Iraqi Journal of Statistical Sciences

www.stats.mosuljournals.com

An application of two classification methods: hierarchicalclustering and factor analysis to the plays PUBG

Amira W. Omer Shahen M. Faraj and Soran Husen Mohamad

Statistics and Informatics department- College of Administration & Economics - Salahaddin University – Erbil ,Iraq.
International Business department - College of law and Administrative, University of Halabja – Halabja,Iraq.
Statistics and Informatics department- College of Administration & Economics - Sulaimani University – Sulaimani, Iraq.

Article information

Article history:
Received 3 October , 2022
Accepted 22 November , 2022
Available online 1 June, 2023

Keywords: classification, hierarchical clustering, factor analysis, ROC curve, PUBG, Kurdistan

Correspondence:
Amira W. Omer
Amira.Omer@su.edu.krd

Abstract

The purpose of this study is to compare the results of hierarchical clustering methods and factor analysis in a survey on PUBG play. To achieve this goal, a statistical sample including n = 261 individuals living in Iraqi Kurdistan was selected. These people have completed a researcher- made questionnaire about PUBG game through Google Form and the 35 variables of the questions. The aim of this study is to classify the variables by both method of factor analysis and hierarchical clustering in order to determine the association in their results . The results of comparing the two methods with a Chi-square- Test =115.986 and a df=25 confirmed the significant agreement of the results ($P<0.05$) as well as there is a statistically significant association between the results of centroid linkagehierarchical clustering and factor analysis. Also, the area under the receiver operating characteristic curve (ROC curve) with an overlap of 0.804 confirmed the similarity of the results.

DOI: https://doi.org/10.33899/iqjoss.2023.178680 , ©Authors, 2023, College of Computer and Mathematical Science, University of Mosul ,Iraq.

1.Introduction

Data classification methods are one of the most widely used in statistical methods. The purpose of these methods is to examine their similarities and differences. Classification is the scientific work of data to predict the values of a classified scale (objective or class) by constructing a model based on one or more numbers and / or batches (predictor or attribute).(Shchemeleva, I. I. 2019).

In this article, we first discuss the theoretical foundations of two popular classification methods: factor analysis and clustering. Then, in the third section and under separate sections, we discuss the method of data collection of this research, which is related to the PUBG game. Also, in this section, you see the results of the implementation of two classification methods mentioned on this data. In the following, we will compare the results with the method of cross-tabulation, Chi-square test and ROC curve. In the

final section, the fourth section, we will discuss the results of comparing these two methods and provide practical suggestions. (Granato.D.et al ,2018)

2.Materials and Methods

In this paper, two methods: factor analysis and hierarchical clustering have been used to classification the questionnaire questions related to the characteristics of PUBG openers. The aim of this research is to compare the results of the two methods. In this section, we will survey the details and differences of these two methods.

2-1: Concept of Factor analysis

Factor analysis is a common name for some multivariate statistical methods whose main purpose is to data summarize and classification. This method examines the internal correlation of a large number of variables and finally classifies and explains them in the form of limited general factors. Factor analysis is also a dependent method in which all variables are considered simultaneously, in other words, in this technique, which can be divided into two types, R-type factor analysis: When factors are calculated from the correlation matrix, then it is called R-type factor analysis. Q-type factor analysis: When factors are calculated from the individual respondent, then it said to be Q-type factor analysis, each of the variables is considered as a dependent variable. (Badaruddoza & Brar, S.K., 2015)

Galton was the first to lay the groundwork for factor analysis, followed by Carl Pearson in the early twentieth century, who proposed a method for factor analysis of a multidimensional geometric space, followed by McDonald's in identifying crimes and their relationship to features. Used. Spearman also introduced mathematical models of this method in 1904. With this research, the principles and foundations of factor analysis were formed and it is widely used by various branches of science such as psychology, economics, sociology, management, medicine and so on (Beaujean & Benson, 2019).

The main purpose of factor analysis is to summarize a large number of variables in a limited number of factors, so that we have the least amount of information loss. In addition, factor analysis has various applications in data analysis. (Sharma, 1996)

Achieving dimensions that are hidden in a wide range of variables but are not easily visible. This type of factor analysis is known as R-type factor analysis. Invent a way to combine and summarize a large number of people in different groups within a large community. This method is known as Q-type factor analysis. (Rencher, 2002). Create a small and completely new set of variables that can be completely used instead of the main variables in subsequent regression or diagnostic analysis. (Yong, A. G., & Pearce, S. 2013)

Like other statistical methods, the first step in factor analysis is problem expression. Any type of variable related to the research problem can be used to perform a factor analysis. Raw data should also be quantitative, but sometimes dummy variables (0 and 1) and non-parametric or qualitative can also be used. (Jiang, D., & Kalyuga, S. 2020)

The first and most important point in applying factor analysis is to calculate the correlation matrix. To do this, it must be determined whether the purpose is to calculate the correlation between the variables or between the respondents. If the summation of variables is to be considered, the correlation between the variables must be calculated; in which case the technique used is called type R factor analysis. However, if the purpose of factor analysis is to combine and classify respondents into different groups, the correlation matrix between respondents is calculated and used; this is called Q-type factor analysis. Of course, this method is less used due to its difficulty and instead methods such as cluster analysis or hierarchical grouping are used.(Yang, X. and Han, H., 2017)

One of the methods for selecting appropriate variables in factor analysis is the use of correlation matrix. Since the basis of factor analysis is based on the correlation between variables but of non-causal type, so in this method the correlation matrix between variables is calculated. This matrix shows the relationship between the variables to form clusters so that the variables within each cluster are correlated with each

other but there is no correlation between the variables in different clusters. (Walter, J., Chesnaux,2019)

2.2: Kaiser-Meyer-Olkin test or KMO

Other methods by which the researcher is able to determine the suitability of the data for factor analysis are the Kaiser-Meyer-Olkin test or KMO. The statistical value of this test always varies between 0 and 1. If the value of this statistic is less than 0.5, the data will not be suitable for factor analysis; if its value is between 0.5 to 0.69, factor analysis can be done with more caution, and finally, if the value of this statistic is more than 0.7, we can say that correlation Are present in the data are suitable for factor analysis. Running the Kaiser-Meyer-Olkin (KMO) Test (Hill, B. D.,2011)

The formula for the KMO test is $$MO_j = \frac{\sum_{i \neq j} h_{ij}^2}{\sum_{i \neq j} h_{ij}^2 + \sum_{i \neq j} U_{ij}} \quad (1)$$

Where:

$h_{ij}$is the correlation matrix,

$U_{ij}$is the partial covariance matrix,

Σ = summation notation ("add up")

2-3: Factor Analysis Methods

There are different models in factor analysis, the most widely used of which are the two methods of principal component analysis and common factor analysis. The choice of each of these models depends on the goal of the researcher.

2.3.1:Principal component analysis

The principal component analysis model is used when the goal is to summarize the variables and achieve a limited number of variables for forecasting purposes. In contrast, co-factor analysis is used when the goal is to identify factors or dimensions that are not easily identifiable.(Everitt, 2005). The first principal component, PC1, is defined as the linear combination of the original variables, $x_1, x_2, \ldots . x_k$ that accounts for the maximal amount of the variance of the x variables amongst all such linear combinations. The second principal component, PC2, is defined as the linear combination of the original variables $x_1, x_2, \ldots . x_k$, that accounts for a maximal amount of the remaining variance subject to being uncorrelated with PC2. Subsequent components are defined similarly.(Anderson, 1984)

The matrix of data $X_{n \times m}$contains (m) columns of variables and (n) rows of observations. Matrix of data calculated by correlation matrix and the application of PCA by the correlations matrix to get m of characteristic roots that symbolizes by λ_{ij} in decreasing order $\lambda_1 > \lambda_2 > \cdots > \lambda_k$ which represent variations of summary factors. [(Anderson, 1984), (Afifi and Clark,1984)]. Explain of characteristic vector placed as factors in the linear combination of the original variables to give (PC_{ij}) of the value i of the principal component j

$$PC_{ij} = a_{1j} X_{1i} + a_{2j} X_{2i} + \cdots . a_{kj} X_{ki} = \sum_{i=1}^{k} a_{ij} X_{ik} \quad (2)$$

Where:

PC_{ij} represents the principal component j

a_{ij} represents the coefficient of variable (i) of the component (j) which are values of Eigen

Vectors accompanying of Eigen roots. In the correlation matrix user, the variation j equal to Var (PC_{ij}) = λ_i and λ_i represents a characteristic root of the principal component j. It can be obtained by summary factors to get factors $F_1, F_2, \ldots . F_k$ by dividing each principal component of the standard deviation as follows:

$$F_i = \frac{PC_j}{\sqrt{Var(PC_j)}} = \frac{PC_j}{\sqrt{\lambda_i}} \quad (3)$$

Where:

Fi is a factor (i) and it can be represented as (PC_j) as follows:

$$PC_j = F_i\sqrt{\lambda_i}$$

Transfer Principe Component Model to the Factor Model

$X_i = \sum_{j=1}^{m} a_{ij} PC_j \quad , \; i = 1,2,3, \dots, k \, , \;\; j = 1,2, \dots \dots, m$

$X_i = \sum_{j=1}^{m} a_{ij} F_i \sqrt{\lambda_i} \quad , \; L_{ij} = a_{ij}\sqrt{\lambda_i}$

$$X_i = \sum_{j=1}^{m} F_i L_{ij} \qquad (4)$$

In addition to choosing the analysis model, the researcher must determine how the factors should be extracted. There are two general ways to do this, the orthogonal factor method and the inclined factor method. In the method, it is assumed that each factor is independent of the other factors and the correlation between the factors is contractually considered to be zero. But the inclined factor method assumes that the main variables are correlated and therefore the factors must also have some degree of correlation. (Dombrowski, S. C et, al ,2021)

The choice of each of these methods depends on the objectives of the research, if the goal is to summarize the number of variables, regardless of how significant the results of the extracted factors will be, or if the goal is to form a set of uncorrelated variables to perform regression and pre-methods, the orthogonal method will be a good choice. On the other hand, if the goal is to achieve meaningful factors, the inclined method is suggested. .(Shrestha,N,2021)

One of the important points in factor analysis is determining the number of extractable factors. The factor can usually be derived from the number of variables included in the analysis, but the latter factors usually play a very small role in explaining the issue. So we need to determine the number of factors needed. Although there is no exact basis for this work. The mathematical process used to obtain a factor solution form a correlation matrix is such that each successive factor, each of which is uncorrelated with the other factors, accounts for as much of the variance of the observed variables as possible. (The amount of variance accounted for by each factor is shown by a quantity called the eigenvalue, which is equal to the sum of the squared loadings for a given factor, as will be discussed below). This often means that all the variables have substantial lodgings on the first factor; i.e., that coefficient $a_{11}, a_{21}, \dots . a_{nm}$ are all greater than some arbitrary value each as .3 or .4. while this initial solution is consistent with the aim of accounting for as much and possible of the total variance of the observed variables with as few factors as possible, the initial pattern is often adjusted so that each individual variable has substantial loading on as few factors as possible (preferably only one). After determining the number of factors to be extracted, by interpreting the significance of factor loads, these factors are interpreted. (Badaruddoza & Brar, S.K, 2018)

3:Cluster Analysis

3.1: Cluster analysis

Cluster analysis is a statistical method for grouping data or observations, according to their similarity or degree of proximity. Through cluster analysis, data or observations are divided into homogeneous and distinct categories. This method is used to segment customers based on their similarities. Often in cluster analysis, decisions about the number of clusters are made based on the Bayesian criterion and the Akaike criterion which used to determine the suitable method. An answer obtained at the level of at least the Bayesian criterion and the can represent the best balance between accuracy and complexity, which considers the most important effects and does not underestimate their importance(Jarman, A. M. 2020). Also, another way to decide on the number of clusters is to use the distance ratio. The optimal number of clusters is when a large change in distance ratio is observed.(Veletić, J., & Olsen, R. V. 2021)

The term cluster analysis was first used by (Tryon in 2019) for the methods of a group of objects that were similar. Cluster analysis is a shortcut tool for data analysis that aims to organize different objects into groups whose maximum degree of connection between two objects is maximum and otherwise minimum if they belong to the same group. In other words, cluster analysis shows the structure of data without explaining what exists (Tryon, 2019).

3.2: Clustering Methods

In clustering analysis, grouping of a set of objects is done in such a way that the objects in a group (called clusters) are more similar than other categories (clusters). This is the main task of exploratory data mining and is a common method for analyzing statistical data that is used in many fields including machine learning, pattern recognition, image analysis, data retrieval, bioinformatics, data compression and computer graphics. In addition, cluster analysis itself is not a specific algorithm, but a general process and can be obtained by different algorithms that understand what constitutes a cluster and how they work. .(Knote, R.,2019)

There are several ways to do clustering. One of the most widely used of these methods is the hierarchical method (Veletić, J., & Olsen, R. V. 2021). This method, which works based on the distance matrix, has many options that help the researcher to choose. In this research, we applied the hierarchical clustering method. The details of this method will be detailed in the following sections.(Briggs, C,et al,2020)

3.2.1: Hierarchical Cluster Analysis

Hierarchical clustering refers to the way in which observations and data are categorized and grouped hierarchically. There are points that set this method apart from other clustering methods, and there is a top-down (or bottom-up) look at this technique. One of the most widely used statistical methods, known as "unsupervised learning" is clustering analysis which applies clustering methods to explore data and find hidden patterns or groupings in data. In this method, unlike k-mean clustering, each observation may be in more than one cluster because clusters are formed based on different levels of distance. Therefore, each cluster may be subset of another cluster at a distance level. Clustering, however, is a method that classifies observations into similar groups using "Features" or "Attributes" of observations. (Wu, C.,et al , 2021)

Choosing the right features for this job is one of the important issues to consider. On the other hand, data standardization is also proposed so that the measurement scale of the attribute or attribute does not cause the distance function to deviate. Due to the high computational burden of clustering methods, the use or creation of techniques that can generate clustering responses with appropriate accuracy in a shorter time is also one of the most recent research topics in machine learning, especially in the big data space. (Zheng, W. ,2022)

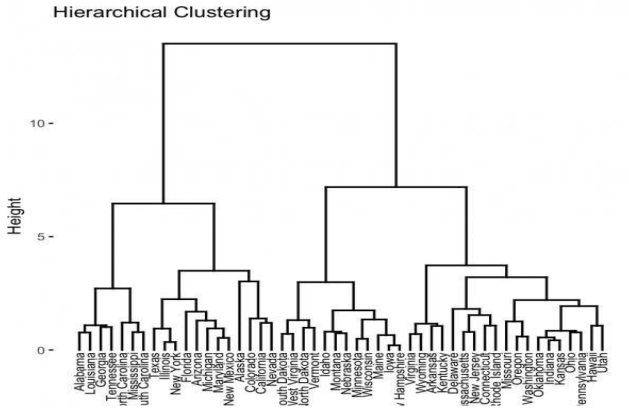


Figure 1. dendrogram in clustering show the clusters

Hierarchical clustering with integrative method: If the view of this chart is from bottom to top (Bottom-Up), according to the height of the chart (Height) at the lower level, the clusters are a subset of higher level clusters, so it seems that the lower clusters are combined. And create higher level clusters. This method of hierarchical clustering is known as the "agglomerative" method. This method is usually abbreviated to HAC, which stands for Hierarchical Clustering.(Nielsen, F. 2016)

Hierarchical clustering by division method: Conversely, if the view is top-down, the top clusters are broken down into other sub-clusters until we reach clusters with only one member. In this way, the largest cluster, which includes all observations, is divided into the smallest clusters, which contain only one observation. This method is called "Divisive". Since this method has limitations, it is less used in hierarchical clustering. Therefore, in this paper, we introduce the aggregation method, but we get acquainted with the commands in statistical software that perform division clustering.(jarman, A. M. 2020)

The time complexity of the integrated hierarchical clustering in the HAC algorithm, the time complexity is equal to smallest distance between two pints o(n2) and the memory space required is equal to o(n2). Therefore, as the volume of data increases, the speed and memory space for performing clustering operations increases dramatically. For this reason, this algorithm is not usually used for "Big Data" clustering.(Yildirim, P., & Birant, D. 2017)

To perform the calculations related to this clustering method, we need two distance (similarity) criteria. The amount of distance between pairs of observations. The distance between the clusters. In the first case, distance functions can be used for quantitative or qualitative data. (Wu et al., 2021).

So if the data is small, for example, the Euclidean distance or the Manhattan distance can be used. For qualitative data, simple matching or "Hamming Distance" for data can also be used.(Ma, Y.et al ,2021)

Usually, a "Distance Matrix" or "Similarity Matrix" is used to speed up computations before starting the aggregation hierarchical clustering process. This matrix shows the distance between each pair of observations. Of course, the type of function by which the distance should be measured affects the values in this matrix.(Xu, D., & Tian, Y. ,2015)

In HAC clustering, according to the values of this matrix, the observations or clusters that have the least distance (most similarity) are merged to form a new cluster. In the next step, the distance between the new observations or clusters is calculated by the updated distance matrix and the integration work continues so that only one cluster remains (Yildirim, P., & Birant, D. 2017)

The table 1. introduces the method of calculating the distance between observations for quantitative data:

Table 1. Methods to calculate distance between two observations

The function	Formula
Euclidean Distance	$\|x_i - x_j\| = \sqrt{(a_i - a)^2 (b_i - b_j)^2}$
Manhattan Distance	$\|x_i - x_j\| = \|a_i - a_j\| + \|b_i - b_j\|$
Maximum Distance	$\|x_i - x_j\| = \max(\|a_i - a_j\| + \|b_i - b_j\|)$
Mahalanobis distance , where S is the covariance matrix and x_i and x_j are variables vector of x_i and x_j	$\sqrt{(x_i - x_j)^T S^{-1} (x_i - x_j)}$

x_i, x_j : are ith and jth observation, where I and j are indicated, a and b are feature variables,

As a comparison between these distance functions, it can be seen that Euclidean distance has been used more in studies related to psychology, computer science and business and other fields. (Sharma, S.1996)

It should be noted, however, that the use of distance between pairs of observations is used in most clustering methods. But the point that distinguishes hierarchical clustering from other methods is the measurement of the distance between clusters. In this way, the two clusters that are most similar (least distant) to each other are merged to form a new cluster. So at each stage it is only possible to combine two clusters. These steps are known as "Merge Levels". (Yim, O and Ramdeen, K. T. 2015)

In the following, we will examine the methods of measuring the distance between clusters, which, of course, are best used in hierarchical clustering. Various criteria can be used to measure the distance between clusters. For example, the distance can be calculated based on the distance between the nearest

or farthest observations between two clusters. Each of these criteria has its own advantages and disadvantages. However, according to the data structure (existence of outliers), the pattern of placement (dispersion) of observations in each cluster and… can be the basis for choosing one of the link methods is (Centroid) because this method has minimum AIC and BIC with other methods. (Shen, J. J. 2007). (Christopher .D. et al ,2008)

Table 2. Some linkage methods

The name of the linkage method	**Formula**
Complete-Linkage	$\max_{ij}\{d(a_i, b_j)\}$
Single-Linkage	$\min_{ij}\{d(a_i, b_j)\}$
UPGMA (Unweighted Pair Group Method with Arithmetic Mean)	$\frac{1}{\|a\|\|b\|}\sum_{i=1}^{a}\sum_{j=1}^{b} d(a_i, b_j)$
Centroid	$d(a_i, b_j)$
Ward	Calculation base on objective function and minimization of variance of hybrid clusters

$a_1, a_2, , a_k$ = Observations from cluster 1
$b_1, b_2, , b_k$ = Observations from cluster 2
$d(a_i, b_j)\}$= Distance between a subject with observation vector (a) and a subject with observation vector (b)
$\|.\|$ = Euclidean norm

Centroid based linkage approach

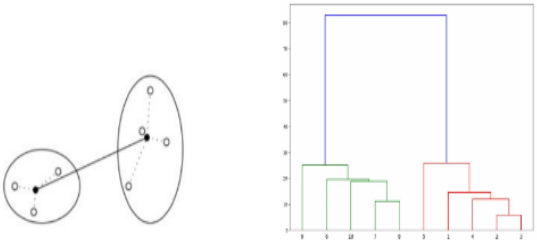
The basic idea of centroid linkage method is to take the distance between the centroids of the data points in clusters. If among the pair of clusters the first one have points p, q, r, s, t and the second one has the points w, x, y, z then to find the distance between the clusters would be the distance between the centroid found for the data points (p,q,r,s,t) and centroid found for the data points (w,x,y,z). Unlike above single, complete and average linkage method, the distance is calculated once rather than between each and every points of the clusters. Which is shown in figure 2(a). The formula to calculate the centroid of a finite set of k points $x_1, x_2, \ldots x_n$ is straightforward

$$C = \frac{x_1, x_2, \ldots x_n}{k} \quad (6)$$

In most cases, the points will be n-dimensional and the centroid should be calculated as taking the points as vertices in a simplex and the formula would be if the n vertices are: $v_1, v_2, \ldots v_n$ which are vectors . (Protter, M. H., & Morrey, C. B. ,1977)

$$C = \frac{1}{n+1}\sum_{i=0}^{n} V_i \quad (7)$$

Centroid linkage clustering results somewhat similar to average linkage clustering but centroid linkage method has a bad characteristics: possibility of inversion and that's why it is dangerous to use in hierarchical clustering which needs further research (Christopher .D. et al ,2008). The resulting dendrogram found after applying centroid linkage method is shown in figure 2(b)



2(a) 2(b)

Figure 2 : Centroid linkage clustering

4. STATISTICAL ANALYSIS

4-1: The Data Collection

The sample of this study includes n = 261 observation of the Iraqi Kurdistan region. This was from a population of residents who play PUBG and administered to Google through a questionnaire containing questions related to demographic features and 35 questions related to PUBG game. The data were analyzed using (SPSS24) software in two parts: descriptive and inferential. In the next section, we will explain in detail the results as well as the comparison results of the classification methods.

4-2- Data analysis

In this section, first provide a brief descriptive report of the respondents' situation. Then we will examine the results of comparing classification methods.

4-2-1- Descriptive Analysis

The respondents were 261 residents of the Iraqi Kurdistan region who answered a researcher- made questionnaire about the PUBG computer game through Google Form and a simple random selection. How to distribute the frequency of age, gender, etc. of the respondents is as figure (3,4):

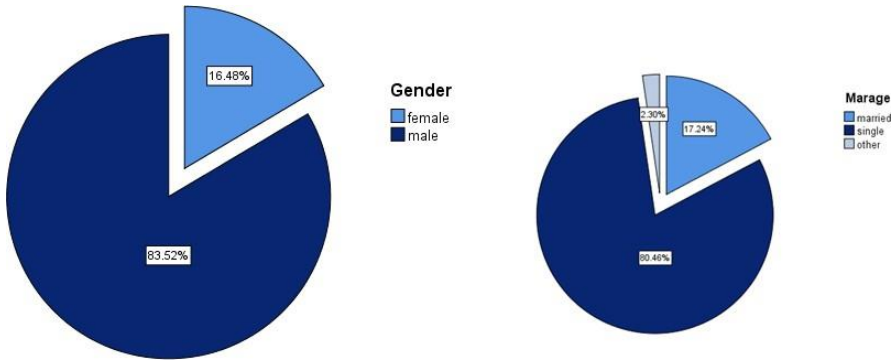
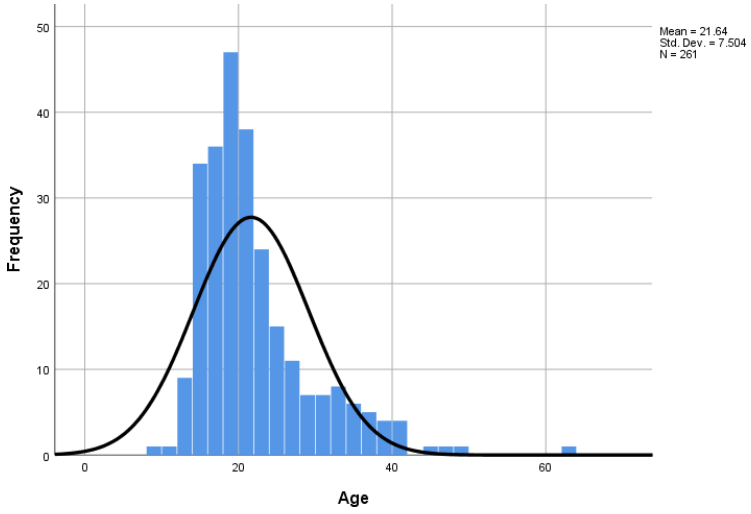


Figure 3. The pie chart for gender and marriage frequency of subjects

Figure 4. The normality curve



As shown in the diagrams above, most of the respondents were single and young boys with an average age of less than 22 years. The small number of older respondents means that the age distribution of the data is right-skewed.

4-2-2: Inferential Analysis

In the data analysis step, we implemented 35 variables resulting from the implementation of the PUBG questionnaire among 261 people in two methods of classification; hierarchical clustering and factor analysis.

Hierarchical Cluster Analysis

In this part of the study, the variables was descripted in table (7), we used 35 variables by hierarchical clustering method and by applying the Euclidean distance matrix to find the smallest distances and also Centroid linkage method to calculate new distances for clusters resulting from integration. As can be seen in the results shown in the figure(5), these variables can be classified into 6 main clusters. In the following sections, after factor analysis, we will compare these results

Figure 5. The dendrogram for hierarchical clustering Centroid linkage

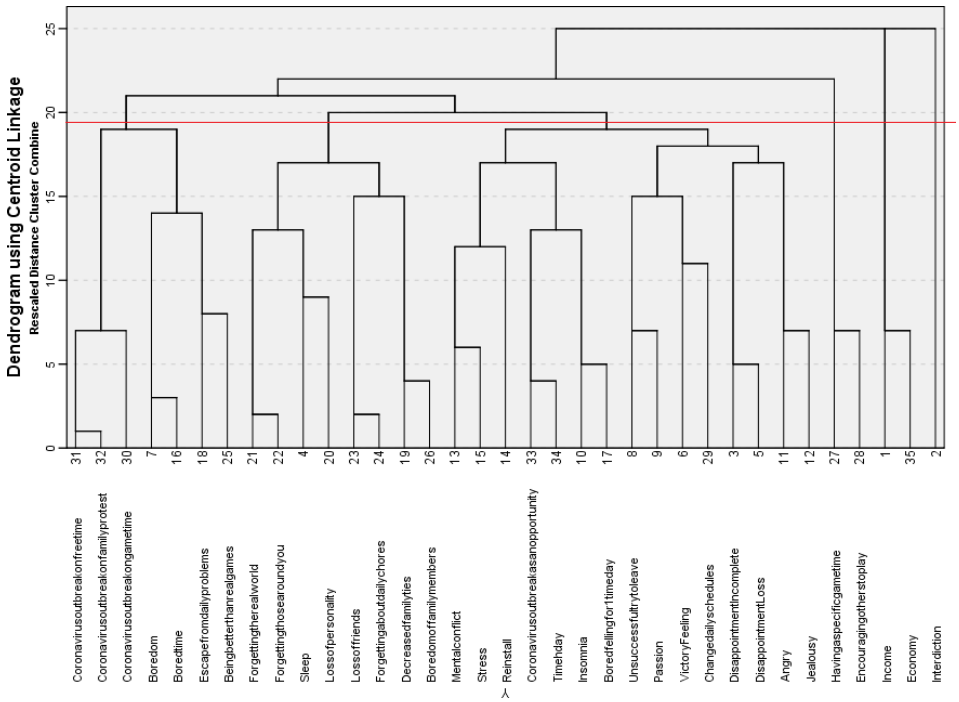


Table 3. Hierarchical clustering Centroid linkage results

Agglomeration Schedule

Stage	Cluster Combine		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	33	34	.809	0	0	10
2	23	24	.750	0	0	19
3	25	26	.735	0	0	23
4	9	18	.690	0	0	21
5	21	28	.661	0	0	23
6	3	35	.649	0	0	20
7	12	19	.632	0	0	20
8	5	7	.625	0	0	25
9	15	17	.582	0	0	18
10	32	33	.552	0	1	28
11	13	14	.527	0	0	25
12	29	30	.521	0	0	32
13	1	2	.520	0	0	33
14	10	11	.513	0	0	22
15	20	27	.500	0	0	21
16	6	22	.458	0	0	19
17	8	31	.377	0	0	22
18	15	16	.312	9	0	26
19	6	23	.281	16	2	24
20	3	12	.268	6	7	26
21	9	20	.243	4	15	28
22	8	10	.214	17	14	27
23	21	25	.211	5	3	24
24	6	21	.127	19	23	30
25	5	13	.126	8	11	27
26	3	15	.089	20	18	29
27	5	8	.046	25	22	29
28	9	32	.033	21	10	31
29	3	5	.022	26	27	30
30	3	6	-.008	29	24	31
31	3	9	-.042	30	28	32
32	3	29	-.096	31	12	33
33	1	3	-.212	13	32	34
34	1	4	-.251	33	0	0

Factor Analysis

Based on the same data and 35 variables, we performed factor analysis using varimax method. The results are as follows:

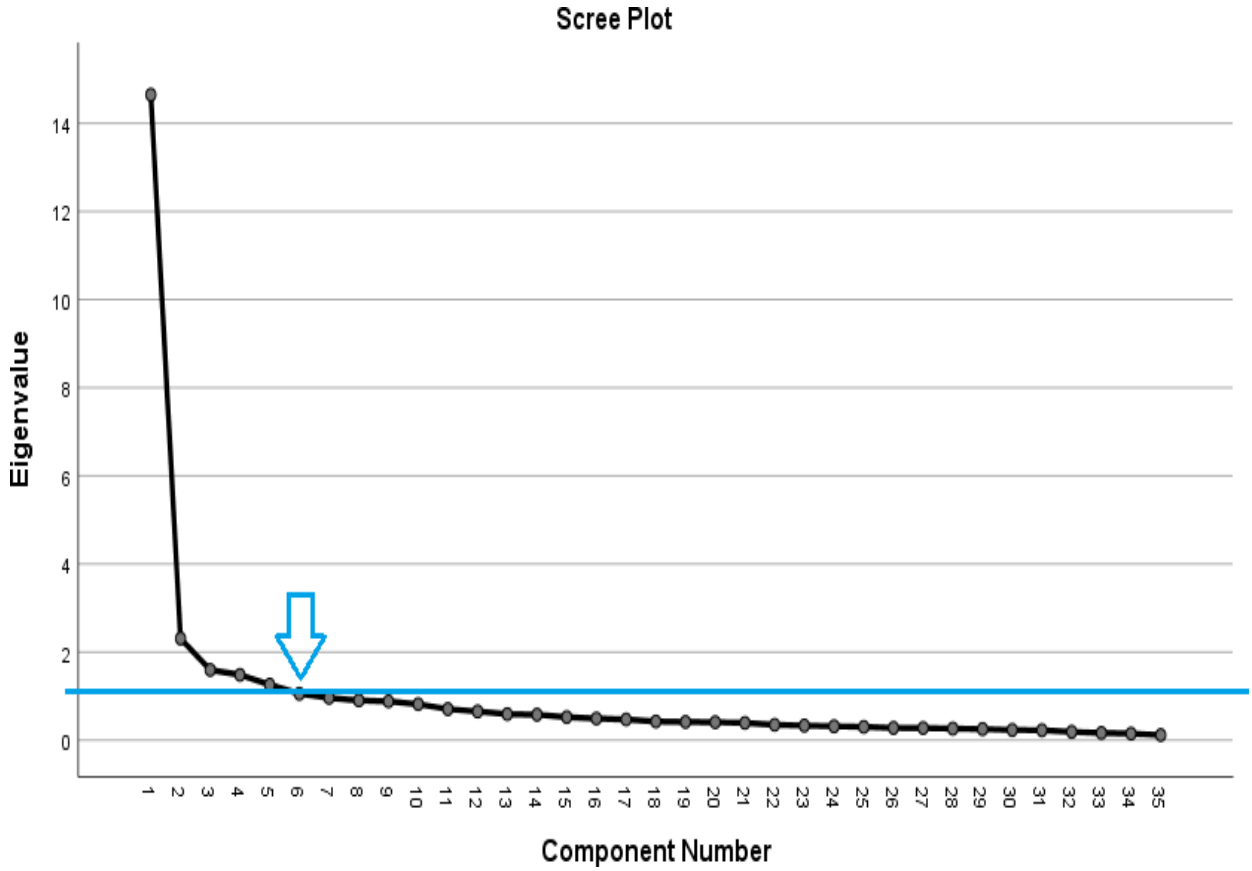
First, based on the results of KMO test (Table 4), it can be seen that the adequacy of the sample size for factor analysis is confirmed. (P <0.05)

Table 4. KMO test

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.942
Bartlett's Test of Sphericity	Approx. Chi-Square	5818.82
	Df	595
	Sig.	.000

It can also be found from the scree plot (Fig. 6) that according to the approximate smoothing of the diagram at point 6, the number of principal components can be 6 factors. However, later in the selection stage, we will see the following variables of each factor, which will have only 3 main factors in the subset.

Figure 6. The Scree plot for recognizing the number of factors



In the Table 5 related to the portion of controlling the variation of the principal factors, it can be seen that these 6 factors will control 63.847% of the total variance.

Table 5. Total Variance Explained by factors

	Total ained					
	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	14.648	41.852	41.852	14.648	41.852	41.852
2	2.307	6.592	48.443	2.307	6.592	48.443
3	1.594	4.555	52.998	1.594	4.555	52.998
4	1.483	4.237	57.235	1.483	4.237	57.235
5	1.260	3.601	60.836	1.260	3.601	60.836
6	1.054	3.011	63.847	1.054	3.011	63.847
⋮	⋮	⋮	⋮			
⋮	⋮	⋮	⋮			
⋮	⋮	⋮	⋮			
⋮	⋮	⋮	⋮			
35	.107	0.1561	100.00			

In the table of components matrix (Table 6) based on the factor load of each variable in the main factors and of course by removing the weak coefficients (between -0.3 to 0.3), we determined in which factor each variable is located.

Table 6. Component Matrix.

Component Matrix[a]							
	Component						
	1	2	3	4	5	6	Factor
Economy					.634		5
Income					.727		5
Time(h/day)	.787						1
Interdiction						.481	6
Disappointment/Incomplete			.661				3
Sleep	.702						1
Disappointment/Loss					.457		5
Victory/Feeling			.633				3
Boredom	.587						1
Unsuccessful try to leave			.465				3
Passion			.667				3
Insomnia			.747				3
Angry			.656				3
Jealousy			.621				3
Mental conflict			.747				3
Reinstall						.453	6
Stress			.694				3
Bored time	.633						1
Bored felling for 1time/day	.777						1
Escape from daily problems	.610						1
Decreased family ties		.734					2
Loss of personality		.636					2
Forgetting the real world		.743					2
Forgetting those around you		.751					2
Loss of friends		.707					2
Forgetting about daily chores		.768					2
Being better than real games	.706						1
Boredom of family members	.748						1
Having a specific game time				.669			4
Encouraging others to play				.606			4
Change daily schedules	.660						1
Coronavirus outbreak on game time	.700						1
Coronavirus outbreak on free time	.753						1
Coronavirus outbreak on family protest	.742						1
Coronavirus outbreak as an opportunity			.761				3
Extraction Method: Principal Component Analysis.							
a. 6 components extracted.							

A comparison of the results of factor analysis and cluster analysis

The results of the two methods of factor analysis classification and hierarchical clustering are shown in Table 7.

Table 7. Comparison between factor analysis and cluster analysis

Variables	Factor Analysis	Clustering	Results
Economy	5	5	No difference
Income	5	5	No difference
Time(h/day)	1	3	Difference
Interdiction	6	6	No difference
Disappointment/Incomplet	3	3	No difference
Sleep	1	2	Difference
Disappointment/Loss	5	3	Difference
Victory/Feeling	3	3	No difference
Boredom	1	1	No difference
Unsuccessful try to leave	3	3	No difference
Passion	3	3	No difference
Insomnia	3	3	No difference
Angry	3	3	No difference
Jealousy	3	3	No difference
Mental conflict	3	3	No difference
Reinstall	6	3	Difference
Stress	3	3	No difference
Bored time	1	1	No difference
Bored felling for 1time/day	1	3	No difference
Escape from daily problems	1	1	No difference
Decreased family ties	2	2	No difference
Loss of personality	2	2	No difference
Forgetting the real world	2	2	No difference
Forgetting those around you	2	2	No difference
Loss of friends	2	2	No difference
Forgetting about daily chores	2	2	No difference
Being better than real games	1	1	No difference
Boredom of family members	1	2	Difference
Having a specific game time	4	4	No difference
Encouraging others to play	4	4	No difference
Change daily schedules	1	3	difference
Coronavirus outbreak on game time	1	1	No difference
Coronavirus outbreak on free time	1	1	No difference
Coronavirus outbreak on family protest	1	1	No difference
Coronavirus outbreak as an opportunity	3	3	No difference

The results of matching these two methods are given in the cross-frequency table (Table 8). Although there is a visual correlation of the results, it is better to use the Chi-square hypothesis test for this purpose.

Table 8. Association between results of Factor Analysis and Clustering Crosstab

Factor Analysis * Clustering Crosstabulation								
Count		Clustering						Total
		1.00	2.00	3.00	4.00	5.00	6.00	
Factor Analysis	1.00	7	2	3	0	0	0	12
	3.00	0	6	0	0	0	0	6
	2.00	0	0	10	0	0	0	10
	4.00	0	0	0	2	0	0	2
	5.00	0	0	1	0	2	0	3
	6.00	0	0	1	0	0	0	2
Total		7	8	15	2	2	1	35
Pearson Chi-Square			115.986		df	25	p-value	0.000

As in the Table 8 can be seen that the association between results of factor analysis and cluster analysis. As a results shows that there is a statistically significant association between results of factor analysis and clustering because the result of p-value was less than the common alpha 0.05 Therefore, the conformity of the results can be accepted.

Another way to determine the accuracy of different classification methods is to use the ROC curve. In Fig. 6 it can be seen that the curve tends to the left and up so that the area under the curve is 0.804 (Table 9). Therefore, the degree of overlap between the two methods in the correct diagnosis of classes is very high.

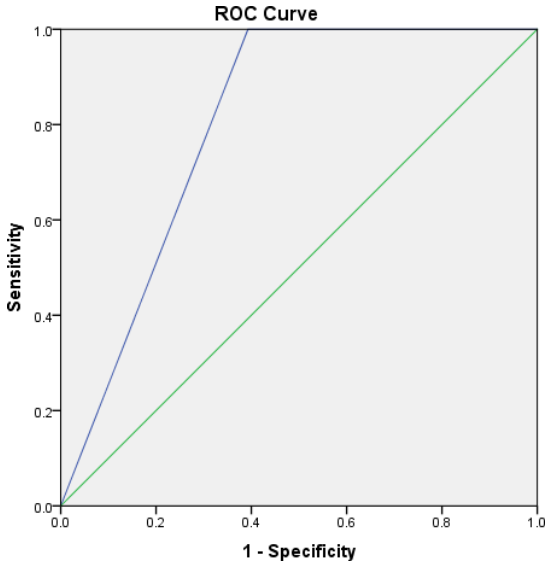


Figure 7. The ROC curve for results

Table 9. Area under the curve

Test Result Variable(s): Factor Analysis
Area
.804

5.CONCLUSIONS AND RECOMMENDATIONS

In this section, we will provide general results and some practical suggestions for all those who are interested in doing statistical work based on the classification of variables.

5-1- Conclusions

As observed in the third section, there is a significant association between the result of Centroid linkage hierarchical clustering and factor analysis by Varimex method. Therefore, researchers can safely choose one of these methods. However, since hierarchical clustering is of course much simpler with the Centroid linkage and with a more detailed dendrogram diagram as well as the possibility of selecting the threshold to change the number of main clusters, the hierarchical clustering method is proposed. And also the results of the factor analysis and cluster analysis shows that there are six significant factors and six important clusters from which to determine the variables affecting migration and the ratio explain is 63.847% of the total variance. This analysis shows there is a quite difference between the results of factor analysis and cluster analysis which the number of factors equal to the number of clusters, as well as in terms of inclusion, the factor analysis and cluster analysis have the same variables in each cluster and factor with the same sequence except the factor and cluster but (6) variables was difference place of cluster and factor with the same sequence except the factor and cluster .

5-2- Recommendations

As stated in the previous section, since there is no significant difference between the results of the two methods, the researcher is free to choose each method based on his interest. We suggest to researchers that the comparisons to be made between other methods of cluster analysis and the other ways the factor analysis. And also we recommend the possibility of using factor analysis in the classification gives the difference results as cluster analysis, especially when the application analyzes

6.References

1. Afifi, A. A., & Clark, V. (1984). Computer aided multivariate analysis lifetime learning publications belmont.
2. Anderson, T. W., & Anderson, T. W. (1958). An introduction to multivariate statistical analysis (Vol. 2). New York: John Wiley & Sons.
3. Badaruddoza and Brar, S.K., 2015. Factor analysis of traditional cardiovascular risk traits in Punjabi adolescents in India. Egyptian Journal of Basic and Applied Sciences, 2(1), pp.13-18
4. Badaruddoza and Brar, S.K., 2015. Factor analysis of traditional cardiovascular risk traits in Punjabi adolescents in India. Egyptian Journal of Basic and Applied Sciences, 2(1), pp.13-18.
5. Bandalos, D. L., & Finney, S. J. (2018). Factor analysis: Exploratory and confirmatory. In The reviewer's guide to quantitative methods in the social sciences (pp. 98-122). Routledge.
6. Beaujean, A. A., & Benson, N. F. (2019). The one and the many: Enduring legacies of Spearman and Thurstone on intelligence test score interpretation. Applied Measurement in Education, 12(1), 198-215.
7. Briggs, C., Fan, Z., & Andras, P. (2020, July). Federated learning with hierarchical clustering of local updates to improve training on non-IID data. In 2020 International Joint Conference on Neural Networks (IJCNN) (pp. 1-9). IEEE
8. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze,Web book (https://nlp.stanford.edu/IR-book/html/htmledition/centroid-clustering-1.html). Introduction to Information Retrieval, Cambridge University Press. 2008, Table of Contents (Hierarchical clustering)
9. Dombrowski, S. C., McGill, R. J., Canivez, G. L., Watkins, M. W., & Beaujean, A. A. (2021). Factor analysis and variance partitioning in intelligence test research: Clarifying misconceptions. Journal of Psychoeducational Assessment, 39(1), 28-38.
10. Everitt, B., Fienberg, S., Olkin, I., & Casella, G. (2005). An R and S-PLUS companion to multivariate analysis (No. 519.5 E8.). London: Springer.
11. Granato, D., Santos, J. S., Escher, G. B., Ferreira, B. L., & Maggio, R. M. (2018). Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between

bioactive compounds and functional properties in foods: A critical perspective. Trends in Food Science & Technology, 72, 81-90.

12. Hill, B. D. (2011). The sequential Kaiser-Meyer-Olkin procedure as an alternative for determining the number of factors in common-factor analysis: A Monte Carlo simulation. Oklahoma State University.
13. Jarman, A. M. (2020). Hierarchical cluster analysis: Comparison of single linkage, complete linkage, average linkage and centroid linkage method. Georgia Southern University.
14. Jiang, D., & Kalyuga, S. (2020). Confirmatory factor analysis of cognitive load ratings supports a two-factor model. Tutorials in Quantitative Methods for Psychology, 16, 216-225.
15. knote, R., Janson, A., Söllner, M., & Leimeister, J. M. (2019). Classifying smart personal assistants: An empirical cluster analysis
16. Ma, Y., Lin, H., Wang, Y., Huang, H., & He, X. (2021). A multi-stage hierarchical clustering algorithm based on centroid of tree and cut edge constraint. Information Sciences, 557, 194-219.
17. Nielsen, F. (2016). Hierarchical clustering. In Introduction to HPC with MPI for Data Science (pp. 195-211). Springer, Cham
18. Protter, M. H., & Morrey, C. B. (1977). College calculus with analytic geometry. Addison-Wesley.
19. Rencher, A. C. (2002). Methods of Multivariate Analysis, 2nd edn., A John Wiley & Sons. New York
20. Sharma, S. (1996). Applied multivariate techniques.
21. Shchemeleva, I. I. (2019). Social activity of the student youth: Factor and cluster analysis. Sotsiologicheskie issledovaniya, (4), 133-141.
22. Shen, J. J. (2007). Using cluster analysis, cluster validation, and consensus clustering to identify subtypes of pervasive developmental disorders (Doctoral dissertation, M. Sc. thesis of Queen's University, Kingston, Ontario, Canada).
23. Shrestha, N. (2021). Factor analysis as a tool for survey analysis. American Journal of Applied Mathematics and Statistics, 9(1), 4-11.
24. Tryon, R.C. (2019) Cluster Analysis: Correlation Profile and Orthometric (Factor) Analysis for the Isolation of Unities in Mind and Personality. Edwards Brothers, Ann Arbor.
25. Veletić, J., & Olsen, R. V. (2021). Exploring school leadership profiles across the world: a cluster analysis approach to TALIS 2018. International Journal of Leadership in Education, 1-27
26. Walter, J., Chesnaux, R., Gaboury, D. and Cloutier, V., 2019. Subsampling of Regional-Scale Database for improving Multivariate Analysis Interpretation of Groundwater Chemical Evolution and Ion Sources. Geosciences, 9(3), p.139
27. Wu, C., Peng, Q., Lee, J., Leibnitz, K., & Xia, Y. (2021). Effective hierarchical clustering based on structural similarities in nearest neighbor graphs. Knowledge-Based Systems, 228, 107295
28. Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. Annals of Data Science, 2(2), 165-193.
29. Yang, X. and Han, H., 2017. Factors analysis of protein O-glycosylation site prediction. Computational biology and chemistry, 71, pp.258-263.
30. Yildirim, P., & Birant, D. (2017). K-linkage: A new agglomerative approach for hierarchical clustering. Advances in Electrical and Computer Engineering, 17(4), 77-88.
31. Yim, O., & Ramdeen, K. T. (2015). Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data. The quantitative methods for psychology, 11(1), 8-21.
32. Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. Tutorials in quantitative methods for psychology, 9(2), 79-94.
33. Zheng, W. (2022). Cluster Analysis Algorithm in the Analysis of College Students' Mental Health Education. Applied Bionics and Biomechanics, 2022

تطبيق طريقتين للتصنيف: التجميع الهرمي وتحليل العوامل لمسرحيات PUBG

اميرة وائل عمر ، شاهين محمد فرج ، سوران حسين محمد
قسم الاحصاء والمعلوماتية – كلية الادارة والاقتصاد – جامعة صلاح الدين – اربيل – العراق.
قسم الأعمال الدولية – كلية القانون والإدارة ، جامعة حلبجة – حلبجة ، العراق.
قسم الاحصاء والمعلوماتية – كلية الادارة والاقتصاد – جامعة السليمانية – السليمانية – العراق.

الخلاصة:
الغرض من هذه الدراسة هو مقارنة نتائج طرق التجميع الهرمي وتحليل العوامل في مسح على لعبة PUBG. ولتحقيق هذا الهدف تم اختيار عينة إحصائية تضم ن = 261 فردا يعيشون في كردستان العراق. لقد أكمل هؤلاء الأشخاص استبيانًا من إعداد الباحث حول لعبة PUBG من خلال نموذج Google والمتغيرات الخمسة والثلاثين للأسئلة. الهدف من هذه الدراسة هو تصنيف المتغيرات من خلال كل من طريقة تحليل العوامل والتكتل الهرمي من أجل تحديد الارتباط في نتائجها. أكدت نتائج مقارنة الطريقتين مع اختبار Chi–square– = 115.986 و df = 25 التوافق المعنوي للنتائج (P <0.05) بالإضافة إلى وجود ارتباط ذي دلالة إحصائية بين نتائج التكتل الهرمي لربط النقط الوسطى وتحليل العوامل. أيضًا ، أكدت المنطقة الواقعة أسفل منحنى خاصية تشغيل المستقبل (منحنى ROC) مع تداخل قدره 0.804 تشابه النتائج.

الكلمات الدالة: لتصنيف ، التجميع الهرمي ، التحليل العاملي ، منحنى ROC، PUBG ، كردستان