STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **16**, July 2026, pp 716–731.
Published online in International Academic Press (www.IAPress.org)

Cluster Analysis of Traffic Accident Patterns in Iraq: An Exploratory Statistical Study

Hayder Majeed Hammadi

Department of Mathematics, College of Basic Education, University of Misan, Misan, Iraq

Abstract Traffic accidents are considered one of the serious societal phenomena that are no less impactful than security threats such as terrorism, due to the significant losses they cause in lives and property, as well as the depletion of economic and human resources. This phenomenon has become a real challenge to sustainable development efforts, necessitating its analysis using advanced statistical methods to understand its dimensions, identify its patterns, and determine the factors influencing it. Based on this importance, this research reviews the analysis of traffic accidents in Iraqi provinces relying on data from the Central Statistical Organization for the year 2024 and in coordination with the Ministry of Interior, with the aim of providing a quantitative perspective that helps explain spatial variations and guide future traffic policies. Cluster analysis was employed using the statistical software (SPSS Version 23) to explore the structural homogeneity among the provinces based on a set of relevant qualitative and quantitative variables. The researcher used three methods in the practical aspect: (1- Single linkage method, 2- Complete linkage method, 3- Ward's method). These methods were chosen because they yield less stringent clusters, especially Ward's method, which reduces variance when merging groups. The analysis of accidents was based on the nature of the accident (collision, overturning, run-over) as it is more accurate and available comprehensively for all provinces, whereas choosing fatalities or injuries might be linked to other reasons outside the current analysis scope or the type of road, which lacks temporal standardization or spatial accuracy in official records. The analysis was conducted with the aim of classifying the provinces into homogeneous clusters that reveal statistical variations in the pattern of accidents between different regions, thereby enabling the possibility of providing realistic recommendations based on scientific foundations to reduce these accidents or mitigate their impacts. .

Keywords Traffic accidents, Cluster analysis, Iraq, SPSS, Statistical modeling

DOI: 10.19139/soic-2310-5070-3104

1. Introduction

Traffic accidents are one of the most significant problems faced by some Iraqi provinces. There are several factors influencing the increase in accident rates, resulting in damages: the first damage is human harm, whether it be death or injuries, and the second damage is material damage, with financial losses occurring despite the efforts made by the relevant authorities, yet accidents remain high. Addressing this problem should be based on scientific principles and involve various efforts and specializations. After World War II, the world witnessed an increase in the number of vehicles, both civilian and military, as vehicles play an active role in human life by being used for all life needs, which significantly contributed to the reliance on vehicles due to their convenience. But the complete reliance on private vehicles has led to a number of traffic and environmental problems, especially in large cities, the most prominent of which are traffic congestion and traffic accidents. The main cause must be diagnosed using quantitative methods in the analysis by interpreting statistical figures. It is known that accidents occur suddenly, so understanding the causes is required to avoid their occurrence in the future or to reduce them. Drivers today face many problems when driving in crowded and congested cities. The driver needs to be cautious and attentive

*Correspondence to: Hayder Majeed Hammadi (Email: haydar@uomisan.edu.iq). Department of Mathematics, College of Basic Education, University of Misan, Misan, Iraq .

ISSN 2310-5070 (online) ISSN 2311-004X (print)

to traffic signals, car horns, music from passing vehicles, in addition to bicycles. Among the main factors that may lead to this problem are lack of experience, immaturity, inaccurate risk perception, overestimation of driving skills, and recklessness. It is essential to identify the underlying perceptions behind the decision to engage in this behavior and to change it. Perceptions in this context may refer to opinions about road accidents, attitudes toward safety, as well as evaluations of road usage behaviors. These perceptions may reflect people's thoughts on how and why things happen.

Traffic accident data for the year 2024 for all provinces of Iraq, except for the Kurdistan Region, was obtained by the Ministry of Planning / Central Statistical Organization / Directorate of Transport and Communications Statistics. The analysis results showed that the provinces of Anbar and Al-Muthanna are stagnant in the early stages, reflecting a clear convergence in the number of accidents, especially rollover accidents resulting from the nature of the long roads and weak traffic monitoring. The results also showed that most hierarchical analysis methods reached groups with similar outcomes, with minor differences due to variations in the Euclidean distance equations for each method. The variations between the provinces were interpreted based on road characteristics and accident patterns.

The main contributions of this study are:

- Providing a clear statistical classification of Iraqi provinces based on the number and type of accidents.
- Offering a quantitative interpretation of tree diagrams and the characteristics of each cluster.
- Highlighting the provinces most prone to accidents, such as Anbar and Al-Muthanna, and linking the results to specific traffic recommendations.

1.1. Study Problem

One of the main reasons threatening human life is traffic accidents, which cause human and material losses. It can be said that the main problem is that some individuals own more than one vehicle, which negatively affects vehicle density and causes environmental pollution both inside and outside cities. The negative effects of traffic accidents reflect on families and communities in general, causing human losses (deaths or injuries) and material losses. These accidents impose heavy burdens on the country. Therefore, there is a need to study this phenomenon and search for effective solutions to mitigate it.

1.2. Study Objectives

Classifying the provinces according to the indicators included in the study to determine the degree of similarity or dissimilarity between the provinces using cluster analysis. And building a statistical model, interpreting it, and presenting the results and recommendations to help reduce traffic accidents in the provinces.

2. Related Works

In 2006, researchers (Al-Shourbagy et al.) [1] conducted a study to develop and improve the level of traffic safety on the roads within King Saud University to identify the characteristics of traffic accidents and determine their severity on the level of traffic safety. The study concluded several findings, including the lack of information in the examined accident reports regarding hazardous locations where traffic accidents frequently occur. Additionally, the difficulty in searching for accident records poses a challenge in the analysis process to identify the causes of the problem and subsequently find solutions. Furthermore, there is insufficient adherence to traffic laws and regulations by road users, and speeding has increased the rate of accidents.

In 2010, (Aworemi)and others [2]. conducted a study on the factors causing traffic accidents in some states in southwestern Nigeria. Data were collected from 352 participants from four out of six states in southwestern Nigeria, with the aim of determining the relationship between human characteristics, vehicle characteristics, road characteristics, environmental characteristics, and traffic accidents in the study areas. It was concluded that human factors, vehicles, roads, and the environment significantly contributed to traffic accidents in the study area,

accounting for approximately 79.4%. Some human characteristics that contribute to their effectiveness in causing accidents include hesitation, fatigue, lack of experience, physical defects, distraction, speeding, and confusion.

In 2016, researchers (Rosenbloom) and others [3]. published a study on adolescents' pre-driving attitudes toward traffic accidents. Where 326 boys and girls from the ninth grade (aged between 14 and 15 years) were asked to answer open-ended questions related to the causes of traffic accidents. One of the most important points (causes) is the three most common points for traffic accidents. The first point relates to the driver's personal reasons, such as anger. The second is the driver and the law, and the third is driving under the influence of alcohol. The most common response was the necessity of providing more information thru awareness campaigns, advertisements, lectures, courses, and talks in schools and workplaces. Most participants considered high school to be a place for disseminating information and raising awareness. The girls provided significantly more answers than the boys and were more optimistic about the success of interventions to reduce traffic accidents.

In 2022, the researcher (Alka beer) [4]. conducted a study on the analysis of road accidents in the city of Misrata during the period from 2012 to 2021 thru comprehensive statistical processing using official data from the traffic department. The researcher used descriptive analysis and the Mann-Kendall test to analyze the temporal trend. Mathematical models were also constructed using traditional linear regression and artificial neural networks (ANN) to estimate accident and fatality rates based on the variable "vehicle ownership level." The results showed that the total number of accidents during the study period exceeded 3,300, with more than 2,100 fatalities recorded, the majority of whom were males at a rate of 87.5%. The age group (20–44) was the most susceptible to death in accidents.

In 2023, the researcher (Essa) and others [5]. conducted a study comparing hierarchical cluster analysis methods in classifying graduate students at Iraqi universities for the academic year (2019–2020), excluding universities in the Kurdistan region. The sample included (30) Iraqi universities, and three variables were adopted: the number of higher diploma students (X1), master's students (X2), and doctoral students (X3). In 2023, the researcher (Essa) and others conducted a study comparing hierarchical clustering methods in classifying graduate students at Iraqi universities for the academic year (2019–2020), excluding universities in the Kurdistan region. The sample included (30) Iraqi universities, and three variables were adopted: the number of higher diploma students (X1), master's students (X2), and doctoral students (X3). The researchers used the Euclidean Distance as a measure of dissimilarity after standardizing the data, and four methods of hierarchical agglomerative clustering were applied: Single Linkage, Complete Linkage, Median Method, and Ward's Method. For the purpose of statistically comparing these methods, the results showed that the Complete Linkage method is the best among the methods used, followed by the Median method, then the Ward method, while the Single Linkage method came in last place. The results also showed that the best classification was with three clusters, where the University of Baghdad was in the first cluster, the Technical Institutes and Universities Authority in the second cluster, while the third cluster included the rest of the Iraqi universities. This indicates the efficiency of the complete linkage method in representing the true structure of the data and achieving the highest homogeneity between the clusters.

In 2025, the researchers (Hanadi A. Amhimmid) and others [6]. published a comparative study of hierarchical clustering methods: Single Linkage, Average Linkage, and Ward's Method. A comparative study by Amhimmid and others on methods in hierarchical cluster analysis includes: Single Linkage, Average Linkage, and Ward's Method. The study aimed to evaluate the performance of these methods in clustering data with different characteristics in terms of shape, density, and noise level. The analysis results were represented using a dendrogram to illustrate the stages of merging between the clusters. The study results clearly demonstrated that the Ward method excelled in achieving the highest degree of homogeneity within the clusters and the best separation between them, while the single method showed a good ability to detect extended clusters. The centroid method provided a balanced performance between the two methods, making it a suitable option in cases where the nature of the data is unclear. The study concluded that the choice of hierarchical cluster analysis method largely depends on the nature and structure of the data, and confirmed that the Ward method is one of the most efficient methods in applications that require high homogeneity within the clusters and clear differentiation between them.

In 2025, Rahman and others [7]. conducted a study on regional disparities in maternal and child health in Bangladesh, relying on data from the Multiple Indicator Cluster Survey (MICS 2019). The research problem was the clear disparity between regions in terms of maternal and child health levels and access to basic health

services. The researchers relied on a set of health and social variables in the study, which included indicators of child mortality, prenatal and postnatal care, nutrition, vaccination, and basic health services. Where hierarchical cluster analysis was used to classify regions into homogeneous clusters, relying on Euclidean distance and various aggregation methods, most notably Ward's Method. The dendrogram was also used to determine the optimal number of clusters and to spatially interpret the structure of the data. The analysis results showed the possibility of classifying the regions of Bangladesh into several clusters that differ significantly from each other in terms of maternal and child health indicators. The results revealed the presence of clusters with high levels of health deprivation compared to other clusters with better health conditions. The Ward method also proved its efficiency in achieving the highest degree of homogeneity within the clusters and clearer separation between them compared to other methods. The study concluded that cluster analysis is an effective statistical tool in revealing patterns of regional disparity and contributes to supporting decision-makers in directing health policies fairly and toward the most needy areas, based on precise quantitative foundations.

3. Theoretical Aspect

3.1. Cluster Analysis

Cluster analysis aims to classify a sample of observations into two or more different but unknown categories based on the formations of variable categories [5, 8]. Often, the goal of this analysis is to discover a certain pattern that organizes the observations. The purpose of this analysis is to categorize individuals into groups that share common characteristics, allowing someone to easily predict the behaviors or properties of other individuals or items based on the knowledge of the categories to which these items belong, whether they are people or things that share the same characteristics [9, 10], there are set of way in predicting enter to various fields such as medicine [11, 12], geographic [13, 14] ,agricultural [15, 16, 17], and more. Cluster analysis is a useful technique for understanding thru a dataset where the dataset is grouped into a cluster such that the data within the same cluster are similar to each other and dissimilar to the data in other clusters [18].

The use of cluster analysis has proven to be highly successful in many sciences. Its extensive use in many scientific and practical applications, such as population surveys and network linking, has been proven. This method has been employed in numerous cases to divide society into clusters with the aim of prioritizing [19]. The resulting classifications have proven to be highly beneficial, whether in distributing health or commercial brochures, or in identifying healthcare centers or commercial centers and categorizing them into the most effective segments of society.

The main objectives of cluster analysis can be identified as follows:

- Data Exploration Data Exploration: thru cluster analysis, the basic structure of the data and its homogeneous branches can be understood, and this knowledge can then be used to uncover the data.
- Data Reduction Data Reduction: in the case where the researcher faces large amounts of heterogeneous observations, it may be difficult to handle them unless they are grouped into homogeneous sets and each group is represented by a number, after which they can be dealt with [8].
- Diagnosis Diagnosis (Identification): the most important and fundamental step in cluster analysis is identifying the similar characteristics within a single cluster and the different characteristics between clusters by dividing the community or observations into clusters [6].
- Generating hypotheses Generating hypotheses: it benefits from cluster analysis to understand the structure of society and provides hypotheses for that and tests them [20].
- Forecasting Forecasting: forecasting after displaying the clustering results and using them in the future.

3.2. Cluster Analysis Methods

There are two main types of cluster analysis as show in next sections

3.2.1. Hard Cluster Analysis: It involves the membership of an element to only one cluster, and it is the most common method described as one of the suitable techniques for classifying data elements [6]. It is preferred to use it when there is no information available about any of the variables, with several measurements available. It is divided into hierarchical cluster analysis and non-hierarchical cluster analysis [8], [9].

3.2.2. Fuzzy cluster analysis: It is the membership of an element to all clusters but with varying degrees of membership [8].

3.3. Hard Cluster Analysis Methods

The method of cluster analysis differs from other statistical techniques, as the sample size in cluster analysis is not related to statistical inference. This is because the goal of cluster analysis is not to estimate the extent to which the obtained results can be generalized to the population, but rather that the results of cluster analysis are based solely on the sample, i.e., the study variables. There are many clustering methods that have been developed to describe the shape of relationships between elements. These methods are primarily based on the distance matrix or the correlation matrix [5, 12]. There are two main types of acute cluster analysis are hierarchical analysis and non-hierarchical analysis.

3.4. Clustering Steps

- Calculate the distance matrix or the correlation matrix.
- Grouping elements into a dendrogram (linking the two elements that have the shortest distance between them)[22].
- The linking process continues until reaching the dendrogram.

The length of the line in the dendrogram indicates an increase in dissimilarity levels, and thus, an increase in the line length between two variables indicates dissimilarity between their data [9].

3.5. Hierarchical Cluster Analysis

This method is uncomplicated as it does not require prior knowledge of the number of clusters on which the cases are classified [23]. It is also characterized by its ability to be used with small samples and works on classifying the sample items by following one of the two methods as follows:

- Clustering method: In this method, the partial groups of clusters or observations are grouped together to obtain a more comprehensive set. That is, the analysis begins with one cluster for each case, and then the close clusters are combined until we obtain a number of clusters that include a group of elements [24].
- The divisive method: It starts with a single cluster that includes all the subsets and observations, and this cluster is divided into smaller and smaller clusters. After grouping the elements into clusters, the similarity between the clusters must be calculated, and there are many different methods used for the purpose of calculating similarity [8].

Hierarchical agglomerative clustering uses several methods called linkage methods.

3.5.1. Single Linkage Method: It is also called the nearest neighbor method. This method is primarily based on the assumption that the two most similar elements form the nucleus of the cluster by finding the shortest distance between any two elements of the clusters [25]. The remaining units are then added to this nucleus sequentially based on their similarity to the elements of the cluster nucleus, with the most similar being added first, followed by the less similar in order. According to the following formula:

$$D(R,Q) = \min\{D(A,Q),\ D(B,Q)\} \tag{1}$$

Since: Represents one of the measures of Euclidean distance $D(R,Q)$

3.5.2. Complete Linkage Method This method is also called the complete linkage method or the farthest neighbor [24, 26].

3.5.3. Centroid linkage method: The centroid linkage method is considered the Euclidean distance between the centroids of the two clusters [25].

$$D(A,B) = d(\bar{y}_A, \bar{y}_B) \tag{2}$$

3.5.4. Median Method: This method relies on determining the similarity between any two clusters as the average distance between the closest clusters [23].

3.5.5. Ward's Method: It is considered the optimal and best method because it uses the analysis of variance approach to determine the distances between clusters, meaning it attempts to minimize the sum of squares for any cluster of the two merged clusters at each step, as proposed by Ward in 1963 [24]. This missing information can be calculated by determining the total sum of the squares of the deviations of each element from the mean of the cluster to which it belongs, according to the following formula [27]:

$$\mathrm{SSE}_A = \sum_{i=1}^{n} (y_i - \bar{y}_A)'(y_i - \bar{y}_A) \tag{3}$$

$$\mathrm{SSE}_B = \sum_{i=1}^{n} (y_i - \bar{y}_B)'(y_i - \bar{y}_B) \tag{4}$$

$$\mathrm{SSE}_{AB} = \sum_{i=1}^{n} (y_i - \bar{y}_{AB})'(y_i - \bar{y}_{AB}) \tag{5}$$

Since the two clusters are linked, it reduces the increase in the square of the distances and expresses the amount of that increase as follows:

$$I_{AB} = \mathrm{SSE}_{AB} - (\mathrm{SSE}_A + \mathrm{SSE}_B) \tag{6}$$

The three methods were chosen as indicated for each of them:
1. Single Linkage: It relies on the closest distance between any two points in different clusters.
2. Complete Linkage: It relies on the farthest distance between any two points in different clusters.
3. Ward's Method: It attempts to minimize the variance within each cluster when merging the groups.

Because each linkage method yields slightly different results due to the varying mechanisms for calculating distances between clusters.

3.5.6. Data Matrix (1): The set of elements can be represented by a matrix, where there are (n) rows representing the observations of the elements or samples and (p) columns with a dimension of (n*p), which represent the properties or distinguishing features of those elements. The matrix X is called the model matrix [10]. Hierarchical clustering uses the original data matrix, but many hierarchical clustering methods use the similarity matrix or the dissimilarity matrix, which are both referred to as the proximity matrix. The proximity matrix is a symmetric matrix of order (m*n) that contains all the pairwise differences or similarities between any two elements [9, 10].

3.5.7. Distance Measurement: The distance measure is one of the most commonly used metrics in most clustering methods, as the fundamental idea of clustering elements is based on the concept of distance. Clusters should contain elements separated by relatively small distances compared to the distances between the clusters. These distances depend on single or multiple dimensions [9]. Data is often converted into a distance matrix (D) or a similarity matrix (S) for clustering n elements by determining the distance between the centers of the clusters and the groups of observations. Among the most important distance measures used are. It is the Euclidean measureis the most

common and widely used metric for calculating distances between elements in a multidimensional space, and the distance is calculated using the following formula [21]:

$$d_E(x_i, x_j) = \|x_i - x_j\| = \sqrt{\sum_{d=1}^{p}(x_{id} - x_{jd})^2} \tag{7}$$

Since x_i and x_j represent the i^{th} and j^{th} elements in a p-dimensional space.

4. The practical side

The Central Statistical Organization prepares an annual report on traffic accidents based on police station records in coordination with the Ministry of Interior. Data is collected from reports sent by the Police Affairs Agency / Criminal Statistics Directorate, which gathers accident records according to the numbers of these stations and their geographical units across all Iraqi governorates [28]. The World Health Organization seeks to provide indicators on the total number of fatalities, and this report includes only the fatal accidents that occurred during the incident. The number of recorded traffic accidents reached 11,763 in 2024 compared to 11,552 in 2023, an increase of 0.018%.

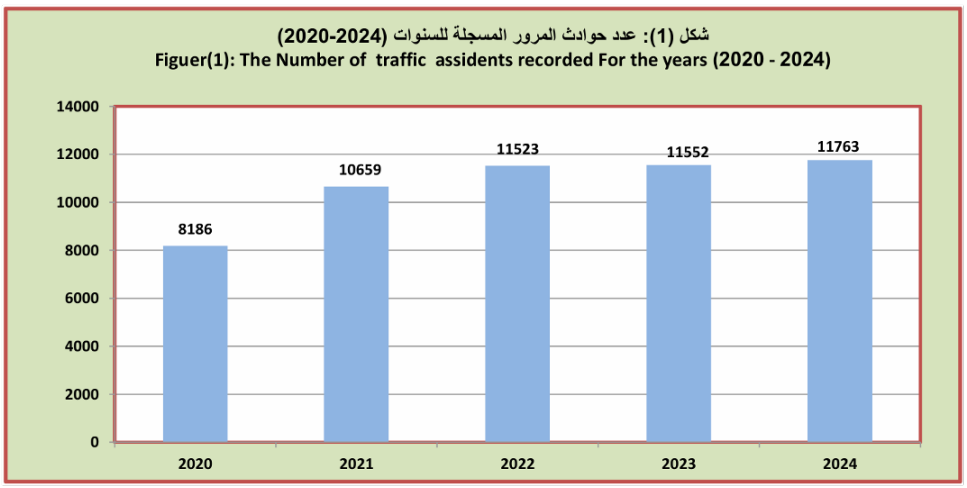

Figure 1. Number of recorded incidents for the years 2020-2024

4.1. Study Variables and Coding

This study relied on accident type indicators as aggregation variables to ensure consistency and completeness of reporting across the provinces.

$$Y_1 = \text{Collision}, \qquad Y_2 = \text{Overturning}, \qquad Y_3 = \text{Trampling}$$

The provinces have been coded as follows:

4.2. Distance Measurement and Standardization

The Euclidean distance metric was adopted in all clustering methods. Before calculating the distances, the data was standardized using the Z-score standardization. (Z-score Standardization) This step was necessary to prevent variables with large values (such as Baghdad due to the high number of accidents there) from significantly affecting distance calculations. Standardization ensures that each indicator contributes equally to the clustering structure.

Table 1. Iraqi Provinces Coding (0–15)

Code	Province
1	Nineveh
2	Kirkuk
3	Diyala
4	Anbar
5	Baghdad
6	Babylon
7	Karbala
8	Wasit
9	Salah al-Din
10	Najaf
11	Qadisiya
12	Al-Muthanna
13	Dhi Qar
14	Maysan
15	Basra

4.3. Hierarchical Aggregation Results

4.3.1. Single linkage method: (Nearest neighbor Method): Steps for clustering according to the single linkage method based on the nature of the incident and at the governorate level for the year 2024. Provincial level for the year 2024

Table 2. Steps of aggregation according to the single linkage method

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	4	12	0.309	0	0	5
2	9	14	0.335	0	0	3
3	2	9	0.344	0	2	4
4	2	13	0.383	3	0	5
5	2	4	0.493	4	1	7
6	3	6	0.619	0	0	11
7	1	2	0.648	0	5	8
8	1	7	0.785	7	0	9
9	1	10	0.819	8	0	10
10	1	11	0.886	9	0	11
11	1	3	1.307	10	6	12
12	1	8	1.845	11	0	13
13	1	15	1.974	12	0	14
14	1	5	3.303	13	0	0

The table includes the aggregation steps as it illustrates the stages of clustering the provinces into clusters based on the distances between them, which were determined in the proximity matrix. The first column shows the step number (stage), the second shows the combined clusters of provinces (cluster combined), and the coefficients indicate the distance between the linked provinces. The fourth column shows whether or not the province is present in a previous cluster (stage cluster appears), and the last column represents the subsequent step in which the new province for each cluster will appear. The results of the cluster analysis, which determines the similarity or dissimilarity between the units (provinces), were obtained, with similarity expressed thru the derived distance

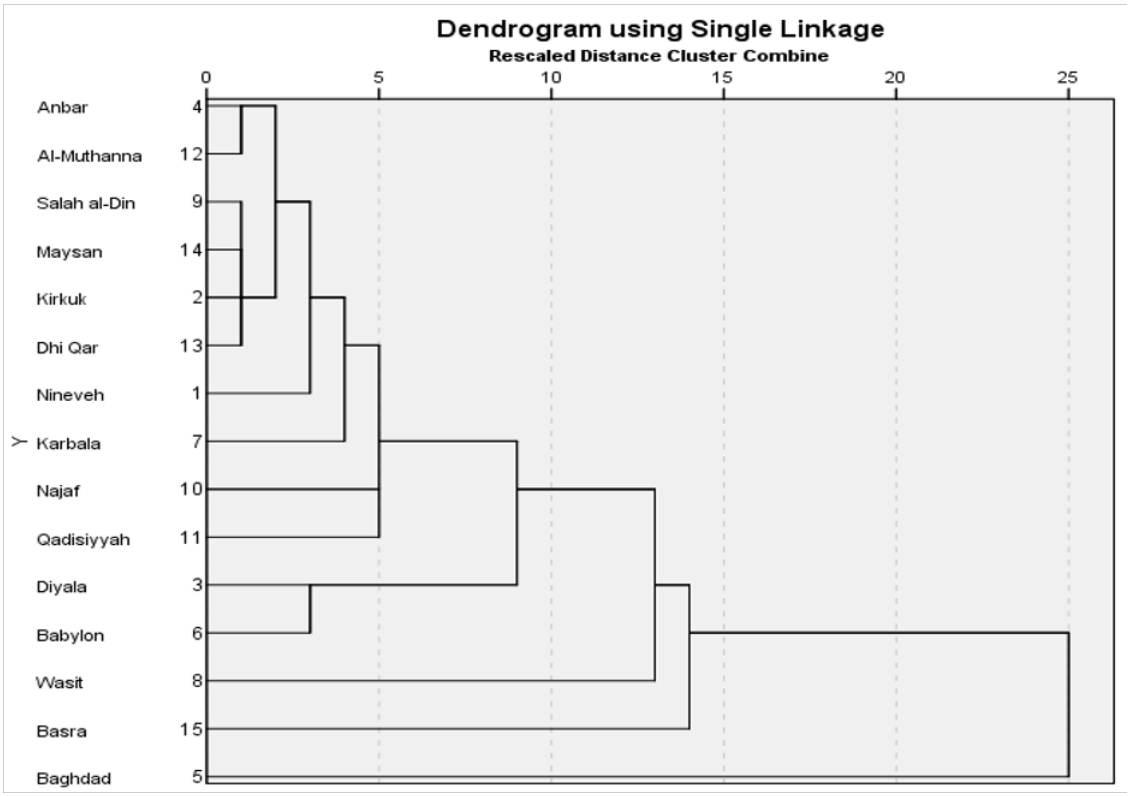

Figure 2. Tree diagram according to the single linkage method for the nature of the incident At the provincial level for the year 2024

between the specified objectives. The smallest difference in coefficients in the first step was estimated between Al-Anbar Governorate (unit 4) and Al-Muthanna Governorate (unit 12) with a Euclidean distance of (0.309), which is the shortest compared to others. The fourth column indicates that neither of the two governorates was present in any previous step. As for the fifth column, it shows that in the subsequent fifth step, a new governorate was linked to one of these two governorates, which represents the linking of Anbar Governorate (unit 4) with Kirkuk Governorate (unit 2).

Moving to the second row, we find that the proximity and linkage have become between Salah al-Din Governorate (item 9) and Maysan Governorate (item 14) with a distance of (0.335). It is clear from the fourth column that neither of the two governorates was present in any previous step. The fifth column indicates that in the next step (the third), there will be a linkage and proximity between Salah al-Din Governorate (item 9) and Kirkuk Governorate (item 2).

The clustering process continued from the least distant to the most distant in an ascending manner, where the largest proximity coefficient was (3.303) between Nineveh Governorate (item 1) and Baghdad Governorate (item 5).

The table shows the degree of homogeneity in the included groups; a small coefficient value indicates that the group is homogeneous, while a large coefficient value indicates that the homogeneity between the groups is lower.

Figure (2) shows the provinces that were linked together at each step of the analysis, with distances divided at the top of this figure to measure how close each province is to the others or how close the groups are to each other.

The length of the line indicates an increase in the degrees of similarity, and there are several nodes in the tree where each node represents the merging of nearby provinces. This is evident from the following tree structure:

Baghdad Governorate is the farthest from the other governorates, as it merged with the rest of the tree at the highest value of merging distances. This indicates that Baghdad Governorate is the most distinct according to the indicators used in the analysis.

Next is Al-Anbar Governorate, which represents the second cluster in terms of distance from the other governorates, as it appears isolated at the top of the tree and far from any small group, indicating its clear separation from the other governorates.

Followed by a group of provinces that merge at levels higher than the average in terms of distance, the closest of which in order are Basra Province and Dhi Qar Province, as they merged with the rest of the tree at medium distances compared to the other provinces, indicating a relative distinction for these two provinces.

Another close-knit group appears, including: Maysan, Salah al-Din, and Sulaymaniyah, where these provinces merge at a relatively early stage and within small distances, reflecting their clear similarity.

Thus, the provinces can be classified into three main categories in terms of similarity according to the single linkage method:

1. Highly distinct provinces: Baghdad, Anbar. Highly distinct provinces: Baghdad, Anbar.
2. Provinces with moderate distance: Basra, Dhi Qar. Provinces with moderate distance: Basra, Dhi Qar.
3. Very close provinces: Nineveh, Kirkuk, Diyala, Babylon, Wasit, Najaf, Al-Muthanna, Karbala, Maysan, Salah al-Din, Sulaymaniyah. Very close provinces: Nineveh, Kirkuk, Diyala, Babylon, Wasit, Najaf, Al-Muthanna, Karbala, Maysan, Salah ad-Din, Sulaymaniyah.

Based on the appropriate pruning of the tree, four main clusters can be adopted as follows:

- The first cluster: Baghdad alone.
- The second cluster: Anbar.
- The third cluster: Basra and Dhi Qar.
- The fourth cluster: the remaining closely situated governorates, which are (Nineveh, Kirkuk, Diyala, Babylon, Wasit, Najaf, Al-Muthanna, Karbala, Maysan, Salah ad-Din, Sulaymaniyah).

The number of clusters is determined based on the principle of the largest increase in rescaled distance, where a clear increase in distance is observed when moving from four clusters to three. This makes adopting four clusters the statistically most suitable choice according to the single linkage method, achieving the highest degree of separation between groups while maintaining homogeneity within each cluster.

4.3.2. Centroid linkage method The steps of clustering according to the centroid linkage method based on the nature of the incident and at the provincial level for the year 2024.

Table (2) includes the clustering steps as it illustrates the stages of clustering the provinces into clusters based on the distances between them, which were found in the proximity matrix. The smallest difference in transactions in the first step was estimated between Al-Anbar Governorate (unit 4) and Al-Muthanna Governorate (unit 12) with an Euclidean distance of (0.309), which is the shortest compared to others. The fourth column shows that neither of the two governorates was present in any previous step. As for the fifth column, it shows that in the subsequent fifth step, a new governorate was linked to one of these two governorates, which represents the linking of Al-Anbar governorate (unit 4) with Kirkuk governorate (unit 2).

Moving on to the second row, we find that the proximity and linkage have become between Salah al-Din Governorate (item 9) and Maysan Governorate (item 14) with a distance of (0.335). It is clear from the fourth column that neither of the two governorates was present in any previous step. The fifth column indicates that in the next step (the third), there will be a linkage and proximity between Salah al-Din Governorate (item 9) and Dhi Qar Governorate (item 13). The clustering process continued from the least distant to the most distant in an ascending manner, with the largest proximity coefficient being (3.329) between Nineveh Governorate (item 1) and Baghdad Governorate (item 5). This table shows the degree of homogeneity in the included groups, as a small coefficient value indicates that the group is homogeneous, while a large coefficient value indicates that the homogeneity between the groups is lower.

Table 3. shows the aggregation method according to the central linking method.

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1.	4	12	.309	0	0	5
2.	9	14	.335	0	0	3
3.	9	13	.315	2	0	4
4.	2	9	.289	0	3	5
5.	2	4	.436	4	1	10
6.	3	6	.619	0	0	11
7.	1	7	.785	0	0	8
8.	1	10	.782	7	0	9
9.	1	11	.771	8	0	10
10.	1	2	.801	9	5	11
11.	1	3	1.263	10	6	12
12.	1	15	2.093	11	0	13
13.	1	8	2.736	12	0	14
14.	1	5	3.329	13	0	0

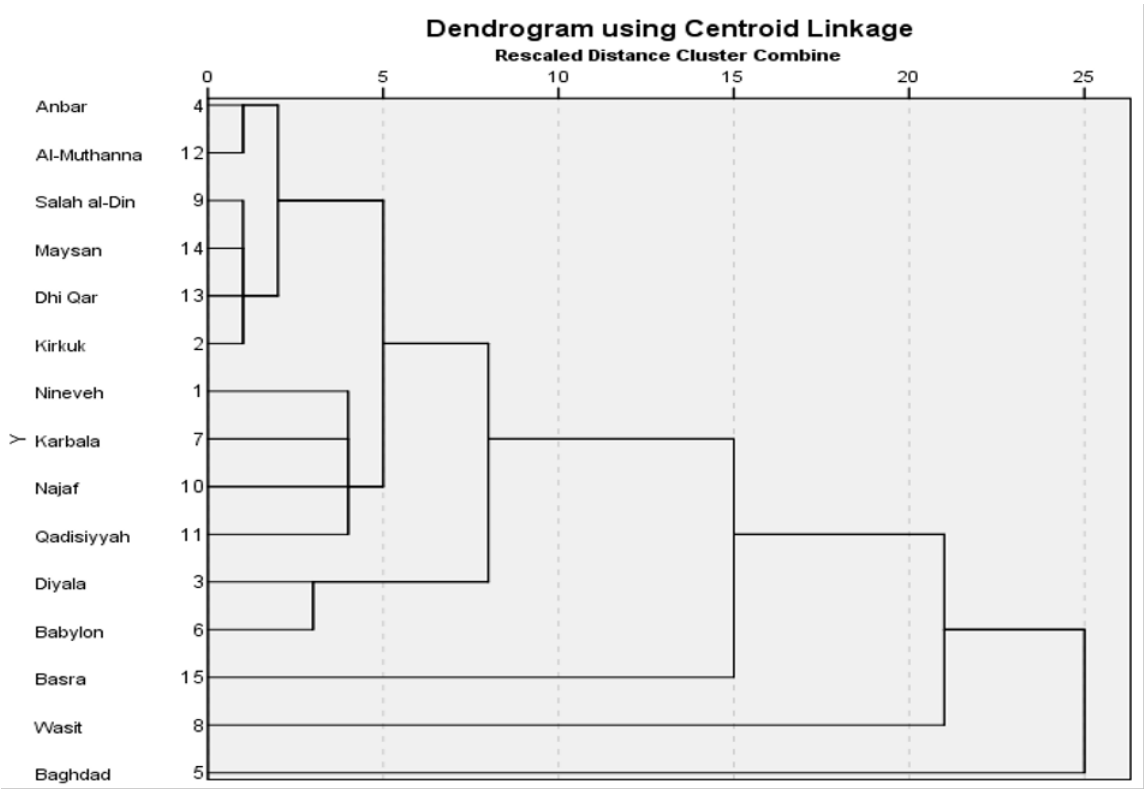

Figure 3. Tree diagram according to the central linking method for the nature of the incident

Figure (3) illustrates the dendrogram resulting from hierarchical cluster analysis using the central linkage method, showing that the provinces linked together at each stage of the analysis are based on the distances between the cluster centers. The length of the horizontal line indicates the degree of dissimilarity; the longer the line, the greater the distance, while a shorter line indicates a higher degree of similarity. It is clear from the following

dendrogram: Baghdad Governorate is considered the most distinct among the other governorates, as it appears at the bottom of the chart alone and merges with the other governorates at the highest level of distance. Following it is Wasit Governorate, which also merges at a relatively large distance compared to the other governorates. Then comes Basra Governorate, whose merging occurs at a medium to high distance, reflecting a unique pattern that distinguishes it from the other governorates, albeit to a lesser degree than Baghdad and Wasit. As for the intermediate stages of the analysis, Dhi Qar Governorate appears, merging at a medium distance, reflecting a moderate degree of variation compared to the other governorates. In contrast, a group of provinces emerges that have merged at relatively small distances, indicating a strong similarity among them. Where a homogeneous group consists of the provinces of Nineveh and Kirkuk, which are considered among the closest provinces in this analysis.

Another relatively close group is formed by Diyala and Babel, where these two provinces merged early in the analysis, indicating a significant similarity in their characteristics. Another relatively close group appears, including the provinces of Anbar, Al-Muthanna, Maysan, Salah al-Din, and Sulaymaniyah, where these provinces merged at relatively close levels of distance, reflecting a general proximity among them compared to the more distinct provinces.

Thus, the provinces can be classified into three main categories in terms of similarity based on the central linkage method:

1. Highly distinct provinces: Baghdad, Wasit. Highly distinct governorates: Baghdad, Wasit.
2. Provinces with moderate distance: Basra, Dhi Qar. Provinces with moderate distance: Basra, Dhi Qar.
3. Relatively close provinces: Nineveh, Kirkuk, Diyala, Babylon, Najaf, Karbala, Anbar, Muthanna, Maysan, Salah al-Din, Sulaymaniyah. Relatively close provinces: Nineveh, Kirkuk, Diyala, Babylon, Najaf, Karbala, Anbar, Al-Muthanna, Maysan, Salah al-Din, Sulaymaniyah.

Based on the appropriate branching of the tree, three main clusters can be adopted as follows:

- The first cluster: Baghdad and Wasit.
- The second cluster: Basra and Dhi Qar.
- The third cluster: the remaining neighboring provinces, which are (Nineveh, Kirkuk, Diyala, Babylon, Najaf, Karbala, Anbar, Al-Muthanna, Maysan, Salah al-Din, Sulaymaniyah).

4.3.3. Ward's Method: The steps for clustering according to Ward's hierarchical method based on the nature of the incident and at the provincial level for the year 2024.

Table 4. shows the clustering method according to the Ward hierarchical method.

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	4	12	0.155	0	0	9
2	9	14	0.322	0	0	3
3	9	13	0.533	2	0	4
4	2	9	0.750	0	3	9
5	3	6	1.059	0	0	10
6	1	7	1.452	0	0	7
7	1	10	1.973	6	0	8
8	1	11	2.551	7	0	11
9	2	4	3.133	4	1	13
10	3	8	4.341	5	0	12
11	1	15	5.969	8	0	13
12	3	5	8.165	10	0	14
13	1	2	10.518	11	9	14
14	1	3	14.739	13	12	0

Thru Table (3) and Figure (4) of the Ward linkage method, Al-Anbar Governorate (unit 4)

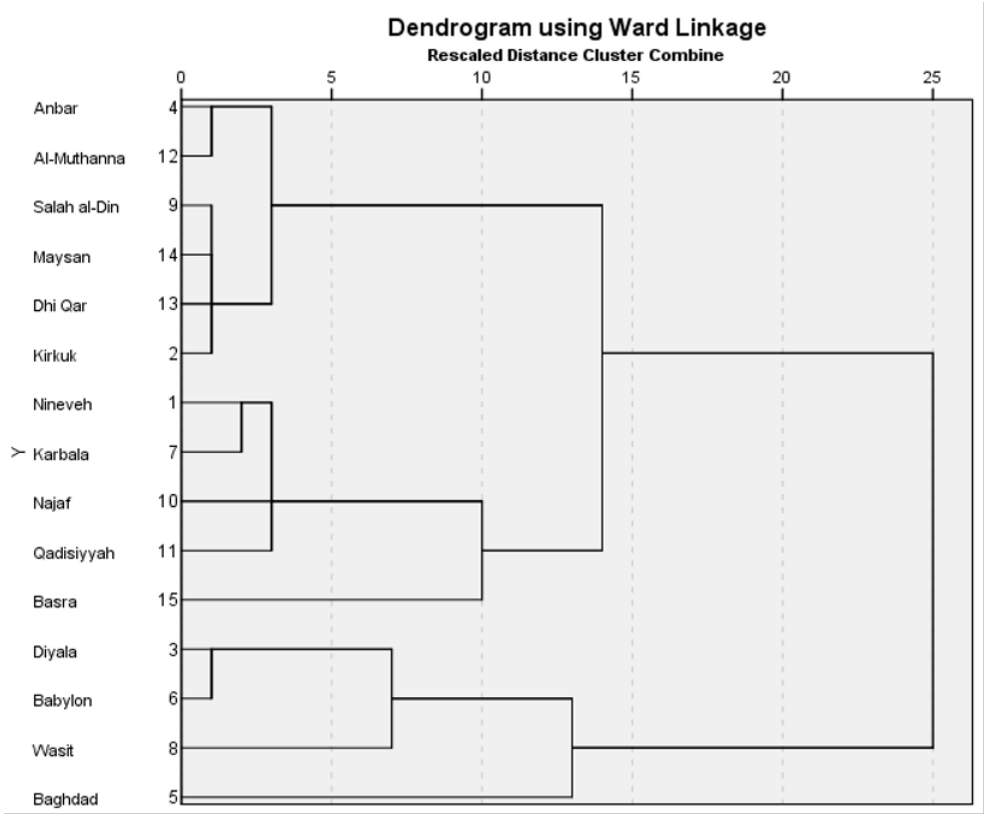

Figure 4. Tree diagram according to the Ward linkage method for the nature of the incident

Table (3) includes the clustering steps as it illustrates the stages of clustering the provinces into clusters based on the distances between them, which were found in the proximity matrix. The smallest difference in transactions in the first step was estimated between Al-Anbar Governorate (unit 4) and Al-Muthanna Governorate (unit 12) with an Euclidean distance of (0.155), which is the shortest among others. The fourth column indicates that neither of the two governorates was present in any previous step. As for the fifth column, it shows that in the subsequent step (the ninth), a new governorate was linked to one of these two governorates, which represents the linking of Anbar Governorate (unit 4) with Kirkuk Governorate (unit 2).

The clustering process continued from the least distant to the most distant in an ascending manner, with the largest proximity coefficient being (14.739) between Nineveh Governorate (item 1) and Dhi Qar Governorate (item 13).

The table illustrates the degree of homogeneity within the included groups; a small coefficient value indicates that the group is homogeneous, while a large coefficient value indicates less homogeneity between the groups.

Figure (4) shows the dendrogram resulting from hierarchical cluster analysis using the Ward Linkage method, which relies on reducing the variance within the clusters at each step of the merging process. The distances listed at the top of the figure within the range (0–25) represent a measure of how close or far the provinces or groups are from each other. The length of the horizontal line indicates the degree of dissimilarity; the longer the line, the greater the distance, while a shorter line indicates a higher degree of similarity.

Where Baghdad Governorate was the farthest from the other governorates, appearing at the bottom of the chart and merging with the rest of the tree at the highest distance value, indicating that it is the most distinct and different according to the variables used in the analysis.

Next is Wasit Province, which also appears at a relatively high level of integration, indicating its clear distinction compared to the other provinces, although its degree of divergence is less than that of Baghdad Province. Next

comes Basra Governorate, which integrates at a medium to high distance, indicating that it possesses a distinct pattern that sets it apart from most governorates, but to a lesser degree than Baghdad and Wasit.

In the intermediate stages of the analysis, Dhi Qar Governorate appears, merging at a moderate distance, reflecting a moderate degree of differentiation compared to the more distinct governorates. In contrast, a group of provinces emerges that have merged at relatively small distances, indicating a strong similarity among them. Where a homogeneous group is formed that includes the provinces of Nineveh and Kirkuk, which merged at an early stage, indicating a significant similarity in their characteristics. Another relatively close group is formed by Diyala and Babel, where these two provinces merged at a relatively small distance, reflecting a clear proximity between them.

Another relatively homogeneous group appears, comprising the provinces of Anbar, Al-Muthanna, Maysan, Salah al-Din, and Sulaymaniyah, where these provinces merged at relatively close distances, reflecting a general homogeneity among them compared to the more distinct provinces.

Thus, the provinces can be classified into three main categories in terms of similarity according to the Ward method:

1. Highly distinct provinces: Baghdad, Wasit. Highly distinct governorates: Baghdad, Wasit.
2. Provinces with moderate distance: Basra, Dhi Qar. Provinces with moderate distance: Basra, Dhi Qar.
3. Relatively close provinces: Nineveh, Kirkuk, Diyala, Babylon, Najaf, Karbala, Anbar, Muthanna, Maysan, Salah al-Din, Sulaymaniyah. Relatively close provinces: Nineveh, Kirkuk, Diyala, Babylon, Najaf, Karbala, Anbar, Al-Muthanna, Maysan, Salah al-Din, Sulaymaniyah.

Based on the appropriate branching of the tree, three main clusters can be adopted as follows:

- The first cluster: Baghdad and Wasit.
- The second cluster: Basra and Dhi Qar.
- The third cluster: the remaining closely situated governorates, which are (Nineveh, Kirkuk, Diyala, Babylon, Najaf, Karbala, Anbar, Muthanna, Missan, Salah al-Din, Sulaymaniyah).

The determination of the number of clusters is based on the principle of the largest increase in rescaled distance, as the transition from three to two clusters requires a relatively large distance, indicating that dividing the data into three clusters is statistically the most appropriate according to Ward's method, as it achieves the highest degree of homogeneity within the clusters and the greatest separation between them.

4.4. Comparison of provincial membership within the clusters using the three linkage methods

This comparative table shows that the majority of provinces (such as Nineveh, Kirkuk, Anbar, Karbala, Najaf, Qadisiyyah, Al-Muthanna, Dhi Qar, Maysan, Salah al-Din) maintained a clear stability in their membership within cluster (1) across the three methods, indicating a strong homogeneity in their statistical characteristics. It also shows that Baghdad Governorate formed an independent cluster (4) in all methods, reflecting its high distinction from the other governorates. In contrast, differences are observed in the distribution of some provinces such as Diyala, Babel, Wasit, and Basra, where these provinces shifted between more than one cluster depending on the variation in the linkage methodology, reflecting their sensitivity to the nature of the algorithm used in clustering.

5. Conclusions and Recommendations

The results of the hierarchical cluster analysis of traffic accidents by the nature of the incident (collision, rollover, run-over) indicate that the provinces of Anbar and Al-Muthanna clustered together in the early stages across most of the statistical methods used, reflecting a clear convergence in the pattern and number of accidents between them. This similarity is attributed to the characteristics of long highways and weak traffic control, which contributes to the increase in rollover accidents. The Schedule Agglomeration tables also showed convergence in the merging steps between the different methods, with slight differences in the clustering sequence. This is attributed to the variation in the equations for calculating the Euclidean distance without affecting the overall structure of the clusters. It

Table 5. Cluster Memberships Using Different Linkage Methods

Governorate	Single Linkage	Complete Linkage	Ward
Nineveh	1	1	1
Kirkuk	1	1	1
Diyala	1	3	2
Anbar	1	1	1
Baghdad	4	4	4
Babylon	1	3	2
Karbala	1	1	1
Wasit	2	3	2
Salah al-Din	1	1	1
Najaf	1	1	1
Qadisiya	1	1	1
Al-Muthanna	1	1	1
Dhi Qar	1	1	1
Maysan	1	1	1
Basra	3	2	3

was observed that Anbar Governorate is one of the most affected by incidents, as confirmed by distance measures that reflect the degree of homogeneity and the beginnings of clustering between the governorates. Thru the results, it was found that most hierarchical analysis methods yielded similar results with limited differences due to the varying methodologies adopted in each method. The results also showed that Baghdad Governorate formed an independent cluster, indicating a distinct pattern of incidents compared to other governorates. Meanwhile, Basra Governorate formed its own cluster, suggesting a relatively independent pattern of incidents. As for the provinces of Diyala, Babel, and Wasit, they shifted between different clusters depending on the linkage method, but they mostly settled within a medium-risk cluster. Meanwhile, the other homogeneous provinces settled within a single cluster across all methods, indicating a uniform pattern of incidents.

Based on these results, the study recommends implementing targeted field interventions in the provinces of Anbar and Al-Muthanna, particularly the installation of rumble strips on long, straight desert roads to reduce loss-of-control accidents. The study also recommends, regarding Baghdad Governorate, the adoption of intelligent traffic management systems to control high-accident intersections, and the allocation of traffic awareness programs for vehicle drivers within major cities. Regarding Basra Governorate, it is recommended to tighten control over the roads leading to the ports and regulate the movement of heavy trucks at specific times, along with conducting regular inspections of heavy vehicles due to their high contribution to accidents.

It is suggested to adopt unified traffic policies and implement standard traffic safety programs, along with strengthening ambulance and emergency teams on external roads. The study emphasizes the necessity of tightening the enforcement of mandatory traffic safety measures, particularly the adherence to wearing seat belts, respecting speed limits, and using child protection systems. The study also recommends expanding the scope of analysis in the future by comparing other statistical classification methods, such as using discriminant analysis, and integrating non-hierarchical methods, particularly the K-means algorithm alongside hierarchical methods. Additionally, it suggests studying data spanning multiple years and conducting regional comparisons with countries of similar contexts by building unified databases, which would enhance the accuracy of statistical interpretation and support decision-making in reducing traffic accidents.

REFERENCES

1. M. F. Al-Shorbaji and A. B. S. Al-Ghamdi, *An exploratory study on improving the level of traffic safety on the roads within King Saud University*, King Saud University, 2006.

2. J. R. Aworemi, I. A. Abdul-Azeez, and S. O. Olabode, *Analytical study of causal factors of road traffic accidents in southwest Nigeria*, Educational Research, vol. 1, no. 4, pp. 118–124, 2010.
3. T. Rosenbloom, A. Ben-Eliyahu, and D. Nemrodov, *Causes of road accidents according to the perception of adolescents who do not yet drive*, North American Journal of Psychology, vol. 18, no. 3, pp. 487–500, 2016.
4. A. M. Elder, *Statistical analysis of road accidents: A case study on the city of Misrata*, International Journal of Engineering and Information Technology (IJEIT), vol. 10, no. 1, pp. 100–106, 2022.
5. A. K. Essa, L. S. H. Fadhil, and D. H. Shihab, *A comparison between the hierarchical clustering methods...*, Periodicals of Engineering and Natural Sciences (PEN), vol. 11, no. 1, pp. 174–185, 2023.
6. H. A. Amhimmid et al., *Comparative study of four methods in hierarchical cluster analysis*, Libyan Journal of Medical and Applied Sciences, vol. 3, no. 4, pp. 11–16, 2025.
7. M. Rahman et al., *Examining regional disparities in maternal and child health in Bangladesh using cluster analysis*, Scientific Reports, vol. 15, no. 1, p. 32741, 2025.
8. K. Genets, G. Wants, and T. Vanhoof, *Traffic road clustering and profiling using accident data*, University of Limburg, Belgium, 2003.
9. M. R. Anderberg, *Cluster Analysis for Applications*, Academic Press, 2014.
10. A. Jaeger and D. Banks, *Cluster analysis: A modern statistical review*, Wiley Interdisciplinary Reviews: Computational Statistics, vol. 15, no. 3, e1597, 2023.
11. A. Abed Mohammed, P. Sumari, and K. Attabi, *Hybrid K-means and Principal Component Analysis (PCA) for Diabetes Prediction*, International Journal of Computing and Digital Systems, vol. 15, no. 1, pp. 1719–1728, Jun. 2024, doi: 10.12785/ijcds/1501121.
12. L. S. Ashour et al., *Non-overlapping Patch-Based Pre-trained CNN for Breast Cancer Classification*, Iraqi Journal for Computer Science and Mathematics, vol. 6, no. 2, Jun. 2025, doi: 10.52866/2788-7421.1271.
13. M. M. Abd Zaid, A. A. Mohammed, and P. Sumari, *Classification of Geographical Land Structure Using CNN and Transfer Learning*, Journal of Computer Science, vol. 20, no. 12, pp. 1580–1592, 2024.
14. M. M. Abd Zaid, A. Abed Mohammed, and P. Sumari, *Classification of Road Features Using CNN and Transfer Learning*, International Journal of Computing and Digital Systems, vol. 17, no. 1, pp. 1–12, 2025.
15. P. Sumari et al., *Optimizing Lemongrass Disease Detection...*, Statistics, Optimization and Information Computing, vol. 14, no. 4, pp. 2022–2040, 2025.
16. A. A. Mohammed et al., *Durian Fruit Diseases Detection Using CNN and Pre-trained Models*, Springer, 2025, pp. 97–116.
17. B. H. Jawad et al., *Hybrid ANN Activation Function to Reduce Water Wastage in Agricultural Irrigation*, IEEE Access, vol. 13, pp. 93302–93322, 2025.
18. A. C. Rencher, *Methods of Multivariate Analysis*, 2nd ed., John Wiley & Sons, 2002.
19. Sh. Al-Jubouri and S. H. Abd, *Multivariate Analysis*, Dar Al-Kutub, University of Baghdad, 2000.
20. J. Wu, *Cluster analysis and K-means clustering: An introduction*, Springer, pp. 1–16, 2012.
21. M. N. Postorino and G. M. L. Sarné, *Cluster analysis for traffic accident research*, University of Reggio Calabria, Italy, 2002.
22. D. B. F. Filho et al., *Cluster analysis for political scientists*, Sociological Methods Research, vol. 5, no. 15, Aug. 2014.
23. W. Hardle and L. Simar, *Applied Multivariate Statistical Analysis*, Springer-Verlag, 2003.
24. J. Bu et al., *Comparative study of hydrochemical classification...*, International Journal of Environmental Research and Public Health, vol. 17, no. 24, p. 9515, 2020.
25. S. S. Muthanna and M. M. Riyad, *Cluster analysis and nearest neighbor method...*, Iraqi Journal of Statistical Sciences, vol. 21, pp. 35–52, 2012.
26. N. H. Timm, *Applied Multivariate Analysis*, Springer, New York, 2002.
27. C. R. Essary, L. M. Fischer, and E. Irlbeck, *A guide to hierarchical cluster analysis in agricultural communications research*, Journal of Applied Communications, vol. 106, no. 3, p. 3, 2022.
28. Ministry of Planning, Central Statistical Organization, *Statistics of Recorded Traffic Accidents for the Year 2015*, Baghdad, Iraq, 2015.