	4
3	18	26	25.456	0	0	7
4	16	27	29.967	0	2	9
5	7	9	57.324	0	1	9
6	24	28	60.745	0	0	14
7	18	29	64.661	3	0	15
8	3	4	83.193	0	0	20
9	7	16	107.564	5	4	19
10	11	12	109.046	0	0	23
11	5	14	111.346	0	0	16
12	19	20	115.191	0	0	16
13	13	25	130.050	0	0	15
14	23	24	170.429	0	6	21
15	13	18	177.353	13	7	19
16	5	19	184.030	11	12	18
17	2	17	240.772	0	0	22
18	5	15	275.065	16	0	20
19	7	13	315.000	9	15	24
20	3	5	324.920	8	18	26
21	21	23	341.615	0	14	24
22	2	8	398.313	17	0	25
23	10	11	415.996	0	10	25
24	7	21	743.888	19	21	26
25	2	10	837.317	22	23	27
26	3	7	1336.420	20	24	27
27	2	3	2524.186	25	26	28
28	2	6	3037.908	27	0	29
29	1	2	6724.299	0	28	0

In Table 3, we note the first stage between case 9 and case 22, the coefficient is 6, and in the second stage between 27 and 30, the coefficient is 17, if we move to the tenth stage between 11 and 12, the coefficient is 109,046, and in the 26th stage we notice the value of the coefficient between 3 and 7 is 1336.420 This is a big shift in the Euclidian distance between the stages.

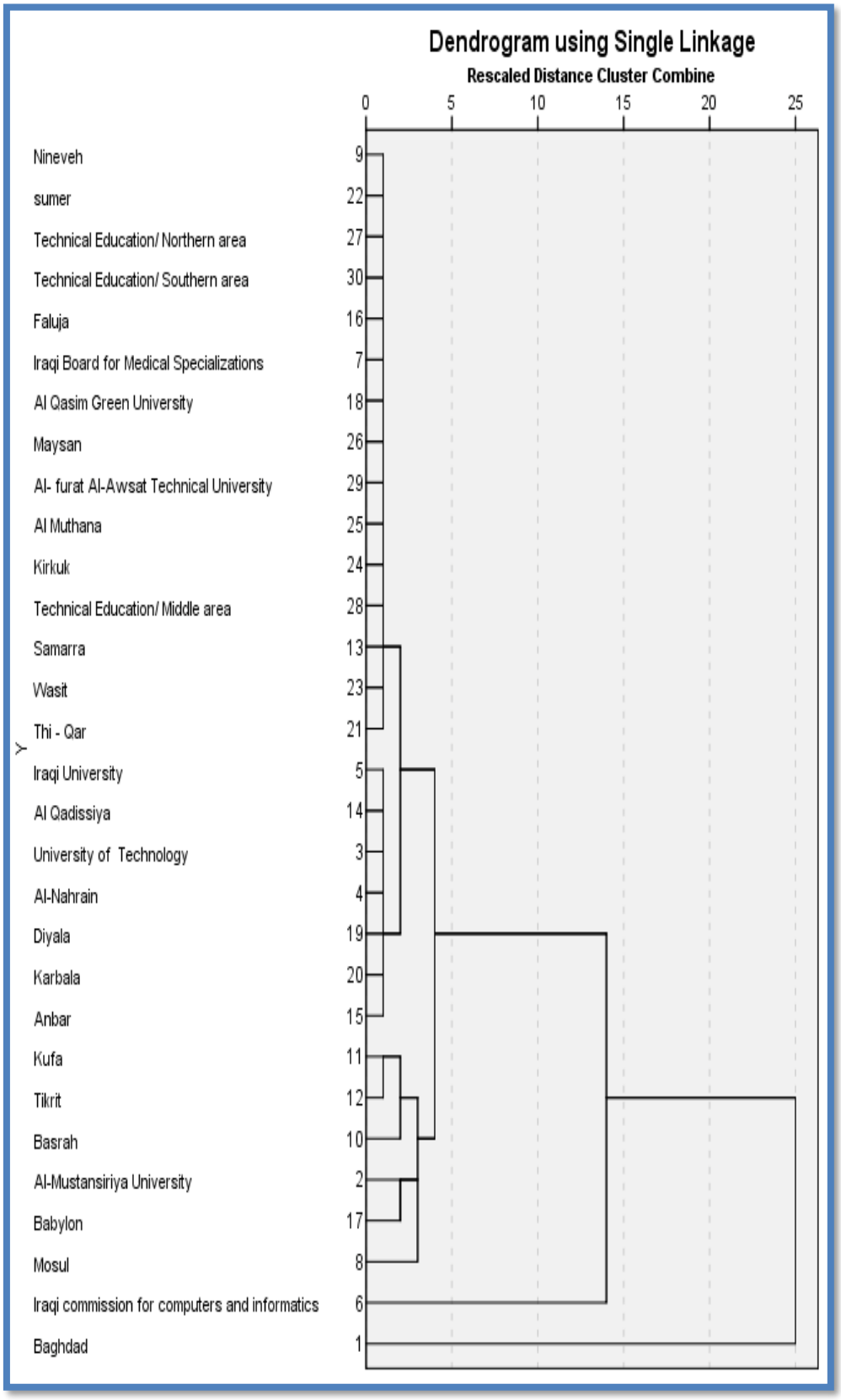

Figure 2. Dendrogram using single linkage method

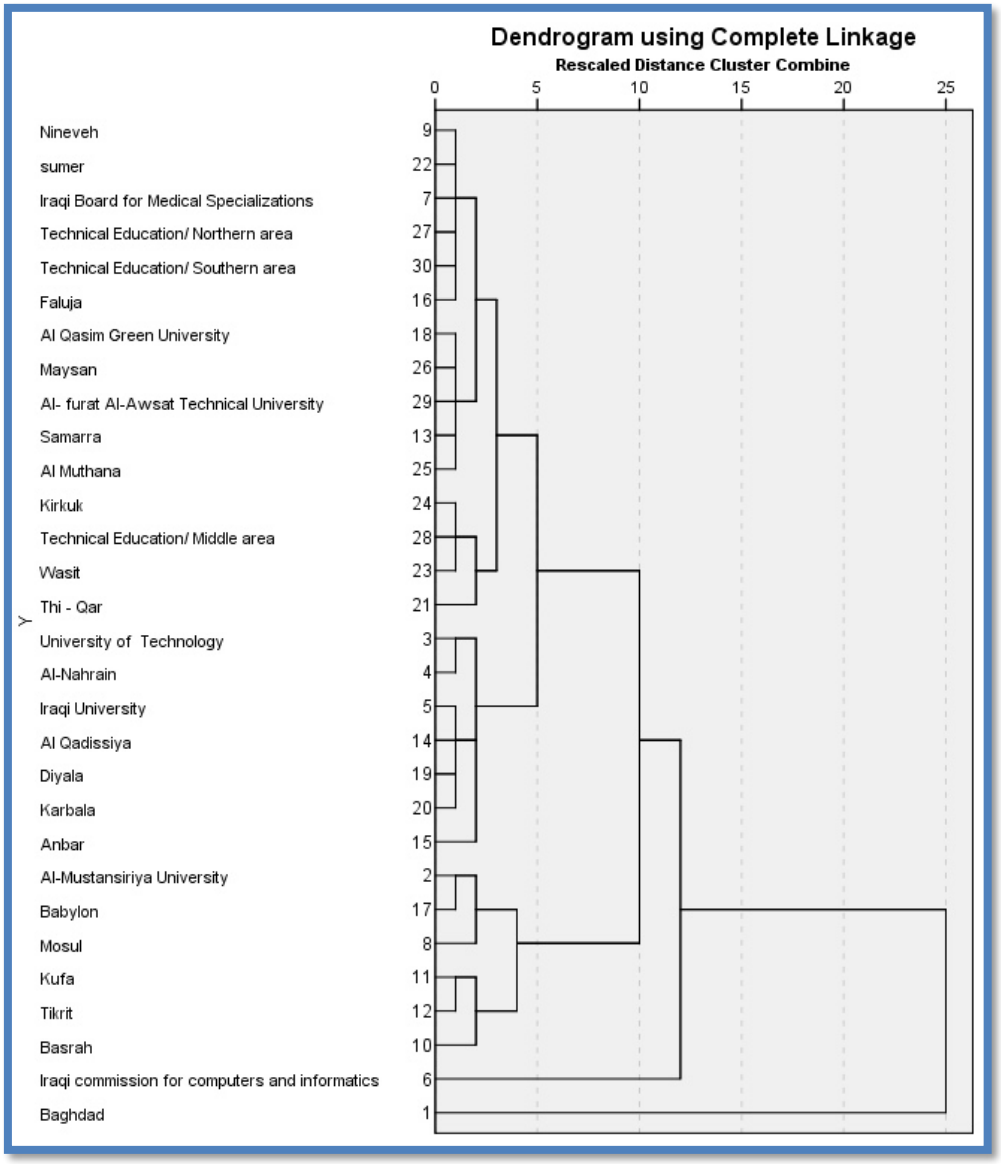

Figure 3. Dendrogram using complete linkage method

Table 4. Agglomeration schedule median method

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	9	22	6.000	0	0	7
2	27	30	17.804	0	0	3
3	16	27	19.873	0	2	5
4	18	26	25.456	0	0	6
5	7	16	46.132	0	3	7
6	18	29	49.424	4	0	11
7	7	9	54.750	5	1	11
8	24	28	60.745	0	0	15
9	3	4	83.193	0	0	10
10	3	19	107.545	9	0	17
11	7	18	108.088	7	6	18
12	11	12	109.046	0	0	24
13	5	14	111.346	0	0	14
14	5	20	101.565	13	0	17

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
15	24	25	113.164	8	0	16
16	13	24	107.478	0	15	18
17	3	5	126.210	10	14	20
18	7	13	152.220	11	16	22
19	21	23	173.046	0	0	22
20	3	15	190.065	17	0	26
21	2	17	240.772	0	0	23
22	7	21	318.505	18	19	26
23	2	8	321.234	21	0	25
24	10	11	350.093	0	12	25
25	2	10	398.306	23	24	27
26	3	7	560.844	20	22	27
27	2	3	917.937	25	26	28
28	2	6	2144.520	27	0	29
29	1	2	4836.730	0	28	0

In Table 4, we notice the initial stages, but there is a large shift in distance, but as the stages progress, the value of the coefficient increases. For example, in the seventh stage, the coefficient between case 7 and case 9 was 54.750, while in stage 28 the coefficient was 2144,520. This is a leap for the initial stages.

Table 5. Agglomeration schedule ward's method

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	9	22	3.000	0	0	7
2	27	30	11.902	0	0	4
3	18	26	24.630	0	0	6
4	16	27	37.879	0	2	14
5	24	28	68.252	0	0	15
6	18	29	101.201	3	0	19
7	7	9	136.890	0	1	14
8	3	4	178.486	0	0	21
9	11	12	233.009	0	0	22
10	5	14	288.682	0	0	16
11	19	20	346.278	0	0	16
12	13	25	411.303	0	0	15
13	21	23	497.826	0	0	23
14	7	16	586.702	7	4	19
15	13	24	681.516	12	5	23
16	5	19	780.031	10	11	18
17	2	17	900.417	0	0	20
18	5	15	1038.786	16	0	21
19	7	18	1228.079	14	6	25
20	2	8	1442.235	17	0	24
21	3	5	1666.319	8	18	26
22	10	11	1899.714	0	9	24
23	13	21	2183.560	15	13	25
24	2	10	2699.147	20	22	27
25	7	13	3551.005	19	23	29
26	3	6	5667.950	21	0	27
27	2	3	8160.543	24	26	28
28	1	2	12584.869	0	27	29
29	1	7	19806.108	28	25	0

In the table above, we notice that the distances moved significantly from the first stage to the fourth stage, and then began to increase. These coefficients were in the advanced stages 10, 15, 20, and then stage 2.

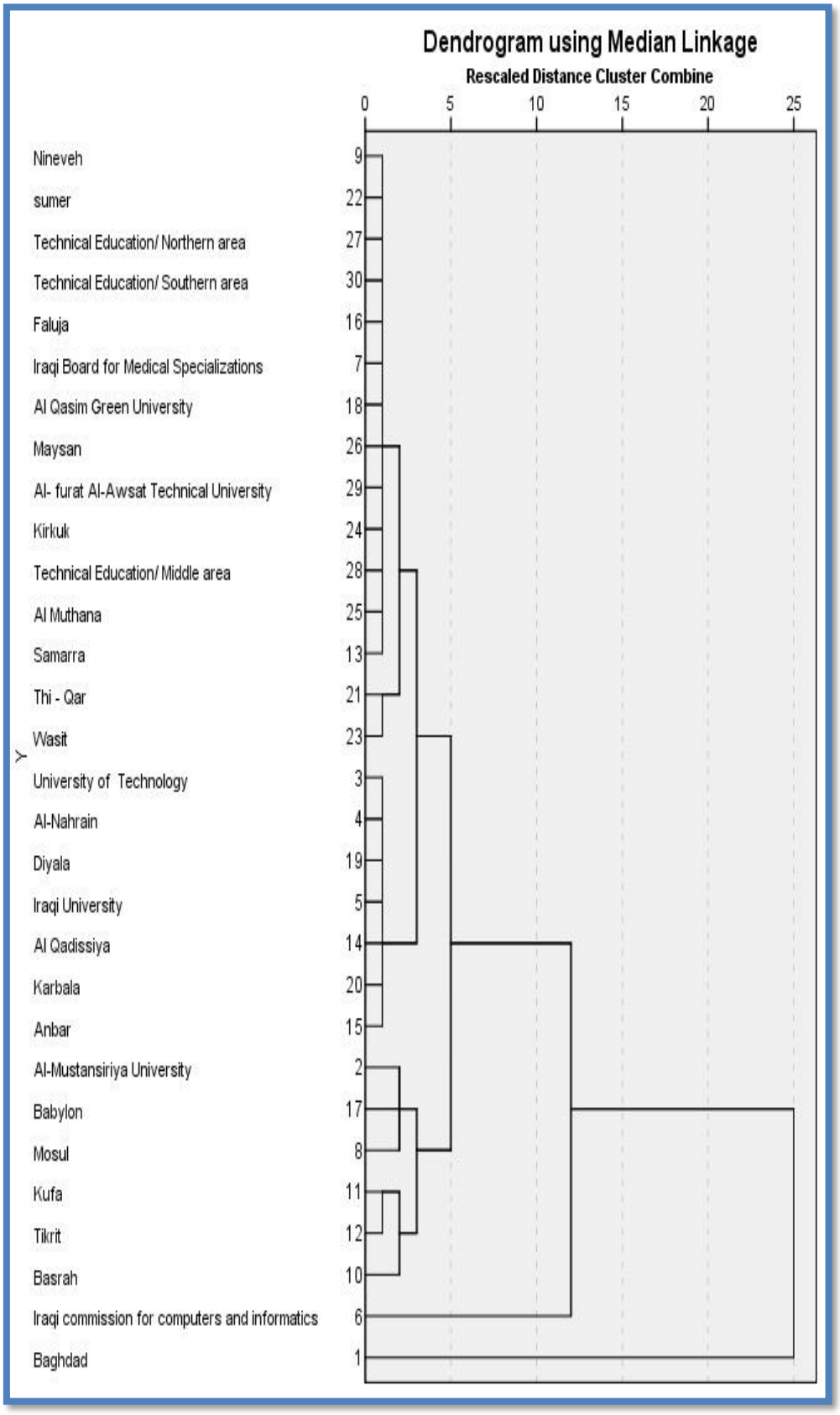

Figure 4. Dendrogram using Median Method

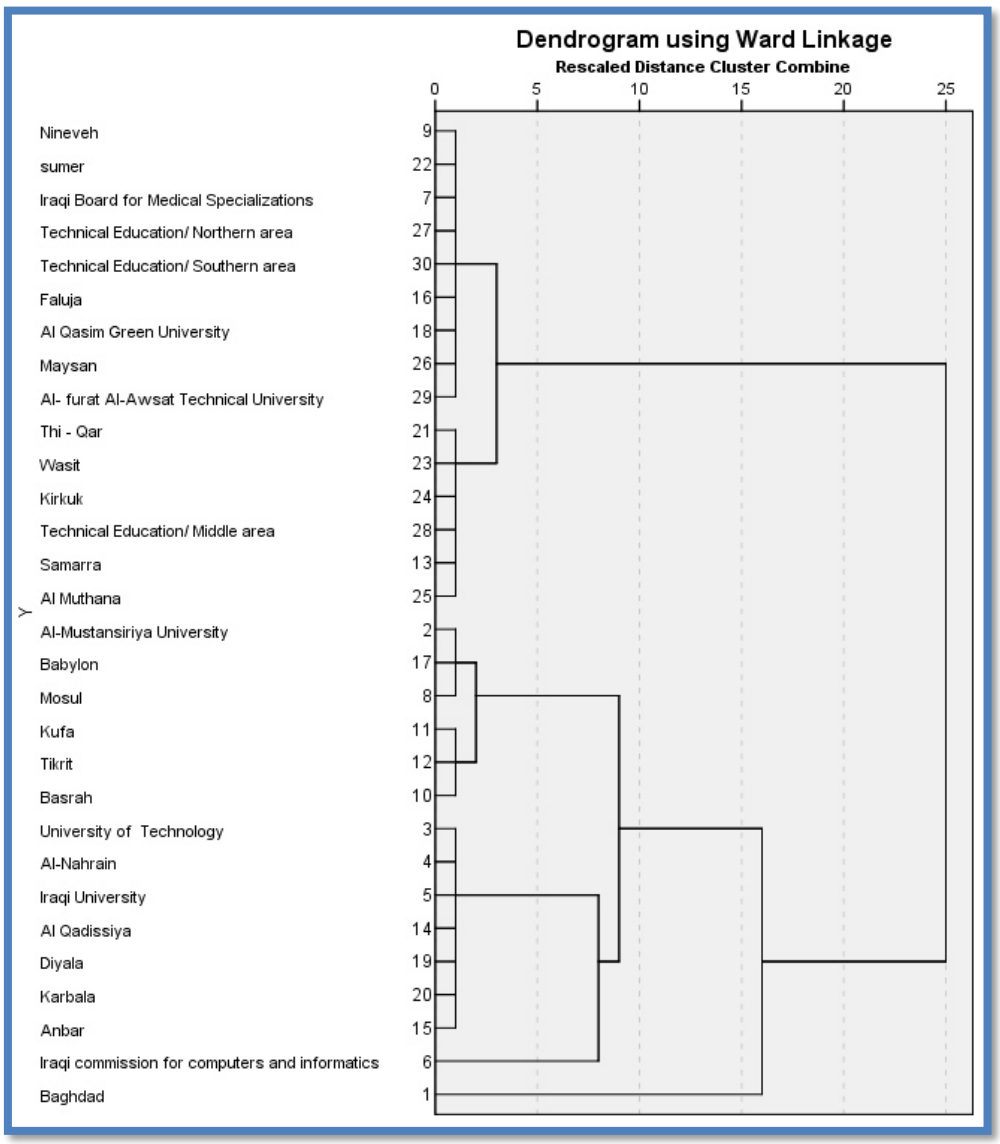

Figure 5. Dendrogram using Ward's Method

To test the best method among the hierarchical methods, CCC , DC(0.1) and DC(0.5) were used, and the ready-made statistical program (NCSS) version (22) was used.
Table 6 shows the best method of the hierarchical methods used in this research, if the comparison criterion of CCC and DC is used, then the good method is the complete method, because the value of the delta is small, the smaller the method is, then the median method is followed:

Table 6. Hierarchical methods with CCC & DC

Hierarchical Methods	CCC	DC(0.5)	DC(1.0)
Single Linkage Method (Nearest Neighbor)	0.947771	0.815927	0.679880
Complete Linkage Method (Furthest Neighbor)	**0.952061**	**0.268770**	**0.288973**
Median Method	0.964261	0.407260	0.396267
Ward's Method	0.830458	0.744886	0.692918

2. Conclusion

After obtaining the results of the study on the thirty Iraqi universities, it is necessary to mention the most important outcomes of the study:

1. The aggregative hierarchical methods are convergent in the classification of universities for each cluster.
2. The results showed that the best method was the perfect connection, because the value of the delta was the smallest value than the rest of the methods as in Table 6, the more the value of the delta is close to zero can be better
3. One of the good comparisons that are used in mask analysis through recent studies is the comparison criterion, the shrink correlation coefficient and the delta coefficient to highlight the best methods in the multiple analysis.
4. We note the classification of Iraqi universities according to postgraduate students through the method of full linkage. The number of clusters is three. It is necessary in the classification process to know the number of clusters for each method, if the first cluster consists of the University of Baghdad, and the second cluster is Iraqi commission for computers and informatics and contracts. The third guarantees the rest of the universities. We will mention the university code for ease (2, 3, 4, 5, 7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23, 24, 25, 26, 27, 28, 29, 30).
5. The universities in each cluster are more homogeneous in terms of composition among them.

Declaration of competing interest

The authors declare that they have no known financial or non-financial competing interests in any material discussed in this paper.

Funding information

No funding was received from any financial organization to conduct this research.

References

[1] M. S. Aldenderfer and R. K. Blast, "Cluster Analysis", 1st ed., Sage Puplications, Newbury, Park, Call, pp.9-59, 1984.

[2] A. Afifi and V. Clark, "Computer-Aided Multivariate Analysis", 3rd ed.Chapman & Hall, New York, pp.361-391, 1996.

[3] L. S. Stanley, "Notes on Cluster Analysis", 2001. .Available at: http://www.statsoftince.com/textbook/stcluan.html.

[4] B. S. Everitt, S. Landau and M. Leese, " Cluster Analysis " , 4th Edition,Edward Arnold, pp.450-462, 2001.

[5] A. C. Rencher, "Methods of Multivariate Analysis", Second Edition, Brigham Young University, 2002.

[6] N. H. Timm, " Applied Multivariate Analysis" , Springer-Verlag New York, Inc.USA, 2002.

[7] R. A. Johnson and D. W. Wichern, "Applied Multivariate statistical analysis, Uppre Saddle River(NJ);prentice-Hall, 2002.

[8] W. Hardle and L. Simer, " Applied Multivariate Statistical Analysis ", Springer , Berli, 2003.

[9] B. S. Everitt , S. Landau, M. Leese, D. Stahl, "Cluster Analysis" 5th Edition, John Wiley &Sons, p.77, 2011.

[10] S. Saracli, et al "Comparison of hierarchical cluster analysis methods by cophenetic correlation " , Journal of Inequalities and Applications, pp.1-8, 2013

[11] K.B. Yao, "A comparison of clustering methods for unsupervised anomaly detection in network traffic", Ph.D. Thesis, University of Copenhagen, 2006.

[12] A. Mohammed, A. G. AL-Rawi, " Using Some of Hierarchical Approach of Cluster Analysis for Classification of Agricultural Lands by Area and the Amount of Production for some Agricultural Crops in the Iraqi Governorates for the Years (2005) and (2010)" ; Al-Rafidain University College of Science Journal ; pp.52-74, 2019.

[13] A. Dh. Ahmed, D.I. Mahdi, S.N. Abood, "Classification The Iraq Provinces according to some variants of the health sector," *AL-Qadisiyah Journal For Administrative and Economic sciences* 17, no. 3, pp. 271-285, 2015.

[14] T. M. Abbas, S. M. Hameed, Q. N. Naif, "The use of cluster analysis to analyse the factors affecting the heart ",*Al-Nahrain Journal of Science* ,13 no. 3, pp. 58-65, 2010.