JOURNAL OF UNIVERSITY OF BABYLON
For Pure and Applied Sciences (JUBPAS)

ARTICLE
Vol.29; No.3. October-December | 2021

Comparison of the Two Methods of Cluster Analysis (Non-Hierarchical and Hierarchical) in the Classification of Laboratory Quality (GLP) at the University of Babylon

Aasha Abdulkhleq Ismael [1], Zainab Abood Ahmed AL-Bairmani[2]
Israa Faleh Fadel AL-Masferi [3]

1Administration & Economics, Tikrit University, aasha.a.alkalek@tu.edu.iq, Tikrit, Salahaddin.
2Dep. Studies & Planning, University of Babylon, zanab.abboud@uobabylon.edu.iq, Hillah, Babil.
3 Administration & Economics, University of Babylon, isra.faleh@uobabylon.edu.iq, Hillah, Babil.
* email: zanab.abboud@uobabylon.edu.iq; mobile: 07700822025

Received: 5/4/2021 **Accepted: 23/5/2021** **Published: 1/11/2021**

Abstract

In this research, the two methods of cluster analysis (**non-hierarchical/ K-Means method and hierarchical/ Ward Linkage method**) were used to classify the quality of accreditation of educational laboratories in the faculties of the University of Babylon, which numbered (**17**) laboratories, and it was concluded that the variable (x_4**: training course programs for laboratory workers**) has the largest differences (**F=12.232**) It is followed by the variable (x_5**: maintenance programs for devices and equipment**) with statistically significant differences (**F=4.125**), while the least variable has a significant difference (x_7: **Certificates of appreciation and letters of thanks to the laboratory workers**) with significant differences (**F=.020**), in addition to the presence of six laboratories (**4, 5, 6, 9, 10, 16**) joined the first cluster in both methods.

Key words:

Cluster, hierarchical, non-hierarchical, K-Means, Ward Linkage.

Citation:

Aasha Abdulkhleq Ismael [1], Zainab Abood Ahmed AL-Bairmani[2] Israa Faleh Fadel AL-Masferi[3]. Comparison of the Two Methods of Cluster Analysis (Non-Hierarchical and Hierarchical) in the Classification of Laboratory Quality (GLP) at the University of Babylon. Journal of University of Babylon for Pure and applied science (**JUBPAS**). October-December , 2021. Vol.29; No.3; p:34-43.

INTRODUCTION

Cluster analysis: is a type of data reduction technique that reduces data such as factor analysis that reduces the variables included in the model and turns them into factors, Likewise, the discriminatory analysis works to classify new cases into groups that were previously identified according to specific criteria, But cluster analysis is unique to these techniques because its job is to reduce data by classifying them into homogeneous groups and defining them without having to know the group

membership or the number of potential groups in advance, Cluster analysis also allows many options regarding the algorithm for combining aggregates and with each option results in a different clustering structure, Therefore, cluster analysis is a convenient statistical tool for exploring the basic structures in different types of data sets [1].

Cluster analysis develops methods and tools for classifying a group of observations. This is done by grouping similar vocabulary according to some appropriate criteria. Cluster analysis is used in many fields and sciences (medical, natural, social,...etc [2].

Uses of cluster analysis [3]:

Data disclosure: Cluster analysis is a way to know the structure of the data.

Diagnosis: Through the analysis process, the observations are divided into clusters, so that we can diagnose (identify) each observation.

Classification: Through the analysis, the data can be divided and summarized into the least possible number of clusters.

Generating hypotheses: Cluster analysis can provide hypotheses about the structure of the community from which the data were taken.

Prediction: The results obtained from the analysis (clusters) can be predicted later.

This is why most researchers have dealt with these fields. As for this research, researchers will take a topic that no one has previously discussed, which is educational laboratories in universities because of their important and essential role in the student's march in terms of developing his skill, scientific thinking and proper preparation for the experiment according to standard standards and for all disciplines and including Fits with modern science and how to identify problems and conclusions, explain the experiment process, as well as improve or develop skills with great scientific diversity, and the results of the Laboratory Accreditation Quality Form in the Faculties of the University of Babylon have been relied upon, which includes a number of indicators required for each laboratory and according to GLP standards.

Previous studies

- In (2008) the researcher (Al-Shakraji, Thanun Yunus) used hierarchical cluster analysis to classify observations into homogeneous groups through the use of complete linkage method on data for players consisting of two groups, each group comprising (48) players, and the results formed (4) clusters For the first group, the second group is also (4) clusters, and each of the clusters of the two groups is formed from a number of homogeneous observations between them [3].
- In (2011) the two researchers (Rashid, Aseel and Mahdi, Nabaa) used hierarchical and non-hierarchical cluster analysis methods in analyzing the reality of education in Iraq, and the results came that the hierarchical method is better than

the non-hierarchical method, and Baghdad governorate surpassed the rest of the governorates in providing better Services in the field of education, while the variable (number of universities) had a great impact on the formation of clusters[4].

- In (2019) the two researchers (Muhammad, Muhammad Abdul-Wadud and Al-Rawi, Ghalib, Asma) used some hierarchical cluster analysis methods to classify the agricultural lands of the governorates of Iraq according to the area and the amount of production for the years 2005 and 2010. Nineveh and the central region governorates are the best agricultural after classification, and the increase in production was weak [5].

Aim Of The Research

The research aims to employ the two methods of cluster analysis in the classification of educational laboratories at the University of Babylon according to the Laboratory Accreditation Quality Form, and to know the most important variables in that form.

The theoretical side

Conditions that must be met when using cluster analysis:

- When the variables are different in their levels of measurement, a standard measurement of those variables must be made.
- Selecting the necessary variables that have an impact on the cluster analysis and not neglecting them.
- The data should not contain outliers as they affect classification accuracy.

There are two types of cluster analysis methods (Nonhierarchical , hierarchical):

Nonhierarchical Clustering Method: There is one method and it is considered the most used method called the K-Means method. It begins by dividing the data into a similar group, then we search about the K of the averages which depend in the clustering process of the data in question. The goal is to measure the convergence of the data with respect to the mean, after which the data is divided into clusters for each medium cluster of its own, thus relying on the variance of the cluster [6].

Hierarchical Clustering Method: There are a number of methods for hierarchical cluster analysis, and each method has certain characteristics in it and differs from other methods. In general, the hierarchical method is worked out according to two methods: (Agglomerative cluster analysis: Each case is in one cluster, after which similar clusters are gradually collected until we reach the required number of clusters, Divisive cluster analysis: All cases are grouped into one cluster, after which these cases are gradually classified into smaller and smaller clusters).

Hierarchical clustering methods do not require prior knowledge of the number of clusters on the basis of which cases will be classified, and among these methods:(Single Linkage, Complete Linkage, Average Linkage, Centroid Clustering, Ward's Method), Ward's Method This method is called the Minimum Variance Method because it relies on analysis of variance to calculate distances between clusters (It is the way we rely on its results).

There are several ways to measure the similarity or difference between two elements, and the appropriate measurement is based on calculating the distance between the two elements in order to determine the degree of closeness and divergence between them, and the greater the distance, the greater the dissimilarity (difference) and vice versa, and the Euclidean distance function is the most used in most methods Cluster is calculated by the following formula[7]:

$$(X, y) = \sqrt{\sum_{i=1}^{n} (x_i - x_j)^2}$$

x_i, x_j: the two elements to calculate the distance between

The practical side

The Laboratory Accreditation Quality Form includes **(11)** variables according to the GLP:

x_1: Leadership, management, and the number of workers on the laboratory staff,
x_2: Securing collective protection systems.
x_3: Civil design for laboratories.
x_4: Training course programs for laboratory personnel.
x_5: Maintenance programs for devices and equipment.
x_6: Lab addresses.
x_7: Certificates of appreciation and letters of thanks to the laboratory workers.
x_8: Individual protection supplies and visual media.
x_9: Securing laboratory safety supplies.
x_{10}: The number of students within the laboratory.
x_{11}: Procedural modalities.

The laboratories of the University of Babylon are the dependent variable (**Y**: Represent the laboratories of the University of Babylon), and its number is **(17)** laboratories:

(**1**:Chemical laboratories, **2**:Physical laboratories, **3**:Medical Laboratory, **4**:Biological laboratories, **5**:Faculty of Nursing laboratories, **6**:Laboratories of the College of Pharmacy, **7**:Laboratories of the Faculty of Materials Engineering, **8**:Faculty of Science laboratories, **9**:Laboratories of the College of Basic Education, **10**:Laboratories of the Faculty of Information Technology, **11**: Laboratories of the College of Medicine, **12**:Faculty of Dental Laboratories, **13**:Laboratories of the College of Education for Pure Sciences, **14**:Museib College of Engineering

laboratories, **15**:Laboratories of the College of Science for Girls, **16**:Laboratories of the College of Engineering, **17**:Other laboratories).

***First*: The nonhierarchical method (K-Means method)**

Table No.(1) shows the distribution of the members of the groups (clusters) and the distance for each item from the center of the group, as it is noticed that the laboratories are clustered in two clusters:

The first cluster includes most of the laboratories, and these laboratories are (2, 4, 5, 6, 9, 10, 12, 13, 15, 16, 17) with a distance between (2.020 - 5.171).

The second cluster includes the rest of the laboratories (1, 3, 7, 8, 11, 14) with a distance between (5.049-2.024).

Table 1. Cluster Membership

Case Number	ID	Cluster	Distance
1	1	2	2.426
2	2	1	2.020
3	3	2	2.265
4	4	1	2.938
5	5	1	2.730
6	6	1	4.136
7	7	2	5.049
8	8	2	2.024
9	9	1	4.279
10	10	1	4.684
11	11	2	4.338
12	12	1	5.171
13	13	S	2.902
14	14	2	3.632
15	15	1	3.643
16	16	1	2.969
17	17	1	2.020

Table No.(2) shows the distances between the centers of the different final clusters, where the distance between the first cluster and the second cluster is (3.064).

Table 2. Distances between Final Cluster Centers

Cluster	1	2
1		3.064
2	3.064	

Table No.(3) shows One-Way Anova for each of the variables using clusters in order to find out the difference between the variables and according to the averages depending on the value of **F** as the average of the squares between groups in the Cluster column, while the average of the squares within the groups in the Error column, it is noticed that the significance of the test function F has no fundamental meaning here and can be neglected, so it should not be used in testing the hypotheses related to the mean of the groups, so we notice that the variable (x_4: Training course programs for laboratory personnel) has the largest differences (**F=12.232**) followed by the variable (x_5: Maintenance programs for devices and equipment) with significant differences (**F=4.125**), and the lowest percentage difference between the groups was for the variable (x_7: Certificates of appreciation and letters of thanks to the laboratory workers) with significant differences (**F=.020**), and for the variable (x_2: Securing collective protection systems) with significant differences (**F=.057**).

Table 3. ANOVA

Var.	Cluster		Error		F	Sig.
	Mean Square	Df	Mean Square	Df		
x_1	.315	1	2.491	15	.126	.727
x_2	.084	1	1.462	15	.057	.814
x_3	.101	1	.771	15	.131	.722
x_4	24.101	1	1.970	15	12.232	.003
x_5	9.718	1	2.356	15	4.125	.060
x_6	.028	1	.039	15	.726	.407
x_7	.015	1	.745	15	.020	.890
x_8	.483	1	.498	15	.970	.340
x_9	.536	1	.376	15	1.425	.251
x_{10}	.162	1	.954	15	.169	.686
x_{11}	.904	1	2.436	15	.371	.551

Table No.(4) shows the numbers of laboratories in the two different clusters, as it shows that most of the laboratories are in the first cluster **(11)** laboratories, and the second cluster **(6)** laboratories.

Table 4. Number of Cases in each Cluster

Cluster	1	11.000
	2	6.000
Valid		17.000
Missing		.000

***Second*: The hierarchical method (Ward Linkage method)**

From Table (5), which shows the steps of agglomeration schedule:

The first step: The word (2), which represents physical laboratories, was linked with the word (17), which represented other laboratories, as it was the least distance between them.

The second step: The word (1: Chemical laboratories) was linked with the word (16: the laboratories of Engineering), and the distance between them was (**4.534**).

The last step: The word (1: Chemical laboratories) was linked with the word (2: Physical laboratories), as it was the largest distance between them.

In general: The chemical laboratories (1) were linked with five laboratories (16, 6, 7, 3, 2) in order, and also the physical laboratories (2) were linked with three laboratories (17, 15, 2) according to the arrangement. Medical (3) with three laboratories (11, 4, 8) in order.

Table 5. Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	2	17	.000	0	0	5
2	1	16	4.534	0	0	6
3	3	11	10.025	0	0	11
4	8	14	16.078	0	0	14
5	2	15	22.322	1	0	13
6	1	6	29.264	2	0	12
7	12	13	36.265	0	0	13
8	4	5	43.823	0	0	9
9	4	9	52.391	8	0	11
10	7	10	65.868	0	0	12
11	3	4	83.568	3	9	14
12	1	7	103.617	6	10	15
13	2	12	125.298	5	7	16
14	3	8	155.195	11	4	15
15	1	3	198.557	12	14	16
16	1	2	247.920	15	13	0

Table 6. Cluster Membership

Case		4 Clusters	3 Clusters	2 Clusters
1:	1	1	1	1
2:	2	2	2	2
3:	3	3	3	1
4:	4	3	3	1
5:	5	3	3	1
6:	6	1	1	1
7:	7	1	1	1
8:	8	4	3	1
9:	9	3	3	1
10:	10	1	1	1
11:	11	3	3	1
12:	12	2	2	2
13:	13	2	2	2
14:	14	4	3	1
15:	15	2	2	2
16:	16	1	1	1
17:	17	2	2	2

From Table (6), which shows the distribution of laboratories as cluster membership, where it shows that the distribution is in four clusters, then in three clusters, and finally in two clusters:

The case of the four clusters: Five laboratories (1, 6, 7, 10, 16) join the first cluster, and five other laboratories (2, 12, 13, 15, 17) join the second cluster, and also five laboratories (3, 4, 5 9, 11) join the third cluster, and the remaining testers (8, 14) join the fourth cluster.

The case of the three clusters: Five laboratories(1, 6, 7, 10, 16) join the first cluster, and five other laboratories (2, 12, 13, 15, 17) join the second cluster, and the rest of the laboratories (3, 4, 5, 8, 9,11, 14) join the third cluster.

The status of the two clusters: The first cluster includes most of the laboratories, while the second cluster includes five laboratories (2, 12, 13, 15 and 17), Figure No. (1) illustrates this more.

In general: Five laboratories (1,6,7,10,16) joined the first cluster, whether in the case of division into four, three, two clusters.

There are also five other laboratories (2, 12, 13, 15, 17) that are themselves joined to the second cluster, whether in the case of division into four, three, two clusters.

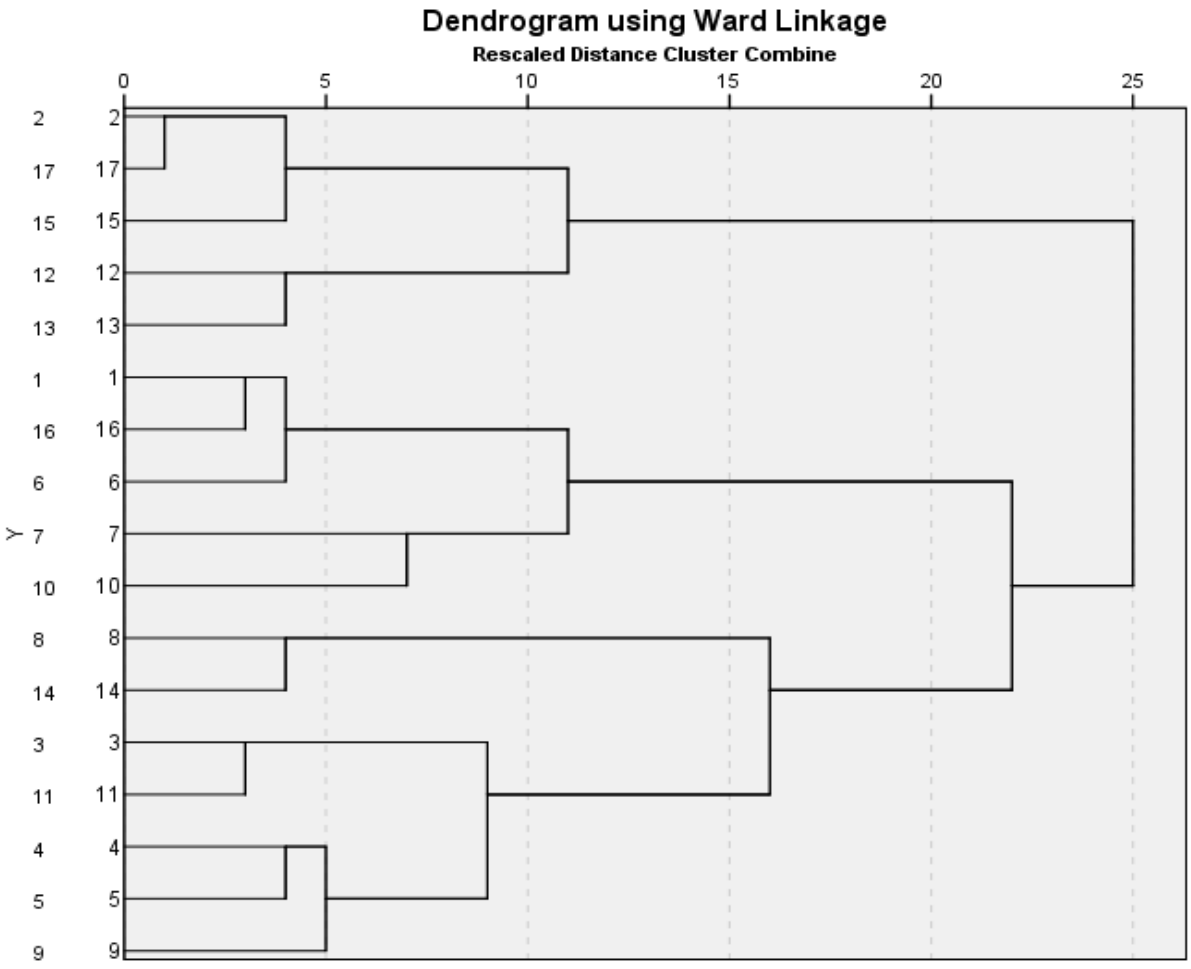

Figure 1. Dendrogram using Ward Linkage

Conclusions

- There are six laboratories (4, 5, 6, 9, 10, 16) that were joined to the first cluster in the two methods.
- The variables with the largest significant differences (x_4, x_5), while the variables have the least significant differences (x_7, x_2).
- Using the hierarchical method, there are five laboratories that join the second cluster in the case of division into two, three or four clusters, in addition to five laboratories of (12) laboratories that join the first cluster in the event of division into two, three or four groups.

Recommendations

- Using other statistical methods (differentiation, factor, ...) and comparing results.
- The use of other methods of hierarchical analysis of the cluster.
- The need to pay attention to educational laboratories in Iraqi universities, and to grant certificates of appreciation to employees of good laboratories.

Conflict of interests.

There are non-conflicts of interest.

References:

[1] Al-Shammari, Muhammad Musa (2020), "Employing the two Methods of Cluster Analysis and Discriminatory Analysis in Classifying Data and Building Discriminatory Functions", *Journal of the Faculty of Education, Al-Azhar University*, Issue 186 / Part One / Egypt.

[2] Hardle, W.,Simar, L. (2003), "Applied Multivariate Statistical Analysis". Berline: Springer.

[3] Alshakerchy, Thanoon (2008), "Using Hierarchical Cluster Analysis to Classify the Observations into Homogenous Groups with Appiication on Basketball Matches", *College of Basic Education Research Journal*, Volume 7, Issue 2/ Iraq.

[4] Rashid, Aseel and Mahdi, Nabaa (2011), "Analyzing the Reality of Education in Iraq Using the Methods of Cluster Analysis", *Al-Qadisiyah Journal for Administrative and Economic Sciences*, Volume 13, Issue 2, Iraq.

[5] Mohammed, Abdulwadood and Dr.Asmaa (2019), "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) ", *Journal of Al-Rafidain University College for Sciences*, Issue 44/ Iraq.

[6] Rencher, A. (2002), "Methods of Multivariate Analysis (2nd). Canada: A Wiley Interscience.

[7] Ashoor, Wafaa (2019), "Classification of Iraqi Governorates by Using Cluster Analysis for 2016", *Journal of Natural, Life and Applied Sciences,* Volume 3 Issue 3/ Iraq.

الخلاصة

في هذا البحث تم استخدام طريقتين للتحليل العنقودي (طريقة غير هرمية/ طريقة K–Means وطريقة ربط هرمي/ وارد) لتصنيف جودة اعتماد المعامل التربوية في كليات جامعة بابل والتي بلغ عددها (17).) المختبرات، واستنتج أن المتغير (x4: برامج الدورة التدريبية للعاملين في المختبرات) له أكبر الفروق (F = 12.232) ويليه المتغير (x5: برامج الصيانة للأجهزة والمعدات) مع وجود فروق ذات دلالة إحصائية (F = 4.125)، بينما أقل متغير له فرق معنوي (x7: شهادات تقدير وخطابات شكر للعاملين في المختبر) مع وجود فروق ذات دلالة إحصائية (F = .020)، بالإضافة إلى وجود ستة مختبرات (4، 5) ، 6 ، 9 ، 10 ، 16) انضم إلى المجموعة الأولى في كلتا الطريقتين.