Transport and Communications Science Journal, Vol. 75, Issue 04 (05/2024), 1673-1687

Transport and Communications Science Journal

APPLYING A TWO-STEP CLUSTER ALGORITHM IN TRAFFIC ACCIDENT DATA ANALYSIS

Khanh Giang Le*, Ho Thi Lan Huong, Van Manh Do, Quang Hoc Tran

University of Transport and Communications, No 3 Cau Giay Street, Hanoi, Vietnam

ARTICLE INFO

TYPE: Research Article
Received: 22/07/2023
Revised: 18/12/2023
Accepted: 30/01/2024
Published online: 15/05/2024
https://doi.org/10.47869/tcsj.75.4.16

* *Corresponding author*
Email: gianglk@utc.edu.vn; Tel: 0983663031

Abstract. Cluster analysis is often employed as the initial stage in organizing heterogeneous data into homogeneous groups. Choosing an effective clustering approach and an ideal number of clusters in a traffic accident dataset might be complex and challenging. This study aims to evaluate the effectiveness of k-means and two-step cluster methods. Subsequently, the two-step cluster method and GIS are applied to analyze the traffic accident datasets from 2015 to 2017 in Hanoi, Vietnam. First, according to the Silhouette score, the two-step cluster method achieved a higher score of 0.563, while the k-means method scored 0.341. A higher Silhouette score indicates more well-defined clusters. Second, the research suggests combining the two-step cluster method with GIS for analyzing traffic accident datasets. The outcome identifies five typical types of accidents in Hanoi. In addition, the locations of various accident types were visually illustrated on a map, enabling traffic officials to recommend precise and urgent countermeasures. Importantly, the clustering results reveal that the two-step cluster method exhibits a significantly higher rate of homogeneous data in the clusters compared to the k-means method. This study demonstrates that the two-step cluster method is not only more effective than the k-means method in terms of clustering ability but also in data pre-processing. The study's results enable authorities to gain a more detailed understanding of typical traffic accident patterns in Hanoi. Besides, the employed methods could potentially be applied to other regions, providing an additional avenue for analysis.

Keywords: Traffic accident, clustering algorithms, two-step cluster, k-means, geographical information system.

1. INTRODUCTION

Currently, one of the most urgent issues in the globe is traffic accidents. In the world, traffic accidents rank as the eighth most common cause of mortality. Traffic accidents kill about 1.25 million people and injure 50 million people each year according to the Road Safety Annual Report 2022 [1]. In Vietnam, there were 11,450 traffic accidents in 2022, resulting in 6,384 deaths and 7,804 injuries [2]. A traffic accident is an unforeseen event that can occur in various situations [3]. Among the various elements influencing traffic accidents are accident types, environmental conditions, roadway designs, vehicle specifications, and driver characteristics, etc. The primary aim of accident data analysis is to identify important factors causing traffic accidents [3,4]. Traffic accident data analysis is crucial for determining characteristics related to serious accidents, providing valuable insights for safe driving recommendations. Traffic accident data analysis can identify several underlying causes of traffic accidents. Analyzing accident data is the key to understanding contributing factors, allowing for the implementation of effective preventive measures to reduce traffic accidents and enhance overall road safety.

In fact, traffic accident data analysis is really challenging because there are many different causes and types of accidents in various situations [5,6]. This leads to many valuable information being hidden while analyzing traffic accident data sets if only traditional analysis methods are used. It is really difficult to determine the important causes and factors that cause traffic accidents. Thus, data mining techniques such as clustering algorithms, classification, and association rule mining are very useful in evaluating the various features relevant to road accidents and in analyzing the various circumstances of the occurrences of accidents. In addition, identifying locations where accidents frequently occur, known as hotspots, is also important in indicating accident-related characteristics [7,8].

Determining accident hotspots can be aided by clustering accident locations [5,8]. The authorities may then be able to evaluate each of those hotspots independently and implement efficient preventive measures with the support of that hotspot analysis [9]. Furthermore, drivers can avoid accidents if they know hotspot locations [10]. The accuracy of hotspot clusters affects awareness, caution, safety, and accident avoidance [5,8,10].

In fact, traffic accident data are frequently diverse, which causes some relationships to be obscured. Traffic accident data can contain both numerical and category data. Thus, several previous research attempted to reduce variability by either concentrating on a very particular traffic accident type or developing distinct models for each traffic accident type. Expert knowledge is frequently used to segregate traffic accident data. It is not assured, however, that each cluster represents a homogeneous set of traffic accidents [11]. Thus, cluster analysis is frequently used as the initial stage in organizing heterogeneous data into homogeneous groups [11-13].

Previous studies have shown that each clustering method has its own advantages and disadvantages. Some methods can only handle small data, while others can handle large data. Some give spatial results, while some give non-spatial results. Several methods can only process numerical data, while other methods can only process category data [12]. Meanwhile, traffic accident data can contain both numerical and category data. Besides, visualizing the location of clusters will help the later analysis process be more detailed and accurate. Therefore, in this paper, the authors will analyze the limitations of widely used methods. Then, it is recommended to apply two-step cluster method combined with geographical

information system (GIS) to analyze and visualize the results as an alternative method to overcome the limitations mentioned above. Experimental evaluations based on real data sets in Hanoi during 2015 to 2017 also demonstrate that this approach outperform others in clustering various types of traffic accidents.

The remainder of this paper is organized as follows. Section 2 reviews earlier studies in applying methods to cluster traffic accident data sets. Next, the methods and data used for the research are mentioned in Section 3. Section 4 shows the results and discussions. Finally, the conclusions, limitations of the study, suggestions, and future works are also illustrated in Section 5.

2. LITERATURE REVIEW

The objective of clustering algorithm is to partition the data into distinct clusters or groups. Within each group, objects are expected to exhibit similarities, while objects in different clusters should demonstrate dissimilarities [13]. Besides, clustering techniques contribute significantly to the field of traffic accident analysis by extracting meaningful patterns and providing actionable insights. These applications help authorities, policymakers, and researchers make informed decisions to improve road safety, allocate resources effectively, and develop targeted interventions tailored to specific accident characteristics. There are four main applications of clustering techniques in traffic accident analysis such as identifying accident hotspot locations (including temporal and spatial analysis) [14-16], classifying accident types [4], understanding contributing factors to accidents [17], assessing and predicting of risks [18].

Hierarchical, grid-based, partitional, density-based, and model-based clustering algorithms are among the five primary kinds of clustering algorithms [12]. Hierarchical algorithms, such as Clustering Using REpresentatives (CURE) and Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH), are effective for summarizing and visualizing data but face scalability challenges due to the need to evaluate numerous events for each decision [12]. The BIRCH algorithm does not work well with non-spherical clusters, cannot work with non-numerical variables, and sensitive to the order of the data record [13]. Grid-based algorithms, including STatistical INformation Grid (STING) and CLustering In QUEst (CLIQUE), are suitable for identifying spherical-shaped clusters. However, only clusters with horizontal or vertical boundaries can be identified, those with oblique boundaries remain undetected. In addition, the choice of input parameters significantly impacts clustering outcomes and poses a challenge in selection [19].

Density-based algorithms, like Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Ordering Points to Identify the Clustering Structure (OPTICS) are prevalent for detecting clusters with arbitrary shapes [12]. In DBSCAN and OPTICS, the computation of density involves counting the number of events within a specified area determined by a bandwidth, and the accuracy of this calculation is markedly influenced by the chosen value for the bandwidth. Although density-based algorithms are robust to outliers, they are sensitive to density variations. They cannot select the best number of clusters automatically as well as does not work well in high dimensional space [13].

Among the partitional algorithms are k-medoids and k-means. Although this approach is easy to implement, seldom used in many applications because it requires predetermined cluster numbers. In addition, these both algorithms are unable to process categorical data

[7,12]. Moreover, the k-means algorithm takes into account the complete dataset and produces clusters with spherical shapes, which may not be ideal for accurately depicting regions prone to traffic accidents [20]. The authors of [7] applied the k-means method to identify high-frequency accident areas. This study also recommends that before analyzing accident data, the use of appropriate clustering techniques will reduce the variability of the data and may help reveal hidden information. Nevertheless, this study applied only partitioning clustering methods. The authors of [21] applied k-means to group the combinations of the four indicators into categories with homogeneous effects on run-off-road injury crashes frequency and severity. This study did not group by types of accident. A disadvantage of this algorithm is that the number of clusters required must be determined in advance, which may result in the dataset being incorrectly clustered. Besides, the study of [4] applied k-modes algorithm, a modification of the k-means algorithm, that swaps clusters for modes. The k-modes algorithm is capable of handling categorical data but lacks the ability to handle both numerical and categorical data simultaneously. Furthermore, like the k-medoids and k-means, the k-modes also need a known number of clusters.

The two-step cluster method enables us to address these limitations. The study of [22] applied the two-step cluster method to cluster the profile of bank customers. However, this approach is still not commonly used in the analysis of traffic accident datasets. Besides, visualizing the locations of clusters and the positions of each traffic accident within each cluster is really challenging. Since a long time ago, many different types of studies on road safety have used geographic information systems (GIS). Because traffic accident locations and its attributes are routinely maintained in a GIS, we may quickly search for and look into possible causes [23]. Consequently, there is potential for data mining techniques and GIS integration in various disciplines, particularly in the analysis of traffic accident.

3. DATA AND METHODS

3.1. Data

3.1.1. Data Collection

The investigation was conducted in Hanoi that is the capital of Vietnam. Currently, with an area of 3,359.82 km² and a population of 8.4 million people, Hanoi is the largest city in Vietnam. The primary modes of transportation in Hanoi include motorbikes, buses, taxis, and an increasing number of private vehicles. According to statistics from the Hanoi Department of Transport, the Hanoi's infrastructure system is serving about 7.9 million vehicles (of which, 1.1 million cars, 6.6 million motorbikes and 0.2 million electric vehicles). The annual average growth rate is over 10% per year for cars, over 3% per year for motorbikes. In addition, there are about 12 million vehicles from other provinces participating in traffic in Hanoi [24].

Hanoi has one of the most complex transportation systems, along with a diversity of serious traffic accident situations. According to the Hanoi Department of Transport, there is a substantial risk of traffic accidents since Hanoi's transportation infrastructure cannot keep up with the current rate of urbanization. Hanoi has been facing a huge challenge that is traffic accidents [25]. The analysis to determine the primary causes of accidents has not yet been taken into account. Therefore, the authors decided to choose Hanoi for investigation.

The traffic accident dataset for this research is drawn from all officially police-reported traffic accidents in Hanoi city over three years (2015−2017) (1132 accidents). Several

previous studies have demonstrated that analyzing traffic accident data collected over a three-year period is appropriate [11, 26-27]. This dataset includes the majority of traffic accidents, which are typically not resolved by the parties involved, involving serious injury or death and reported by the police, instead of encompassing all accidents that happened. Data on slight traffic accidents, often resolved by the parties involved, are either not collected or recorded incompletely. To facilitate the investigation and analysis of this dataset, the authors assume that the user of the first vehicle is responsible for causing the accident (first party), while the user of the second vehicle is identified as the second party (victim). The dataset comprises both categorical and numerical data. The dataset contains significant details such as the date and time of accident occurrences, accident locations, types of accidents, causes of accidents, severity levels, the first party's age, gender, vehicle type, and status, as well as the victim's age, gender, vehicle type, and status. In addition, it includes information on the number of people injured and the level of injury involved.

3.1.2. Data Preparation

Data mining approaches emphasize noise elimination, handling missing values, and discarding irrelevant features. Therefore, firstly, the dataset needs to undergo a cleaning process before the analytical phase can commence. The data preparation procedure renders the dataset accessible for further research [12]. The final pre-processed dataset comprises 18 variables, deemed sufficient for the investigation. Tab. 1 provides an illustration of the dataset's description.

Table 1. The variables of traffic accidents.

Type	Variable
Numerical	Age; Speed limit; Number of victims.
Categorical	Vehicle type; Accident type; Reason; Severity index; Consequence; Gender; Crossroad; Populated area; Road type; Road sort; Surroundings; Weekend; Hour; Season; Road surface.

3.2. Methods

3.2.1. The two-step cluster method

Because the two-step cluster approach in SPSS statistics software is suitable for processing big datasets with both numerical and categorical variables, it was used in this investigation. Moreover, the optimal number of clusters is automatically determined [28]. Pre-clustering, resolving outliers, and clustering are the three processes in this method. First, the dataset is scanned using Euclidian and log-likelihood distance criteria to determine whether the current data can be grouped into an existing cluster or whether it needs to form a new one. Second, values that are inappropriate for any situation will be classified as outliers and removed. Sub-clusters are organized into the desired number of clusters in the third stage. Traditional clustering approaches can be successfully used in this step because the number of sub-clusters is significantly less than the number of original data [29].

Both numerical and categorical variables can use log-likelihood distance. It is assumed that the category variables have multinomial distributions and that the numerical variables have normal distributions to calculate this value. The variables are also independent of one another. The distance between clusters i and j is given as follows [30]:

$$d(i,j) = \xi_i + \xi_j - \xi_{<i,j>} \quad (1)$$

with

$$\xi_s = -N_s \left(\sum_{k=1}^{K^A} \frac{1}{2} log(\hat{\sigma}_k^2 + \hat{\sigma}_{sk}^2) + \sum_{k=1}^{K^B} \hat{E}_{sk} \right) \quad (2)$$

with

$$\hat{E}_{sk} = -\sum_{l=1}^{L_k} \frac{N_{skl}}{N_s} log \frac{N_{skl}}{N_s} \quad (3)$$

where $d(i,j)$ is the distance between clusters i and j; $< i,j >$ index that denotes the cluster generated by co-ordinating clusters i and j; K^A is the total number of numerical variables; K^B is total number of categorical variables; L_k is the number of categories for the *k-th* categorical variable; N_s is the total number of data records in cluster s; N_{skl} is the number of records in cluster s whose categorical variable k takes l category; $\hat{\sigma}_k^2$ the estimated variance (dispersion) of the continuous variable k, for the entire dataset; $\hat{\sigma}_{sk}^2$ the estimated variance of the continuous variable k, in cluster j.

The Schwarz's Bayesian Information Criterion (BIC) indicator is computed for each number of clusters within a specific range to determine the optimal number of clusters [30]. Then, using that indicator, a preliminary calculation of the number of clusters is made. By determining the largest change in distance between the two nearest groups throughout each stage of hierarchical clustering, the preliminary estimate is finally improved.

$$BIC(J) = -2 \sum_{j=1}^{J} \xi_j + m_J \, log(N) \quad (4)$$

with

$$m_J = J \left(2K^A + \sum_{k=1}^{K^B} (L_k - 1) \right) \quad (5)$$

where $BIC(J)$ is total number of clusters; J is the number of clusters; m_J is ratio in j cluster developed during the hierarchical clustering step; ξ_j is the j^{th} cluster variance.

In the last phase, statistics, including the chi-square test for categorical variables and the t-statistics for numerical ones, were used to quantify the relative contribution of each variable to the creation of a cluster.

3.2.2. The k-means method

The k-means algorithm is a cyclic algorithm where clusters are iteratively updated until the most suitable solution is achieved. The general logic of the k-means algorithm involves dividing a dataset composed of n data objects into k clusters, determined based on preliminary information and the researcher's experience. The objective is to maximize intra-cluster similarity while minimizing inter-cluster similarity. Cluster similarity is calculated using the mean value of objects [31]. The summarized k-means procedure is outlined as follows [32]:

Input:

- k: the number of clusters;
- D: a dataset containing n objects.

Output: A set of k clusters.

Implementation of the k-means algorithm:

Step 1: Select the value of k to decide the number of clusters to be formed.

Step 2: Select random k points that will act as cluster centroids.

Step 3: Assign each data point, based on their distance from the randomly selected points (Centroid), to the nearest centroid, which will form the predefined clusters.

Step 4: Place a new centroid of each cluster.

Step 5: Repeat step 3, which reassigns each datapoint to the new closest centroid of each cluster.

Step 6: If any reassignment occurs, then go to step 4; else, go to step 7.

Step 7: Finish

The right number of clusters can be selected by the Within-Cluster-Sum-of-Squares (WCSS) method. WCSS stands for the sum of the squares of distances of the data points in each and every cluster from its centroid. The main idea is to minimize the distance between the data points and the centroid of the clusters. The process is iterated until we reach a minimum value for the sum of distances. Here are the steps to follow for determining the optimal number of clusters using the elbow method [33]:

Step 1: Execute the k-means clustering on a given dataset for different k values (ranging from 1-10).

Step 2: For each value of k, calculate the WCSS value.

Step 3: Plot a curve between WCSS values and the respective number of clusters k.

Step 4: The sharp point of bend or a point (looking like an elbow joint) of the plot, like an arm, will be considered as the optimal value of k.

3.2.3. Cluster quality examination

Evaluating the effectiveness of clustering methods involves considering various metrics and aspects of the clustering results. There are many methods to evaluate the effectiveness of clustering techniques. In this study, the authors employed a commonly applied method, the Silhouette score. The Silhouette's value evaluates how similar an object is to its cluster (cohesion) in comparison to other clusters (separation). Cluster cohesion illustrates the average distance between a sample and all other data points within the same cluster. Conversely, cluster separation describes the average distance between a sample and all other data points in the nearest cluster [34]. This metric spans from −1 to 1, and the Silhouette’s value delineates the following [35]:

- Poor classification from −1.0 to 0.2;
- Fair classification from 0.2 to 0.5;
- Good classification from 0.5 to 1.0.

In other words, a Silhouette’s value of 1 implies that the object is notably distant from other clusters. A Silhouette’s value of 0 signifies that the object lies between two neighboring clusters. Conversely, a Silhouette’s value less than 0 suggests that those objects have been erroneously assigned to the wrong cluster. In simpler terms, a higher value indicates that the object aligns better with its cluster than with other clusters. The Silhouette’s value is computed as follows [34]:

$$S(i) = \frac{b(i)-a(i)}{max(b(i),a(i))} \quad (6)$$

where $S(i)$ is Silhouette coefficients for i^{th} object; $a(i)$ is average of the minimum distance between i^{th} object in the same cluster; $b(i)$ is average of the minimum distance between i^{th} object in a different cluster.

3.2.4. Structured query language (SQL) in GIS

Utilizing spatial data is crucial for analyzing traffic accidents. Visualizing the locations of clusters and the positions of each traffic accident within each cluster is really challenging. It becomes simpler to store, manipulate, query, analyze, and visualize spatial data in a GIS. The preferred data language for relational database management systems is Structured Query Language (SQL). Although it similarly resembles relational algebra, it is based on tuple relational calculus. For managing and accessing databases, SQL is a widely used programming language. After the traffic accident datasets are grouped into homogeneous groups, they are spatially visualized through GIS tools. A subset of data can be defined using SQL to perform analytics. ArcGIS 10.5 software was used to conduct this investigation.

4. RESULTS AND DISCUSSIONS

4.1. The two-step cluster method

The BIC indicator is used to determine how many clusters to select. Tab. 2 displays cluster selection criteria. The optimal number of clusters is selected based on the highest rate of distance measurements. In this case, the highest rate of distance measurements is 1.875 corresponding to the five clusters. As a result, five clusters were automatically selected.

Table 2. Criteria for choosing clusters.

Number of Clusters	BIC	BIC Change[a]	Ratio of BIC Changes[b]	Ratio of Distance Measures[c]
1	73,575.957			
2	64,132.764	-9,443.194	1.000	1.530
3	59,025.764	-5,107.000	.541	1.625
4	57,615.518	-1,410.245	.149	1.064
5	56,476.769	-1,138.750	.121	1.875
6	57,303.115	826.346	-.088	1.384
7	58,752.836	1,449.721	-.154	1.163
8	60,430.484	1,677.648	-.178	1.004
9	62,114.310	1,683.825	-.178	1.154
10	63,983.150	1,868.840	-.198	1.199

Note: a is the modifications relate to the table's former cluster count; b is relative to the change for the two clusters solution, the ratios of changes; c is based on the comparison of the current number of clusters to the previous number of clusters, the distance measure ratios are calculated.

Fig. 1 showed that the quality of the cluster is quite good because the rate of sizes of the largest cluster to the smallest one is 3.33 (smallest compared to the other alternatives). Besides, an advantage of this method is that the centroids for each numerical variable and the frequencies for each categorical variable were presented separately.

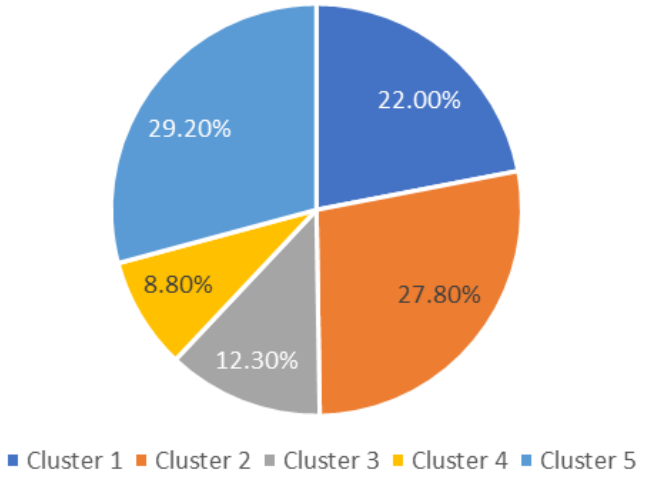

Figure 1. Cluster sizes based on the two-step cluster method.

The five-cluster model provided the distributions of cluster-dependent univariate variables. Consequently, a specific traffic accident type was identified for each cluster. The skewed feature distributions that varied among the clusters were considered to accurately represent each cluster. Traffic accidents can be assigned to multiple clusters based on feature probabilities. Tab. 3 enumerates the attributes and their probabilities in each of the five clusters used to create the model.

Table 3. The likelihood of characteristics within each cluster.

Variables	Values	Clusters (%)				
		1	2	3	4	5
Vehicle type of the first party	Truck, car	80	60	5	0	5
	Motorbike	10	30	86	93	82
The presence of the victim		100	100	100	0	100
Vehicle type of the victim	Motorbike	80	80	89	0	20
	Truck, car	0	5	3	0	70
Consequence of the first party	Fatal	1	2	36	86	90
Consequence of the victim	Fatal	90	87	63	0	2
Road type	Local street	90	8	8	40	20
	National, provincial road	10	90	0	40	75
	Country lane	0	2	92	20	5

As illustrated in Tab. 3, in cluster 1, 80% of the victim's vehicle type is a motorbike, while the first party's vehicle type consists of 80% trucks and cars. Traffic accidents in this cluster predominantly occur on local streets, accounting for 90%. The death rate of victims in this cluster is 90%. Therefore, the authors refer to this cluster as "Traffic accident between a truck/car and a motorbike on local streets".

Cluster 2 shows similarity with cluster 1 regarding to vehicle types. Specifically, the victim's vehicle type in this cluster is predominantly motorcycles, constituting 80%. Meanwhile, the vehicle type of the party causing the accident is also mainly trucks/cars, making up 60%, and motorcycles accounting for the remaining 30%. However, it is

noteworthy that accidents in cluster 2 primarily occurred on national and provincial road, accounting for 90%. The death rate of victims in this cluster is 87%. Consequently, cluster 2 is described as "Traffic accident between a truck/car and a motorbike on national and provincial roads".

Cluster 3 is different from the initial two clusters in terms of the road type, specifically country lanes, constituting 92% of all instances. Besides, the vehicles of both the accident-causing party and the victim in this cluster utilized motorcycles, accounting for approximately 86% and 89%, respectively. Thus, this cluster is characterized as "Traffic accidents involving two motorbikes on country lanes."

In cluster 4, all accidents are self-caused without the involvement of any other vehicles. According to Tab. 3, this cluster stands out from other groups as motorcycles were the vehicle of the user in 93% of the cases. The death rate in this cluster is 86%. The authors label this cluster as "Single-vehicle motorbike accidents."

Cluster 5 is characterized by the fact that the motorbikes (82%) was responsible for causing the accident (due to illegal driving), whereas the second vehicle involved was legally driven. The death rate for motorbike drivers is 90%. Hence, this cluster is described as "Motorbikes causing accidents on streets, provincial, and national roads."

Tab. 3 shows the valuable findings from the five-cluster model. The two-step cluster method enables us investigate the sizes and types of traffic accidents that correlate to each cluster. Besides, this clustering method also identifies important features such as the status of the victims as well as that of the first party. The results further support the idea that when segmenting traffic accident data, the type of vehicles and the kind of roads should be taken into account. Unlike the previous studies, especially, the study of [7] utilized a k-means algorithm to categorize accident locations into three groups: high-frequency, moderate-frequency, and low-frequency. The k-means algorithm employs accident frequency counts as parameters for the clustering of these locations. However, the studies of [14,16,20] show that density-based clustering algorithms and spatial analyses outperform the k-means algorithm when applied to determine accident hotspot locations. The authors of [21] applied k-means to group the combinations of the four indicators into categories with homogeneous effects on run-off-road injury crashes frequency and severity. This study did not group by types of accident.

4.2. The k-means method

The basic requirement for cluster analysis is to determine the number of clusters to be formed by clustering algorithm. The elbow method is applied to find the optimal number of clusters. We generated 10 models for 1 cluster to 10 clusters. Fig. 2 illustrates the outcome of the elbow method for the 10 models generated. It shows that there is a reduction in the values of WCSS with an increase in the number of clusters. Based on Fig. 2, we selected the model with 4 clusters at the elbow joint, as no further improvement was observed beyond this point. Our selection also follows the approach used by previous studies [11,36].

Fig. 3 illustrates the size of 4 clusters based on the k-means method. Fig. 3 showed that the quality of the cluster is not good because the rate of sizes of the largest cluster to the smallest one is 5.88 (higher compared to the two-step cluster method).

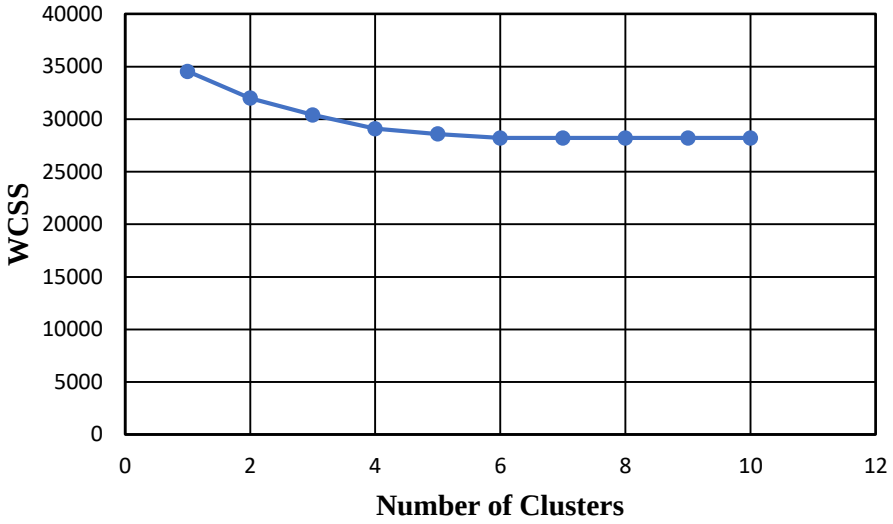

Figure 2. The cluster models generated based on the elbow method.

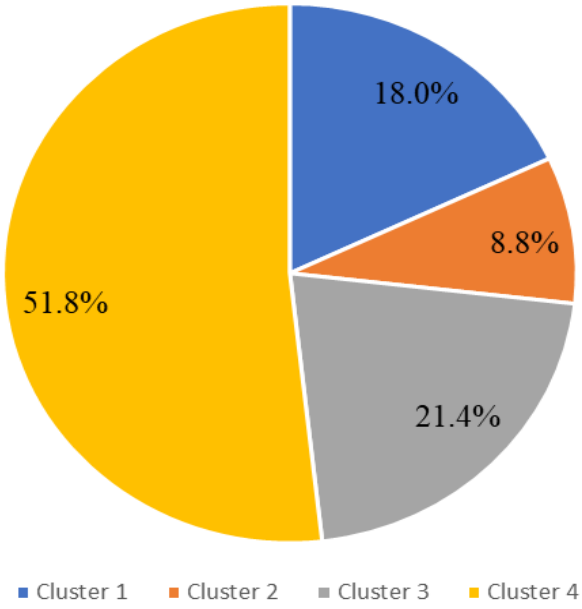

Figure 3. Cluster sizes based on the k-means method.

After creating a number of clusters, we implement the k-means algorithm in Jupyter Notebook programming to segment the accident dataset. Once we achieved the appropriate segmentation of the dataset, our next task was the characterization of each cluster. A thorough analysis of each cluster revealed that accident variables categorizing the clusters were four different accident types. A brief description of the clusters is given below.

Cluster 1 represents traffic accidents involving the first party (cars/trucks), accounting for 61.5%, and the victim (motorbikes), accounting for 62.4%. Cluster 2 is single-vehicle motorbike accidents, accounting for 75%. Cluster 3 represents accidents involving the first party (cars/trucks), accounting for 70%, and the victims (motorbikes), accounting for 79.6%. Cluster 4 includes accidents caused by the first party (motorbikes), accounting for 80.6%,

with the majority of victims being motorbikes (30.3%), followed by trucks (22.6%), and the remaining cases involve pedestrians, cars, and bicycles.

4.3. Cluster quality examination between the two-step cluster and the k-means methods

Evaluating the effectiveness of clustering methods involves considering various metrics and aspects of the clustering results. First, the Silhouette score was applied to evaluate the effectiveness of the two clustering methods, two-step cluster and k-means. The results indicate that the Silhouette score for the two-step cluster is 0.563, whereas the Silhouette score for the k-means is 0.341. A higher Silhouette score denotes better-defined clusters [35]. Specifically, the Silhouette score for the two-step cluster is 0.563, falling within the range from 0.5 to 1.0, resulting in a good classification outcome. Meanwhile, the Silhouette score for the k-means is 0.341, falling within the range from 0.2 to 0.5, resulting in a fair classification outcome. Therefore, the clustering efficiency of the two-step cluster method is better than that of the k-means method.

In addition, we can rely on the clustering results of the two methods to evaluate both the cluster quality and the clustering method. To simplify the analysis of clusters in both methods, the authors emphasize that cluster 1 in both methods shares nearly identical characteristics in terms of vehicle type. Furthermore, in the k-means method, Cluster 2 aligns with cluster 4 in the two-step cluster, while cluster 3 in the k-means method corresponds to cluster 2 in the two-step cluster. Finally, cluster 4 in the k-means method corresponds to the combination of clusters 3 and 5 in the two-step cluster.

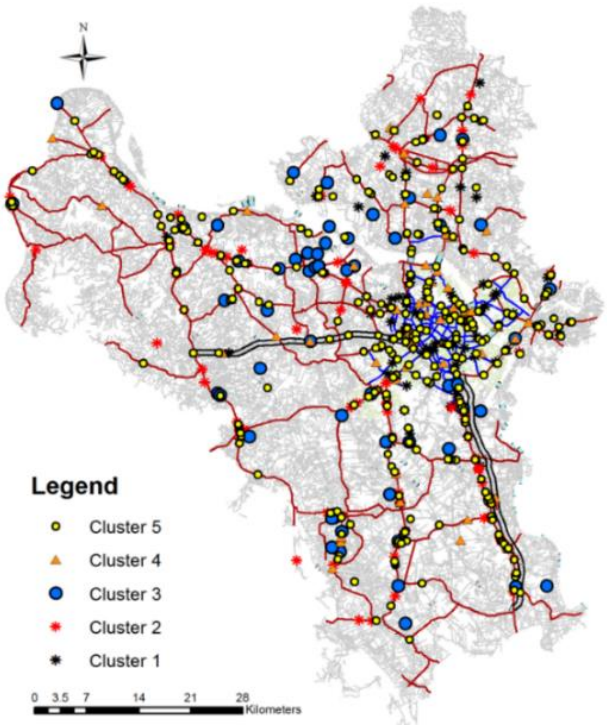

Figure 4. Locations of clusters of traffic accidents studied in Hanoi.

According to the clustering results, we observe that in the k-means method, cluster 1 groups the first party (cars/trucks) for 61.5% and the victim (motorbikes) for 62.4%, whereas in the two-step cluster method, cluster 1 accounts for 80% for the first party (cars/trucks) and 80% for the victim (motorbikes). This indicates that the rate of homogeneous data in cluster 1 using the two-step cluster method is significantly higher than that of the k-means method.

Likewise, in the k-means method, cluster 2 is single-vehicle motorbike accidents, accounting for 75%, whereas in the two-step cluster method, cluster 4 is single-vehicle motorbike accidents, accounting for 100%. It means all accidents are self-caused without the involvement of any other vehicles in this cluster. Although in terms of cluster size, cluster 2 in the k-means method is equal to cluster 4 in the two-step cluster method. Specifically, in the k-means method, cluster 4 is quite large, accounting for 51.8% of the total. Meanwhile, with the two-step cluster, this cluster is separated into 2 smaller groups with distinct characteristics. Importantly, the clusters in the two-step cluster method also have additional important characteristics including road type and mortality rate of victims in each cluster. Hence, the clustering effectiveness of the two-step cluster method surpasses that of the k-means method.

4.4. Location of clusters based on GIS

These clusters were created and displayed in GIS using SQL. The locations of the clusters and the accidents inside each cluster are shown in Fig. 4. The investigation of contributing elements is simple when we are aware of the types and locations of accidents.

5. CONCLUSIONS, SUGGESTIONS, LIMITATIONS, AND FUTURE WORKS

First, the study compared the effectiveness of two clustering methods: k-means and two-step cluster. According to the Silhouette score, the two-step cluster method achieved a higher score of 0.563, while the k-means method scored 0.341. A higher Silhouette score indicates more well-defined clusters. Moreover, based on the clustering results, it becomes evident that the two-step cluster method exhibits a significantly higher rate of homogeneous data in the clusters compared to the k-means method. Consequently, the study concludes that the two-step cluster method is more effective than k-means in terms of clustering ability. Furthermore, the two-step cluster method holds a notable advantage over the k-means, k-modes, and k-medoids method in data preparation, as it not only handles both category and numerical data but also determines the ideal number of clusters automatically.

Second, the research suggests combining the two-step cluster method with GIS for analyzing traffic accident datasets. The outcome identifies five typical types of accidents in Hanoi. In addition, the locations of various accident types were visually illustrated on a map, enabling traffic officials to recommend precise and urgent countermeasures. The study also confirms that segmenting the traffic accident dataset into homogeneous clusters facilitates easier and more accurate identification of the reasons. The outcomes further support data segmentation based on vehicle types, road types, and accident types.

Finally, the data aspect is the limitation of this study. Our dataset is old, and the observation period is not sufficiently long. This may result in findings that fail to capture the diverse situations in traffic accidents. These limitations should be acknowledged and addressed in future studies. Nevertheless, despite these constraints, the authors remain hopeful that the results of this research will assist authorities in gaining a more detailed understanding of typical traffic accident patterns in Hanoi. Moreover, the employed methods could potentially be applied to other regions, offering an additional avenue for analysis.

ACKNOWLEDGMENT

This research is funded by University of Transport and Communications (UTC) under grant number T2022-CT-008TD. The authors would like to give many thanks and acknowledge the support from the project: “Master Plan of Road traffic safety in Hanoi” by the Ministry of Transport, National Traffic Safety Committee in collaboration with Korea Ministry of Infrastructure, Land and Transport, Korea Transportation Institute (KOTI), Korea

Institute of Civil Engineering & Building Technology (KICT), and the Sustainable Urban Development Joint Stock Company (SUD) for providing traffic accident data in Hanoi.

REFERENCES

[1]. Road Safety Annual Report 2022. https://www.itf-oecd.org/road-safety-annual-report-2022, (accessed 15 November 2023).
[2]. In 2022, Handle more than 2.8 million cases of traffic violations and fine more than 4,124 billion VND. https://baochinhphu.vn/nam-2022-xu-ly-hon-28-trieu-truong-hop-vi-pham-giao-thong-phat-tien-hon-4124-ty-dong-102221223112959466.html, (accessed 15 November 2023).
[3]. M. Amiruzzaman, Prediction of traffic-violation using data mining techniques, in Proceedings of the Future Technologies Conference (FTC), Vancouver, Canada, (2018) 15-16. https://doi.org/10.1007/978-3-030-02686-8_23
[4]. S. Kumar, D. Toshniwal, A data mining framework to analyze road accident data, J. Big Data, 2 (2015) 1–18. https://doi.org/10.1186/s40537-015-0035-y
[5]. M. Mashfiq Rizvee, M. Amiruzzaman, M. R. Islam, Data Mining and Visualization to Understand Accident-Prone Areas, in Proceedings of International Joint Conference on Advances in Computational Intelligence, Singapore, (2020) 20–21. https://doi.org/10.48550/arXiv.2103.09062
[6]. S. Pasupathi, V. Shanmuganathan, K. Madasamy, H. R. Yesudhas, M. Kim, Trend analysis using agglomerative hierarchical clustering approach for time series big data, J. Supercomput., 7 (2021) 1–20. https://doi.org/10.1007/s11227-020-03580-9
[7]. S. Kumar, D. Toshniwal, A data mining approach to characterize road accident locations, J. Mod. Transp., 24 (2016) 62–72. https://doi.org/10.1007/s40534-016-0095-5
[8]. T. K. Anderson, Kernel density estimation and k-means clustering to profile road accident hotspots, Accid Anal Prev., 41 (2009) 359–364. https://doi.org/10.1016/j.aap.2008.12.014
[9]. V. Prasannakumar, H. Vijith, R. Charutha, N. Geetha, Spatiotemporal clustering of road accidents: GIS based analysis and assessment, Procedia Soc. Behav. Sci., 21 (2011) 317–325. https://doi.org/10.1016/j.sbspro.2011.07.020
[10]. J. Lu, A. Gan, K. Haleem, W. Wu, Clustering-based roadway segment division for the identification of high-crash locations, J. Transp. Saf. Secur., 5 (2013) 224–239. https://doi.org/10.1080/19439962.2012.730118
[11]. B. Depaire, G. Wets, K. Vanhoof, Traffic accident segmentation by means of latent class clustering, Accid Anal Prev., 40 (2008) 1257-1266. https://doi.org/10.1016/j.aap.2008.01.007
[12]. J. Han, J. Pei, M. Kamber, Data Mining: Concepts and Techniques, Fourth ed., Morgan Kaufmann, 2023.
[13]. C. X. Gao, D. Dwyer, Y. Zhu, C. L. Smith, L. Du, K. M. Filia, J. Bayer, J. M. Menssink, T. Wang, C. Bergmeir, S. Wood, An overview of clustering methods with guidelines for application in mental health research, Psychiatry Res., 327 (2023) 115265. https://doi.org/10.1016/j.psychres.2023.115265
[14]. N. Manap, M. N. Borhan, M. R. M. Yazid, M. K. A. Hambali, A. Rohan, Identification of hotspot segments with a risk of heavy-vehicle accidents based on spatial analysis at controlled-access highway, Sustainability, 13 (2021) 1487. https://doi.org/10.3390/su13031487
[15]. S. S. A. Kazmi, M. Ahmed, R. Mumtaz, Z. Anwar, Spatiotemporal clustering and analysis of road accident hotspots by exploiting GIS technology and Kernel density estimation, Comput J., 65 (2022) 155-176. https://doi.org/10.1093/comjnl/bxz158
[16]. A. Ganjali Khosrowshahi, I. Aghayan, M. M. Kunt, A. A. Choupani. Detecting crash hotspots using grid and density-based spatial clustering, in Proceedings of the Institution of Civil Engineers – Transport, 176 (2023) 200–212. https://doi.org/10.1680/jtran.20.00028
[17]. M. Bonera, R. Mutti, B. Barabino, G. Guastaroba, A. Mor, C. Archetti, C. Filippi, M. G. Speranza, G. Maternini, Identifying clusters and patterns of road crash involving pedestrians and cyclists. A case study on the Province of Brescia (IT), Transp. Res. Procedia, 60 (2022) 512-519. https://doi.org/10.1016/j.trpro.2021.12.066
[18]. K. S. Ng, W. T. Hung, W. G. Wong, An algorithm for assessing the risk of traffic accident. J

Safety Res., 33 (2002) 387-410. https://doi.org/10.1016/S0022-4375(02)00033-6
[19]. C. Zhang, W. Huang, T. Niu, Z. Liu, G. Li, D. Cao, Review of Clustering Technology and Its Application in Coordinating Vehicle Subsystems. Automot. Innov., 6 (2023) 89-115. https://doi.org/10.1007/s42154-022-00205-0
[20]. M. R. Islam, I. J. Jenny, M. Nayon, M. R. Islam, M. Amiruzzaman, M. Abdullah-Al-Wadud, Clustering algorithms to analyze the road traffic crashes, in Proceedings of the 2021 International Conference on Science & Contemporary Technologies (ICSCT), Dhaka, Bangladesh, 5–7 August 2021. https://doi.org/10.48550/arXiv.2108.03490
[21]. J. M. Pardillo-Mayora, C. A. Domínguez-Lira, R. Jurado-Piña, Empirical calibration of a roadside hazardousness index for Spanish two-lane rural roads. Accid Anal Prev., 42 (2010) 2018-2023. https://doi.org/10.1016/j.aap.2010.06.012
[22]. D. Șchiopu, Applying Two-step cluster analysis for identifying bank customers' profile. Buletinul, 62 (2010) 66-75.
[23]. Y. Li, C. Liang, The analysis of spatial pattern and hotspots of aviation accident and ranking the potential risk airports based on GIS platform, J Adv Transp., (2018) 1–12. https://doi.org/10.1155/2018/4027498
[24]. Hanoi urban transport development, Needs a long-term vision and breakthrough approach. https://hanoimoi.vn/phat-trien-giao-thong-do-thi-ha-noi-can-tam-nhin-dai-han-va-cach-lam-dot-pha-640326.html (accessed 01 December 2023).
[25]. More than 400 people died in traffic accidents in 2022 in Hanoi. https://vtv.vn/xa-hoi/ha-noi-hon-400-nguoi-tu-vong-vi-tai-nan-giao-thong-trong-nam-2022-20230106174606764.htm (accessed 01 December 2023).
[26]. D. Endalie, W. T. Abebe, Analysis and Detection of Road Traffic Accident Severity via Data Mining Techniques: Case Study Addis Ababa, Ethiopia. Math. Probl. Eng., 2023. https://doi.org/10.1155/2023/6536768
[27]. H. Z. Selvi, B. Caglar, Using cluster analysis methods for multivariate map of traffic accidents, Open Geosci., 10 (2018) 772-781. https://doi.org/10.1515/geo-2018-0060
[28]. J. P. Verma, Data Analysis in Management with SPSS Software, Springer India, 2013. https://doi.org/10.1007/978-81-322-0786-3
[29]. G. D. Garson, Cluster analysis, Statistical Publishing Associates, 2014.
[30]. J. Bacher, K. Wenzig, M. Vogler, SPSS Two-step Cluster – A First Evaluation, 23 (2004).
[31]. J. MacQueen, Some methods for classification and analysis of multivariate observations, in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probabilities, University of California Press, Berkeley, 1967, 1, 281-296. http://projecteuclid.org/euclid.bsmsp/1200512992
[32]. J. Han, J. G. Lee, M. Kamber, An overview of clustering methods in geographic data analysis, Data Min Knowl Discov., 2 (2009) 149-170. https://doi.org/10.1201/9781420073980
[33]. M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, B. D. Satoto, Integration k-means clustering method and elbow method for identification of the best customer profile cluster, in IOP conference series: materials science and engineering, IOP Publishing, 336 (2018) 012-017. https://doi.org/10.1088/1757-899X/336/1/012017
[34]. Silhouette Coefficient, An Overview, ScienceDirect Topics. https://www.sciencedirect.com/topics/computer-science/silhouette-coefficient (accessed 01 December 2023).
[35]. A. Supandi, A. Saefuddin, I. D. Sulvianti, Two step Cluster Application to Classify Villages in Kabupaten Madiun Based on Village Potential Data, Xplore J. Stat., 10 (2021) 12–26. https://doi.org/10.29244/xplore.v10i1.272
[36]. J. D. Oña, G. López, R. Mujalli, F. J. Calvo, Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks, Accid Anal Prev., 51 (2013) 1-10. https://doi.org/10.1016/j.aap.2012.10.016