PREUČEVANJE PROSTORSKE DIMENZIJE PROMETNIH NESREČ S STATISTIČNIMI IN NESTATISTIČNIMI PRISTOPI ZA PREPOZNAVANJE ŽARIŠČ NESREČ Z GIS

ASSESSING ROAD ACCIDENTS IN SPATIAL CONTEXT VIA STATISTICAL AND NON-STATISTICAL APPROACHES TO DETECT ROAD ACCIDENT HOTSPOT USING GIS

Yegane Khosravi, Farhad Hosseinali, Mostafa Adresi

UDK: 614.86:659.2:004.6:91(55)
Klasifikacija prispevka po COBISS.SI: 1.04
Prispelo: 9. 3. 2022
Sprejeto: 4. 8. 2022

DOI: 10.15292/geodetski-vestnik.2022.03.412-431
PROFESSIONAL ARTICLE
Received: 9. 3. 2022
Accepted: 4. 8. 2022

IZVLEČEK

Prometne nesreče so po vsem svetu eden najpoglavitnejših vzrokov za smrtne žrtve, telesne poškodbe in finančno škodo. Prepoznavanje žarišč nesreč in vzrokov zanje ter izboljšanje stanja na teh žariščih je gospodaren način za izboljšanje varnosti v cestnem prometu. V tej študiji so bile za identifikacijo žarišč nesreč na cesti Dehbala v provinci Yazd v Iranu uporabljene statistične in nestatistične metode združevanja v skupine. V prvem delu študije je uteži kriterijev določal strokovnjak z metodo AHP. Prostorska korelacija naklona in ukrivljenosti je bila izračunana z Moranovim indeksom. Za prepoznavanje žarišč nesreč na podlagi gostote točk so bili uporabljeni Anselin lokalni Moranov indeks, Getis-Ord Gi in ocena gostote jedra. Tako so bile štiri vroče točke nesreč pridobljene z indeksom Anselin Local Moran, tri žariščne točke nesreč pa z Getis-Ord Gi*. Območje, ki je izpostavljeno nesrečam, je bilo pridobljeno z metodo ocene gostote jedra. Trije algoritmi, K-srednjih vrednosti (angl. K-means), K-medoids in DBSCAN, so bili uporabljeni za identifikacijo kritičnih območij ali točk z uporabo nestatističnih metod. Skupine zgostitev točk vsake metode so bile obravnavane kot skupine, kjer je verjetnost nesreč večja. Rezultate statističnih in nestatističnih metod smo med seboj primerjali in dobili končni nabor ogroženih območij. Študija razkriva vpliv geometrijskih značilnosti ceste (naklon in ukrivljenost) na pojavnost nesreč.*

KLJUČNE BESEDE

žarišča prometnih nesreč, prostorska statistika, združevanje v skupine, analiza GIS, Moranov indeks

ABSTRACT

Road accidents are among the most critical causes of fatality, personal injuries, and financial damage worldwide. Identifying accident hotspots and the causes of accidents and improving the condition of these hotspots is an economical way to improve road traffic safety. In this study, to identify the accident hotspots of "Dehbala" road located in Yazd province-Iran, statistical and non-statistical clustering methods were used. First, the weighting of the criteria was performed by an expert using the AHP method. Hence, the spatial correlation of slope and curvature was calculated by Global Moran'I. Anselin Local Moran index and Getis-Ord Gi and Kernel Density Estimation were used to identify accident hotspots based on accident location due to the density of points. As a result, four accident hotspots were obtained by the Anselin Local Moran index, three accident hotspots by Getis-Ord Gi*and one accident-prone area were obtained by Kernel Density Estimation method. Three algorithms, k-means, k-medoids, and DBSCAN, were used to identify accident-prone areas or points using non-statistical methods. The dense cluster of each method was considered as an accident-prone cluster. Then the results of statistical and non-statistical methods were intersected with each other and the final accident-prone area was obtained. This study revealed the effect of geometric charcateristics of the road (slope and curvature) on the occurance of accidents.*

KEY WORDS

Accident hotspots, Spatial statistics, Spatial clustering, GIS analysis, Moran's I

1 INTRODUCTION

Road accidents are among the most important causes of death and severe physical and financial damage worldwide. The social, cultural, and economic effects of accidents have severely threatened the lives of human societies. Over the past decades, an average of approximately 1.2 million people have been killed each year by road accidents, with about 90% of accidents occurring in developing countries (WHO, 2018). However, road safety is a global issue, and it is necessary to find ways to reduce its effect. Road accidents, in addition to annual financial losses, account for about one to three percent of countries' gross national profits, lead to the loss of national capital, and have devastating effects on the development of countries, especially low-income countries (Tessa 2009). Every year, more than 1.3 million people die in traffic accidents. The number of young people between the ages of 15 and 29 who die in road accidents each year is higher than deaths from HIV, AIDS, malaria, tuberculosis, or homicide. Iran is also one of the developing countries, and the number of road accidents in Iran is deplorable. . Unfortunately, Iran is among the deadliest countries in the world in terms of the number of victims of traffic accidents. Although the population of our country is less than one percent of the total population of the world, the percentage of road casualties in the country is about two percent of global deaths and about 1.5 times higher than the worldwide average. The number of people killed in traffic accidents in Iran is unimaginableStatistics released by the WHO on road accident statistics show that Iran is still one of the countries in the dangerous red zone in road driving and ranks fifth in road accident deaths (WHO, 2018).

This study used statistical and non-statistical clustering methods to identify accident hotspots and their effective parameters such as human mistakes, road problems, environmental factors, and car problems. Visualization capabilities, rapid data retrieval, and manipulation in GIS openned possibility for new exploratory data analysis techniques that focus on the „spatial" aspects of data. Therefore, identifying local patterns of spatial communication is essential in this regard (Anselin.L, 1995). Usually, researchers combine GIS and statistical models to evaluate risk of road accidents. However, sometimes the GIS, has been used only as a geographical database to store and represent data about accidents and road characteristics. It has also been used to represent the results of statistical studies of accidents. Another usage of GIS in accident analysing is usingstatistical tools for detecting accident-prone areas (Satria & Castro, 2016).

The primary purpose of identifying accident hotspots is to identify places where traffic accidents occur frequently. So by identifying the hotspots, their problems can be reduced with remedial measures such as geometric redesigning, installing speed reduce signs and other engineering sollutions. Identifying accident hotspots and the causes of accidents and improving conditions at these points are economical ways to improve road safety (Hauer, 2015). In this study, details of the data set and methods used in the analysis are entirely presented. Finally the implementation of the methods is applied to the available data and the results are compared with each other.

The purpose of this paper is to identify accident hotspots using statistical and non-statistical methods. Therefore, various statistical methods such as Getis-Ord Gi*, Anselin Local Moran's I, and Kernel Density estimation based on weights extracted from the AHP[1] method have been used to identify accident hotspots. Non-statistical methods such as K-means, K-medoids, and DBSCAN[2] have also been used to

[1] *Analytic Hierarchy Process*

[2] *Density-Based Spatial Clustering of Applications with Noise*

identify accident hotspots based on the geographical location of accidents. Global Moran index has also been used to determine the spatial autocorrelation of accidents factors. The dense cluster of each method was considered as an accident-prone cluster. Then the results of statistical and non-statistical methods were compared with each other. Finally, an accident-prone area was obtained for the "Dehbala" road.

Researchers and professional experts have used clustering methods broadly in road accidents for efficient accident prediction, detecting road hotspots, and accurate and less costly safety delivery. Frequently studies reported that road accidents at certain geographic places are prevalent. For example, Patil et al. (2020) used K-means clustering algorithm for accidents analysis and hotspot prediction and achieved an accuracy of 81% in accident hotspot prediction (Patil, Prabhu, Walavalkar, & Lobo, 2020). Also, Sinclair and Das used K-means algorithm for identifying patterns and the relation of variables of the data. By combining the results of K-means and KDE[3], they could extract useful information from their data (Sinclair & Das, 2021). Shen et al. (2021) used grid clustering and K-medoids to study the spatial pattern of road accidents in rural roads and then used principal components and K-means to detect road hotspots (Shen, Lu, Long, & Chen, 2019). These studies have used partitional clustering. Despite the easy process of partitional algorithms, they need the number of clusters. The number of clusters should be entered by the user. Therefore, it may cause some problems in the clustering process.

Density-based algorithm possibly works better than the partitional one. Mohammed and Baiee (2020) used the DBSCAN clustering method to detect criminal hotspots in Maryland. The GIS visualized the clustering findings, and the criminal hotspots were found. They mentioned that despite the acceptable accuracy of DBSCAN, they had difficulty with its primary parameters. (Mohammed & Baiee, 2020). Agrawal et al. (2018) used the DBSCAN method to cluster accidents. They adjusted DBSCAN parameters by trial and error and selected all the clusters except noise as accident hotspots (Agrawal, Ruth, Nandini, & Sravani, 2018).

Past studies also have used accident data to identify high spatial concentrations of accidents using GIS. (Amiri, Nadimi, Khalifeh, & Shams, 2021; Manepalli, Bham, & Kandada, 2011; Prasannakumar, Vijith, Charutha, & Geetha, 2011; Steenberghen, Dufays, Thomas, & Flahaut, 2004; Xie & Yan, 2013) have used Getis-Ord Gi*, Moran's index, and KDE to identify road accident hotspots. The results showed that all of these methods could be helpful in different ways. For example, Getis-Ord Gi* can visually show the hotspots and cold spots, Global Moran's I can identify spatial autocorrelation, Anselin local Moran's I can detect clustered areas, and KDE can detect hazardous areas. Likewise, some studies used these methods individually and got acceptable results. Yang et al. (2013) used KDE for traffic rate and economic cost separately and compared the results. They finally extracted places with high economic costs but with low accident rates (Yang, Lu, & Wu, 2013a). Also, Manap et al. used Getis-Ord Gi* individually to determine accident hotspots. They found 16 hotspot locations in their case study (Manap, Borhan, Yazid, Hambali, & Rohan, 2019).

According to previous studies, statistical clustering methods have provided acceptable results; But non -statistical clustering, such as K-means, K-medoids, and DBSCAN, is less commonly used to analyse accidents. Also, none of the studies have used statistical and non-statistical methods to identify accident hotspots. By using these methods simultaneously and identifying accident hotspots by overlapping the results, the shortcomings of each method can be reduced, and more accurate results can be obtained.

[3] *Kernel Density Estimation*

2 METHODS

In the study, the AHP method is used to weigh the parameters. With this method, the parameter that has the greatest impact on accidents is selected by the relevant experts. These parameters include lighting, geometry, and weather. The spatial autocorrelation of the more important parameter is obtained by GMI[4]. Then statistical clustering methods (i.e., Anselin Local Moran's I , Getis-Ord Gi* *, and kernel density estimation) have been used to identify accident hotspots. Also, non-statistical clustering methods (i.e., K-means, K-medoids, and DBSCAN) have been used to identify accident hotspots by their locational feature. These methods are among the most used methods in identifying accident-prone areas. Thus, it is expected that the shortcommings of each method be covered by the others.

The shortcomings of K-means and K-medoids algorithms include selecting the number of clusters by the user, which can be covered by the DBSCAN algorithm. In DBSCAN clustering number of clusters is selected by two main parameters: eps and minpt. Also, K-means and K-medoids have trouble clustering data where clusters are of varying sizes and densities. This problem can be covered by the DBSCAN algorithm, which clusters the data based on density. K-means and K-medoids cannot detect outliers, but DBSCAN can. It does not mean that K-means and K-medoids do not have advantages compared to DBSCAN. Selecting the appropriate eps and minpt in DBSCAN clustering is not easy, but the number of clusters in K-means and K-medoids can be selected by evaluating indexes such as the Davies Bouldin index. They can also cluster the large datasets, which DBSCAN is not suitable for large datasets. (Govender & Sivakumar, 2020; Lytvynenko et al., 2019). Getis-Ord Gi* is very good at identifying hot spots and cold spots while, Anselin Local Moran's I can only identify positive or negative spatial autocorrelation, that is, whether the zones are similar or dissimilar that it can lead to ambiguous results. On the other hand Anselin Local Moran's I can detect outliers while Getis-Ord Gi* can only detect hotspots and cold spots (Levine, 2008). According to the advantages and disadvantages of each method, it is worth to use them together to identify reliable accident-prone points.

2.1 Data

"Dehbala" road is located in the central part of Taft city. "Dehbala" road is mountainous and has a wavy topography. "Dehbala", which is a rural road, leads to "DehBala" village, which is a tourist destination. This road is chosen due to its high traffic on holidays and the mountainous nature of the road. The presence of turns and steep slopes on this mountainous road is a reason to investigate accidents of "Dehbala". The study area is shown in Figure (1) and the accidents are shown in Figure (2).

The road accident data used for the present study has been obtained from the Transportation Organization of Yazd Province. The data set shows the locations of accidents for consecutive years 2014 to 2018 for "Dehbala" road, which are shown as geographical coordinates. Data include 110 fatal, injuries, and damages accidents. These accidents are attributed with detailed information such as place, month, date, day, time, vehicle type, type of accident, and cause of the accident. Also, some of the attributes have been attached separately, such as type of geometry, weather, and lighting condition. The pie chart of accident type is represented in Figure (3). Also the workflow flowchart of the paper is displayed in Figure(4).

[4] *Global Moran's I*

Figure1: Study area

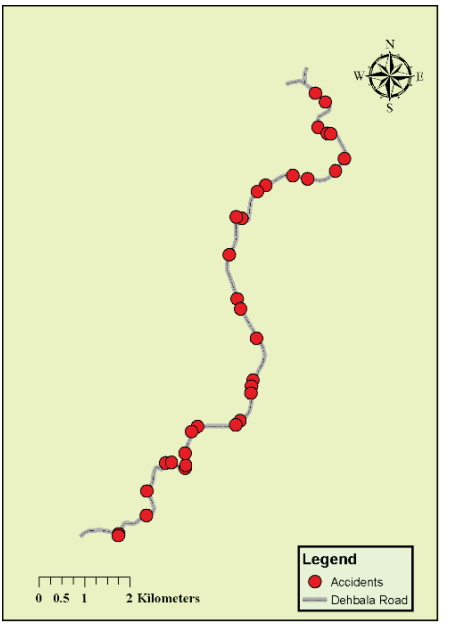

Figure 2: Accident data of "Dehbala" road.

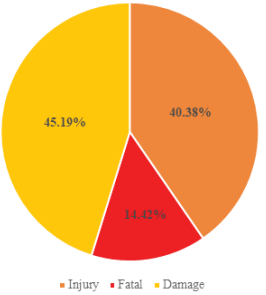

Figure 3: Percentage of accident type in "Dehbala" road.

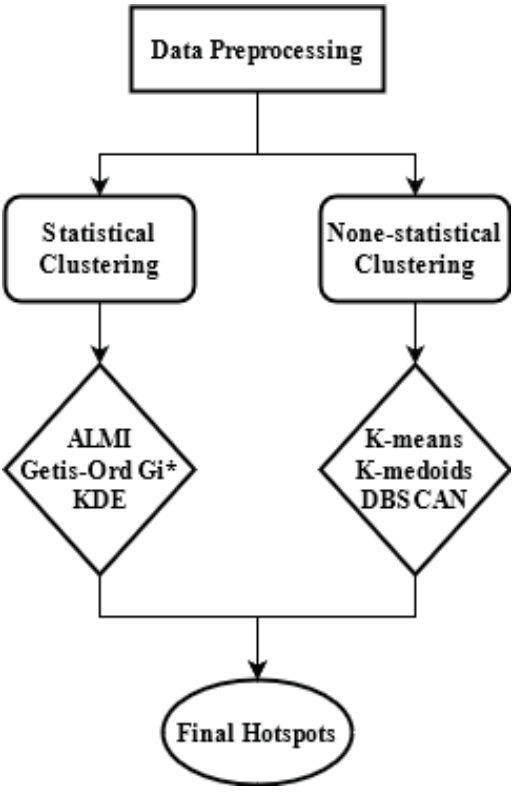

Figure 4: Workflow flowchart.

2.2 AHP

Analytic Hierarchy Process (AHP), which was initially developed in 1980, is a broadly applied multi -criteria decision-making method to determine the weights of criteria and priorities of alternatives in a structured form based on the pairwise comparison. (Liu, Eckert, & Earl, 2020). GIS and AHP have been used together in many applications. In previous studies, GIS has been used predominantly for spatial analysis while AHP has been used for either evaluating different alternatives or weighting different factors to come up with a new solution. In this study, the AHP method has been used for weighting criteria (Macharia, Wathuo, & Mundia, 2015).

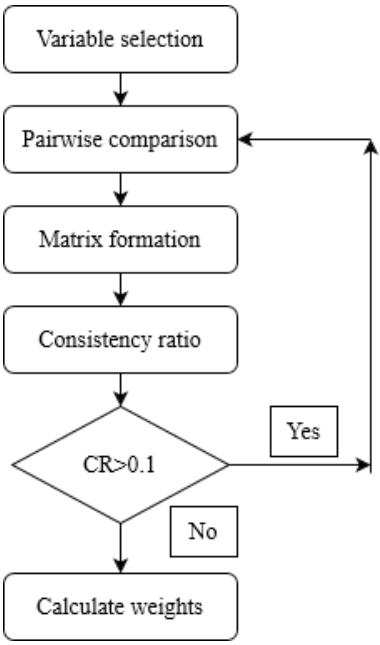

Figure 5: Flowchart for AHP weighting

In this study, the relevant experts compared the selected parameters in pairs, and the final weights are shown in Table (1). As displayed in Figure(5), in the AHP method, first the appropriate parameters must be selected. Then the parameters should be compared in pairs by relevant experts, and its results have to form a matrix to calculate the weights and the consistency ratio. If the consistency ratio (CR) was acceptable weights were calculated. Otherwise, the pairwise comparison was redone (Macharia et al., 2015). Consistency ratio is mentioned in equation (1) and (2) (Teknomo, 2006).

$$CR = \frac{CI}{RI} \tag{1}$$

$$CI = \frac{\lambda_{\max} - n}{n-1} \tag{2}$$

CI is consistency index, λ_{max} is Eigenvalue of matrix and is matrix dimension. Also, is Random Consistency index and its values for different matrix dimensions are shown in table (1) (Teknomo, 2006).

Table 1: RI values for different matrix dimension

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14
RI	0	0	0.52	0.89	1.12	1.26	1.36	1.41	1.46	1.49	1.52	1.54	1.56	1.58

2.3 Clustering

Clustering analysis is the process of dividing a heterogeneous population into several homogeneous subsets or clusters (Pedrycz, 2005). The clustering process is generally defined as an unsupervised classification in which no prior information is obtainable about classes or their number. Clustering, a common statistical data analysis technique, is used in various fields (i.e., pre-processing, Anomaly detection, artificial intelligence, pattern recognition, image interpretation, and segmentation) (Faizan, F., Ismail, & Sultan, 2020). Clustering algorithms are divided into statistical and non-statistical categories, briefly discussed in below.

2.3.1 Statistical clustering

In statistical methods, the null hypothesis is used. The null hypothesis assumes a random distribution for points. If this assumption is rejected, it will indicate the tendency of the points to form clusters or their regular arrangement. Spatial statistics help us understand geographical phenomena's behavior, processes, and patterns, discover their causes and make a more accurate decision (Aghajani, Dezfoulian, Arjroody, & Rezaei, 2017). In this study, the statistical cluster analysis includes Moran correlation analysis, Getis-Ord Gi *, and kernel density estimation.

2.3.1.1 Moran spatial autocorrelation

In many cases, we need to know whether the distribution of our data follows a specific pattern or rule. Evaluation of spatial autocorrelation methods was started mainly by the research of Moran in 1948, Anselin in 1995, and Griffith in 2008 (Anselin.L, 1995; Fischer & Griffith, 2008; Moran, 1948). The first law of geography states: "Everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). This sentence can be expanded to "all places are similar, but nearby places are more similar than distant places" (Schabus & Scholz, 2015).

Spatial Autocorrelation (SA) tests the assumption of randomness. SA is positive if the adjacent regions are similar, and it is negative when the adjacent regions are not similar. So strong SA occurs when the values of a variable, which are geographically close to each other, are related. If the features or the values of the variables related to them are randomly distributed in space, then it is expected that there should be no association between them (Mitra, 2009; Y. Zhang et al., 2022).

In this study, Moran autocorrelation analysis was used to analyze accident hotspots. This autocorrelation analysis is being meaningful by calculating Z-score and P-value parameters. There are two types of Moran index; Global Moran Index (GMI) and Anselin Local Moran Index (ALMI). GMI determines whether, on average, there is autocorrelation between a set of regions; Hence, it is used to describe the characteristics of a variable in an entire region. On the other hand, the ALMI is used to identify local clusters with positive and negative autocorrelations (hot and cold spots) and compare them with adjacent samples (Goodier, 2010; X. Zhang et al., 2021).

2.3.1.1.1 Global Moran Index

Moran's I is one of the most common statistical tools for measuring spatial correlation with returning a non-spatial correlation to a spatial context. The Global Moran Index (GMI) is used to describe the characteristics of a variable throughout the region (O'Sullivan & Unwin, 2010). In this study, the GMI is used to investigate the SA between lighting, the geometry, and weather in 5 consecutive years 2014 to 2018. If the GMI is positive, it indicates a spatial autocorrelation between the parameters. If not, it indicates no significant spatial autocorrelation between the parameters.

GMI analysis shows us whether the data is scattered or clustered. SA evaluates the distribution pattern of features in space by simultaneously considering location and features. GMI calculates the Moran index and evaluates the significance of the calculated index using the standard Z-score and P-Value scores (Wang, Liang, & Wang, 2021).

The value of the GMI is calculated by equation (3) (Morais & Gomes, 2021):

$$I = \frac{n}{s_0} \times \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij}(X_i - \overline{X})(X_j - \overline{X})}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \tag{3}$$

So that X_i and X_j show the values of the variable in places *i* and *j*. Also, $\overline{X}$ shows the average characteristic of each station. W_{ij} is the value of the spatial weight of feature s *i* and *j*. If *i* and *j* are adjacent to each other, the value of W_{ij} is equal to one. If *i* and *j* are not adjacent to each other, the value of W_{ij} is zero. also represents the sum of all elements (Morais & Gomes, 2021).

2.3.1.1.2 Local Moran Index

Local spatial statistics are beneficial for identifying the accident road hotspots. It can help identify and investigate the accumulation of unusual clusters based on a formal assessment of statistical significance and value. The local Moran index was fully developed by Anselin in 1995. It is defined in equation (4) (Morais & Gomes, 2021).

$$I_i = n\left(X_i - \overline{X}\right) \sum_{j=1,\neq i}^{n} W_{ij}\left(X_j - \overline{X}\right) \tag{4}$$

where, *n* represents the total number of variables, X_i and X_j represent the values of the desired parameter in places *i* and j, and W_{ij} represents the value of the spatial weight of feature s in *i* and *j*. A positive value of *I* indicates the spatial clustering of similar values, and a negative value indicates dissimilar spatial clustering**.** The Moran significance test is calculated by z-score. The z-score is defined in equation (5) (Dai, Guo, Zhang, & Li, 2010):

$$Z_t = \frac{I - E(I)}{SD(I)} \tag{5}$$

E(I) indicates the mean of *I* and *SD(I)* indicates the standard deviation (Dai et al., 2010).

2.3.1.2 Getis-Ord-Gi* index

The Gi * index is developed by Getis and Ord in 1995 (Ord & Getis, 1995). It calculates the Getis-Ord

Gi * index for all the features in the data for hotspot analysis. Z-value indicates in which areas the data is clustered with high or low values (Rogerson, 2015). The Z-value is similar to a simple standard deviation, and its high or low value indicates critical points. The P-value also indicates the probability of random processes generated in the observed spatial pattern (Mondal, Singh, & Kumar, 2022). Getis-Ord Gi* is defined in equation (6) (Wang et al., 2021):

$$G_i^* = \frac{\sum_{i=1}^{n} W_{ij} X_j - \frac{1}{n}\sum_{j=1}^{n} X_j \sum_{I=1}^{n} W_{ij}}{\sqrt{\frac{1}{n}\sum_{j=1}^{n} X_j^2 - \left(\frac{1}{n}\sum_{j=1}^{n} X_j\right)^2} \times \sqrt{\frac{1}{n-1}\left[n\sum_{j=1}^{n} W_{ij}^2 - \left(\sum_{j=1}^{n} W_{ij}\right)^2\right]}} \quad (6)$$

where Xj is the value for feature j, n is the sample size, and W_{ij} is the spatial weights between features i and j. A positive value of this statistic indicates a spatial cluster of high values; A negative value indicates a spatial cluster of low values (Mondal et al., 2022).

2.3.1.3 Kernel density estimation

The kernel density estimation (KDE) method was introduced by Rosenblatt in 1956. This method has attracted considerable attention in the non-parametric estimation of density. The most important advantage of this method in accident hotspots detecting is the distribution of accident risk. Hazard distribution can be defined as spreading the probability of an accident in a specific radius around the accident site due to the spatial relationship (Gelb, 2021). The density levels generated show the focal points of accidents (Deshpande, Chanda, & Arkatkar, 2011). The general equation of kernel density is given in equation (7):

$$\lambda(S) = \sum_{i=1}^{n} \frac{1}{\pi r^2} * k * \left(\frac{d_{is}}{h}\right) \quad (7)$$

where λ means area's density, h means bandwidth or KDE search radius, and k means point's weight, which means the number of accidents in the same area (Yang, Lu, & Wu, 2013b).

2.3.2 Non-statistical clustering

Non-statistical clustering includes several categories. These categories include partitional clustering algorithms, density-based clustering algorithms, model-based clustering algorithms, hierarchical clustering algorithms, and fuzzy clustering algorithms. In this study, two partitional (K-means and K-medoids) and one density-based clustering (DBSCAN) methods are used. The following is a detailed description of these methods.

2.3.2.1 K-means clustering

The K-means clustering method is one of the most common and straightforward partitioning clustering techniques. The term K-means was used by MacQueen for the first time in 1967 (MacQueen, 1967). The primary purpose of this clustering is to divide n features into K clusters so that each feature belongs to the cluster with the closest mean. This algorithm has an iterative process and stops and converges until no more points move (Ran, Zhou, Lei, Tepsan, & Deng, 2021). The workflow of the K-means algorithm is shown in Figure (6):

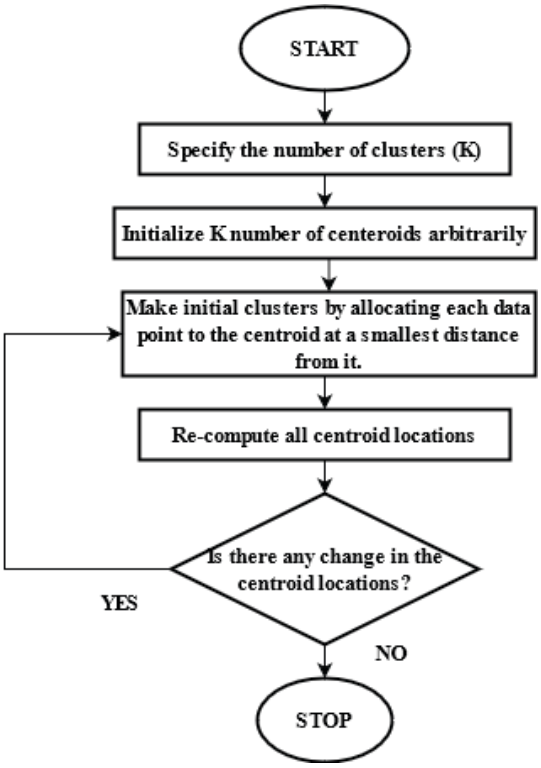

Figure 6: Flowchart of K-means clustering (Sunori et al., 2021).

One of the most important disadvantages of the K-means algorithm is a limitation to those data that uses the mass center concept. However, this limitation is solved using another clustering method called K-medoids.

2.3.2.2 K-medoids clustering

K-medoids is one of the types of partitioning clustering. It is another type of K-means algorithm and is very similar to it. The difference between K-medoids and K-means is in using the Manhattan distance instead of the Euclidean distance used in K-means. Manhattan distance properties make K-medoids less sensitive to noise than K-means (Bradley, Mogg, & Lee, 1997). Another difference between K-medoids and K-means is in the centers of the clusters. The center of the clusters in K-means is not just a real point of the data, but the average of the points in the cluster, whereas K-medoids choose the center of the cluster from the data points, which are real points .

2.3.2.3 DBSCAN

DBSCAN is the basic algorithm of density-based clustering methods. DBSCAN, introduced by Ester in 1996, can find clusters of various shapes and sizes and identify noise-containing and out-of-date data (Ester, Kriegel, Sander, & Xu, 1996). This algorithm requires two primary input parameters (Eps

and MinPts). Eps is the radius of the cluster, and MinPts is the minimum number of points in clusters. Therefore, DBSCAN algorithm is sensitive to its input parameters. The Eps parameter specifies what proximity means for points. If the Eps value is too small, no point is the center point and leads to the identification of most points as noise (Karami & Johansson, 2014).The advantages of the DBSCAN algorithm are as follows (X. Zhang et al., 2021):

- Simplicity and comprehensibility.
- No need to specify the number of clusters by the user.
- Ability to detect clusters with optional shapes.
- Ability to detect outliers and remove noise.

3 RESULTS

AHP method was used to identify the most effective parameter in the occurrence of accidents. Relevant experts used this method to weigh the criteria and sub-criteria. As shown in Table (2), the geometry criterion has more weight. These parameters have been used in this study:

- *Weather*: weather can cause accidents in many cases. For example, rainy and snowy weather increases the risk of accidents due to reduced visibility and slippery roads. Also, strong winds cause the vehicle to get out of control, and fog reduces visibility. In this study, the weather is divided into seven categories: Cloudy, Rainy, Snowy, Clear, Sandstorm, Fog, and Dust.
- *Lighting*: Lighting is provided by sunlight during the day and street lighting at night; Sunrise and sunset times and roadside lamps are important in accidents. Lighting in this research is divided into four categories: day, night, sunrise, and sunset.
- *Geometry*: Road geometry is one of the most crucial factors in road accidents. In this study, road geometry is divided into five types: Only turning, turning uphill / downhill, flat and uphill / downhill.

Table 2: Weighting of criteria and sub-criteria by AHP method by experts.

Criteria	Criteria's weight	Sub-criteria's weight
Lighting	0.333	Night (0.006), Day (0.002), Sunset (0.033) and Sunrise (0.194)
Geometry	0,606	Only turning (0.804), turning uphill / downhill (0.180), flat (0.165) and uphill / downhill (0.554)
Weather	0.606	Cloudy (0.021), Rainy (0.553), Snowy (0.853), Clear (0.014), Stormy (0.255), Fog (0.448) and Dust (0.129)

After collecting experts' opinions and calculating the weight of criteria and sub-criteria by the AHP method, according to Table (2), the most weight was allocated to geometry criteria. This means that geometry, according to relevant experts' ideas, is of great importance in road accidents. Sub-criteria of geometry include slope, curvature, the radius of arcs, etc. In this study, slope and curvature, which are significant sub-criteria of geometry, are investigated.

After identifying the effective parameter in the AHP method, the autocorrelation of its sub-criteria, which includes curvature and slope, is investigated by GMI. As can be seen in Figure (7), both sub-criteria have a cluster property, and their spatial autocorrelation is positive. Their p-values are very small, and their z-values are very large, indicating a spatial and meaningful correlation of the clustering.

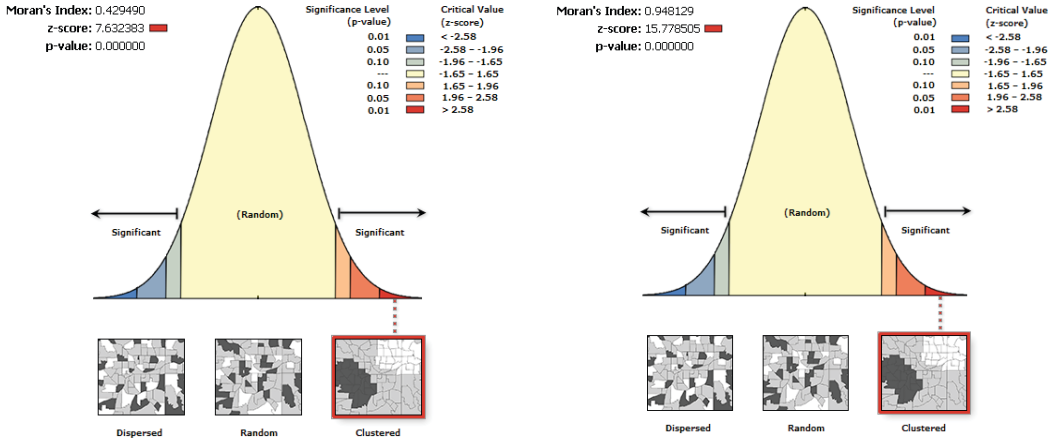

Figure 7: Output of GMI for a)Curvature b)Slope.

In order to determine the accident hotspots by the ALMI, Getis-Ord Gi* and KDE on the „DehBala" road, first, „Collect Events" was used. „Collect events" convert event data, such as crime or accident, to weighted point data. „Collect Events" turns several co-location accidents into a weighty one. The output of „collect event" is shown in Figure (8a). In the ALMI method, we use the Collect Event output to examine the clusters. Its output is shown in Figure (8b). For the Getis-Ord Gi* method, the index values are categorized from 1 to 7 by the natural breaks method. This category reflects the importance of each accident. For example, one is the most accident-prone point (red points), and seven is the safest (blue point). The result is displayed in Figure (9a). Also, the KDE output shows the high-risk are as in Figure (9b).

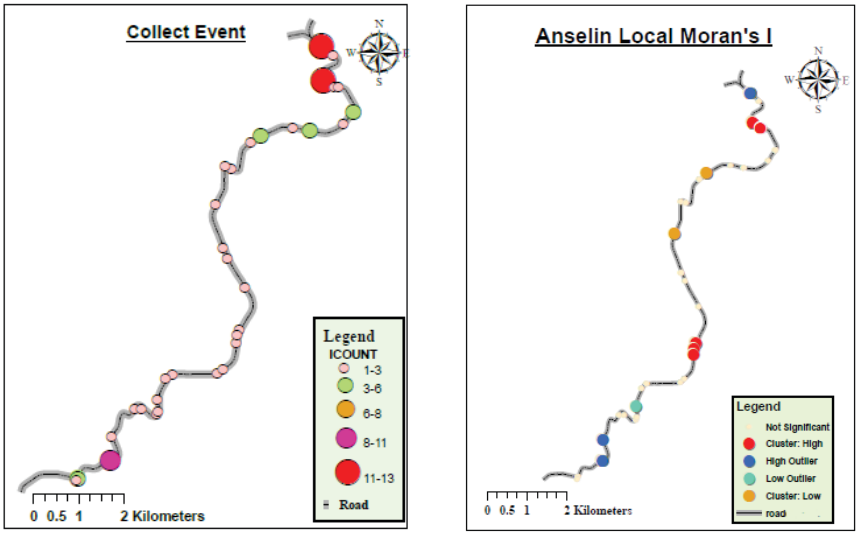

Figure 8: Outputs of a) Collect Events b) ALMI.

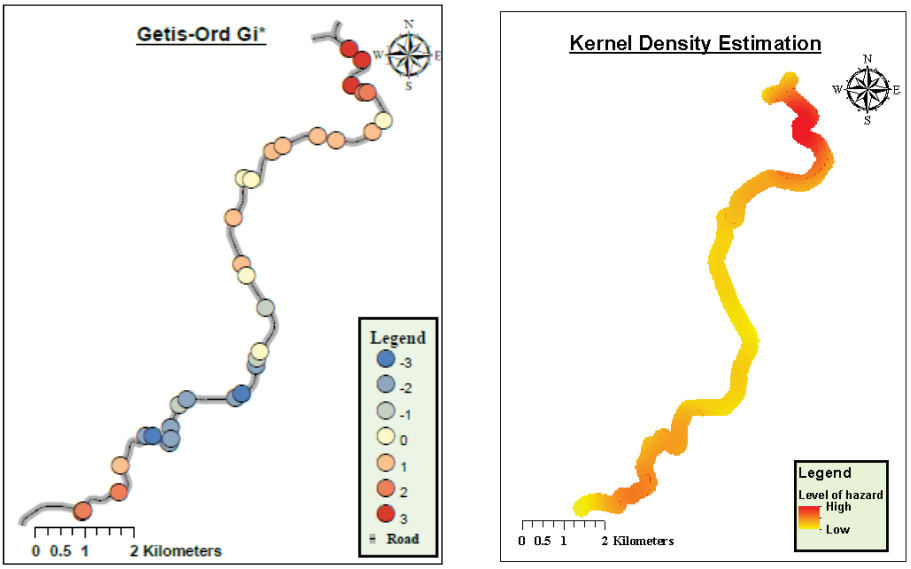

Figure 9: Outputs of a) Getis-Ord Gi* b) KDE.

Euclidean distance is used in all thesestatistical clustering methods. In this study, the Euclidean distance was calculated for the accident points and the results were compared with the difference in distance from the origin. Except for one case, the difference between the Euclidean distance and the distance from the origin was negligible.

In K-means and K-medoids, as mentioned earlier, the number of clusters is first entered by the user, and then the algorithm assigns each accident point to each cluster. Davies Bouldin's validity index was used to find the optimal number of clusters in this study. The Davies Bouldin index is one of the indicators for evaluating the validity of clusters. This index calculates the average similarity between each cluster and its most similar cluster. Therefore, the lower the value of this index, the better clusters are produced (Jumadi Dehotman Sitompul, Salim Sitompul, & Sihombing, 2019). The Davies Bouldin validation index diagram for these two methods is shown in Figure (10).

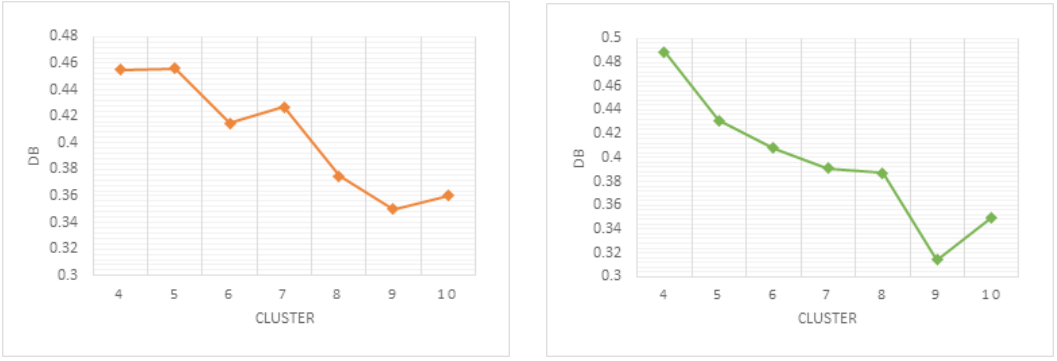

Figure 10: Davies Bouldin index of a)K-means b)K-medoids.

As shown in Figure (10), cluster number 9 is the most appropriate case for both clustering algorithms. The outputs of the K-means and K-medoids algorithms with 9 clusters are shown in Figure (11). For non-statistical clustering methods, distance from the origin has been used to cluster accidents.

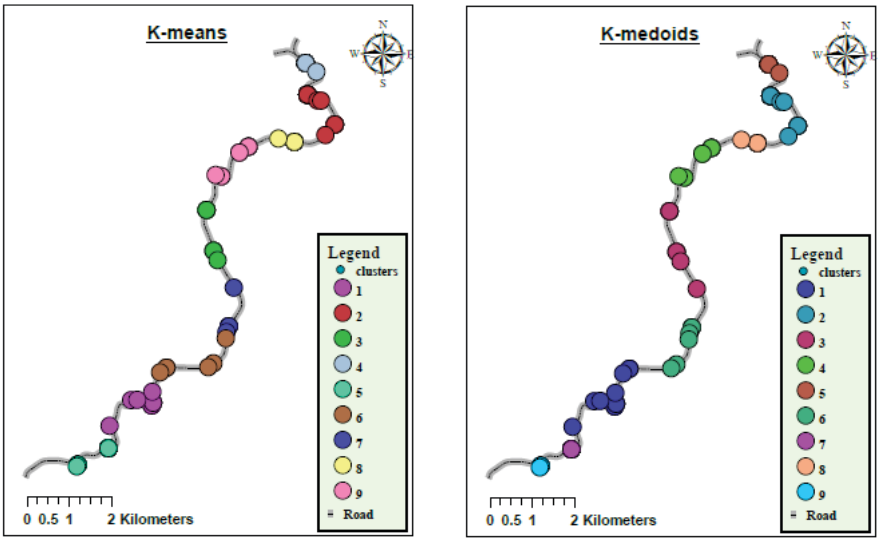

Figure 11: Outputs of a) K-means for 9 cluster b) K-medoids for 9 cluster.

The evaluation index in the DBSCAN method is experimental, and unlike K-means and K-medoids methods, there is no specific evaluation index for its validation. For this method, clusters are obtained by trial and error and changing the two parameters: Eps and MinPts (Ester et al., 1996). After trial and error by changing the algorithm parameters, the DBSCAN clustering output is displayed on the "Dehbala" road data with Eps = 0.9 and MinPts = 5 with 8 clusters in Figure (12).

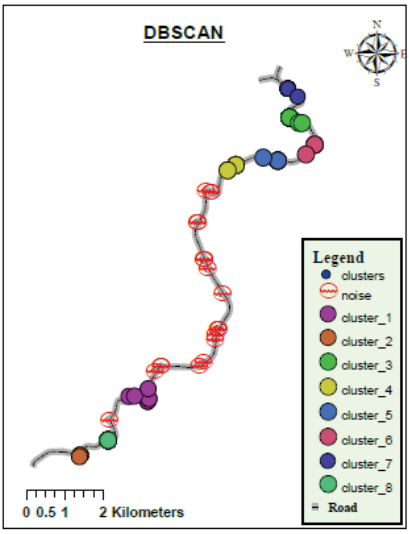

Figure 12: Outputs of DBSCAN algorithm with Eps=0.9 and MinPts=5.

4 DISCUSSION

In this research, different clustering methods have been used for identifying accident-prone areas. For this purpose, K-means, K-medoids, DBSCAN, KDE, ALMI, and Getis-Ord Gi* methods have been used to cluster the accidents. GMI method was also used to identify the effective parameter of accidents on the „Dehbala“ road. Among the applied methods in this study to cluster the accidents, Getis-Ord Gi* and ALMI indexes belong to 2D space while the road extents in 1D space. As long as the road is exactly a straight line, Getis-Ord Gi* and ALMI will work well. However, any twist in the road makes a difference between the eucidean distance and the distance along the road. According to this, the more winding the road, the less valid the 2D indexes. Despite this, the differences can be interpreted (effectless or biased) based on the local geometry of the road at each segment. In this study, the distance between the accident points (both Euclidean and along the road) were examined. The examination revealed that there is one considerable difference at a southern part of the road but it has not affected the final results (Figure 13). The answers of all methods are in agreement. Generally speaking, applied GMI and ALMI indexes are 2D indexes and not always proper for 1D spaces unquestionably.

The main results of this paper are described in detail below:

– The results of the GMI on the criteria in Figure (7) show that this index is positive for both slope and curvature. Both have spatial correlation and seems clustered. So slope and curvature are so prominent in accidents.
– In the ALMI analysis, the index value low clusters and high clusters in Figure (8b) of the road is positive; the desired feature is surrounded by similar features and has formed a cluster. Since the higher, the weight, the more critical the point, high-value clusters are more important for identifying accident-prone areas. Therefore, orange points are more accident-prone in this method.

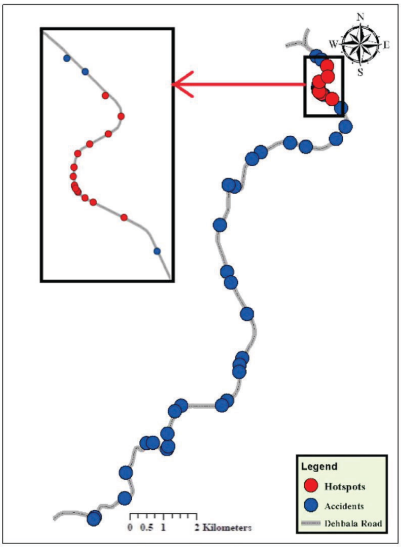

Figure 13: Final hotspot of "Dehbala" road.

– In the statistical analysis of the Gi* index, red points in Figure(9a) have been identified as the most

accident-prone road locations. These areas had a higher z-value than other areas, and the density of accidents in these areas is higher than in other areas.

– In KDE analysis, the initial part of the road, which is shown as red areas in Figure(9b), has a higher density and is known as the most accident-prone area of the road in this method.
– The results of non-statistical clustering output are shown in Table (3). As shown in Table (3), in K-means clustering, cluster number 2 with 25 crashes is the most accidental cluster. In K-medoids clustering, cluster number 2 with 23 crashes is the most accidental cluster. In DBSCAN clustering, cluster number 3 with 24 accidents is the most accident-prone area of the road.
– Finally, to comprehend the final hotspots, the accident-prone clusters of each method were identified and extracted. Then all of extracted clusters were overlapped. The communal area in all these methods was selected as the accident-prone area.. The road area identified as an accident-prone area in all methods defined as final hotspots are shown in Figure (13).

Table 3: Results of a) K-means b) K-medoids c) DBSCAN clustering

K-means	
Cluster	**Number of accidents**
1	15
2	23
3	9
4	12
5	15
6	8
7	10
8	9
9	9

a)

DBSCAN	
Cluster	**Number of accidents**
Noise	19
1	12
2	7
3	24
4	8
5	8
6	6
7	14
8	12

b)

K-medoids	
Cluster	**Number of accidents**
1	12
2	25
3	9
4	15
5	17
6	7
7	6
8	8
9	11

c)

Figure (14) shows the statistics of light, wind speed, and weather parameters at the times of accidents in the accident-prone area. According to Figure (14), most accidents occur during the day, so this factor can not be that much effective on accidents. The wind speed was also normal most of the times.

According to the statistical yearbook of Yazd, the average numbers of snowy, rainy, clear, and stormy days during the year were extracted. In Yazd province the amount of precipitation is too low (less than 59 mm per year) because of its hot and dry climate. Also, sand storms are common in Yazd province (Organization, 2021). The ratio of the number of snowy, rainy, clear, and stormy days during a year is shown in Figure (15). According to Figure (14), although in the accident-prone area, the percentage of accidents was higher in clear and cloudy weather, there is no reason that more accidents occur in clear and cloudy weather, because in Yazd province, 85% of days are clear and cloudy. As a result, according to Figure (14) and Figure (15), the ratio of clear and cloudy weathers at the times of accidents to the whole year is around 0.76, while these ratios for rainy, snowy and stormy days are 2, 7 and 2 respectively. This means that accidents are happened more likely to be occurred in rainy, snowy or stormy weathers.

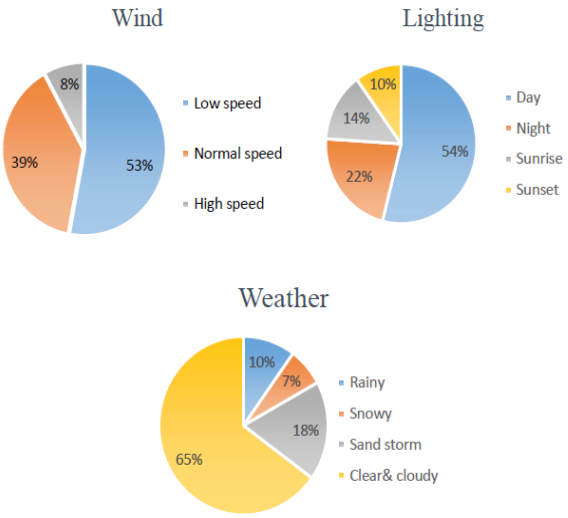

Figure 14: Statistical charts of Wind speed, Lighting, and Weather for accident-prone area

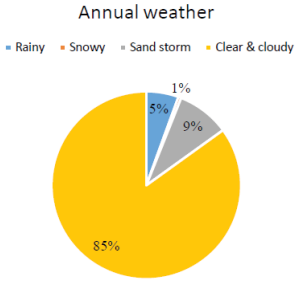

Figure 15: Annual weather in the study area

For Dehbala raod with less than 20 Kilometers length, the waethrer is not very changeable along the road. Terefore, the weather has an uniform impact along the road. Thus, it can be concluded that the accident-prone area is mostly related to the geometry (slope and curvature) of the road. However, rainy, snowy or stormy weathers intensifies the probability of accidents, especially in areas where are prone to accidents.

5 CONCLUSION

More than 1.3 million people die in traffic accidents every year. The number of people killed in traffic accidents in Iran is incredible. Thus, identifying the influential factors in accidents and identifying high-risk areas is one of the necessary measures to reduce accidents. The main objective of this paper was to find the accident-prone areas or hotspots of the "Dehbala" road. First, the weights of the criteria and sub-criteria of lighting, weather, and geometry were determined using the AHP method and the opinion of relevant experts. Geometry had the most weight. The spatial correlation of slope and curva-

ture parameters was investigated using GMI. Slope and curvature both had a positive spatial correlation. Then ALMI, Getis-Ord Gi*, KDE, K-means, K-medoids, and DBSCAN clustering have used to cluster the accident data. The hotspots of ALMI, Getis-Ord Gi*, and KDE were extracted. Also, the clusters with more accidents compared to other clusters, were extracted as accident-prone clusters in K-means, K-medoids, and DBSCAN. Finally, by comparing the results, an area, which was communal in all the methods, was identified as an accident-prone area. By investigating weather conditions at the times of accidents in the accident-prone area, it was revealed that weather condition is also an intensifier factor. Yet, the detected accident-prone part of the road has an intense curve together with high slope that increases the probability of accidents. As a result, regarding that the weather condition is uncontrollable, adjusting the slope and curvature of the accident-prone area of the road is suggested.

This paper is limited to one road, 20 kilometers long along a single county. The categorization of the hotspots was carried out in two levels (hot and cold spots). Further analyses may include longer road sections and more than two levels. It is believed that such analyses could be beneficial for traffic accident prevention and safety improvement in the future.

Acknowledgment

This work was supported by Shahid Rajaee Teacher Training University under grant numbers 3564 and 3567.

Literature and references

Aghajani, M. A., Dezfoulian, R. S., Arjroody, A. R., & Rezaei, M. (2017). Applying GIS to Identify the Spatial and Temporal Patterns of Road Accidents Using Spatial Statistics (case study: Ilam Province, Iran). Transportation Research Procedia, 25, 2126-2138. Retrieved from https://www.sciencedirect.com/science/article/pii/S2352146517307160

Agrawal, K., Ruth, V. M., Nandini, Y., & Sravani, K. (2018). Analysis of road accident locations using DBSCAN algorithm. International Journal of Scientific Research in Science and Technology (IJSRST), 4(8), 462-467.

Amiri, A. M., Nadimi, N., Khalifeh, V., & Shams, M. (2021). GIS-based crash hotspot identification: a comparison among mapping clusters and spatial analysis techniques. International journal of injury control and safety promotion, 28(3), 325-338.

Anselin.L. (1995). Local indicators of spatial association—LISA. Geographical Analysis, 27(2), 93-115.

Bradley, B. P., Mogg, K., & Lee, S. C. (1997). Attentional biases for negative information in induced and naturally occurring dysphoria. Behaviour Research and Therapy, 35(10), 911-927. Retrieved from https://www.sciencedirect.com/science/article/pii/S0005796797000533

Dai, X., Guo, Z., Zhang, L., & Li, D. (2010). Spatio-temporal exploratory analysis of urban surface temperature field in Shanghai, China. Stochastic Environmental Research and Risk Assessment, 24, 247-257.

Deshpande, N., Chanda, I., & Arkatkar, S. S. (2011). Accident Mapping And Analysis Using Geographical Information Systems. International Journal of Earth Sciences and Engineering, 4(6), 342-345.

Ester, M., Kriegel, H., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Paper presented at the KDD.

Faizan, M., F., M., Ismail, S., & Sultan, S. (2020). Applications of Clustering Techniques in Data Mining: A Comparative Study. International Journal of Advanced Computer Science and Applications, 11.

Fischer, M. M., & Griffith, D. A. (2008). Modeling spatial autocorrelation in spatial interaction data: an application to patent citation data in the European Union. Journal of Regional Science, 48(5), 969-989.

Gelb, J. (2021). Spnetwork,A package for network Kernel Density Estimation. The R Journal.

Goodier, J. (2010). The Dictionary of Human Geography. Reference Reviews.

Govender, P., & Sivakumar, V. (2020). Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). Atmospheric Pollution Research, 11(1), 40-56. doi:https://doi.org/10.1016/j.apr.2019.09.009

Hauer, E. (2015). The Art of Regression Modeling in Road Safety: Springer International Publishing.

Jumadi Dehotman Sitompul, B., Salim Sitompul, O., & Sihombing, P. (2019). Enhancement Clustering Evaluation Result of Davies-Bouldin Index with Determining Initial Centroid of K-Means Algorithm. Journal of Physics: Conference Series, 1235(1), 012015. doi:10.1088/1742-6596/1235/1/012015

Karami, A., & Johansson, R. (2014). Choosing DBSCAN Parameters Automatically using Differential Evolution. Int. J. Comput. Appl., 91. doi:10.5120/15890-5059

Levine, N. (2008). CrimeStat: A Spatial Statistical Program for the Analysis of Crime Incidents. In S. Shekhar & H. Xiong (Eds.), Encyclopedia of GIS (pp. 9.4-9.20). Boston, MA: Springer US.

Liu, Y., Eckert, C. M., & Earl, C. (2020). A review of fuzzy AHP methods for decision-making with subjective judgements. Expert Systems with Applications, 161, 113738. doi:https://doi.org/10.1016/j.eswa.2020.113738

Lytvynenko, V., Lurie, I., Krejci, J., Voronenko, M., Savina, N., & Taif, M. A. (2019). Two Step Density-Based Object-Inductive Clustering Algorithm. Paper presented at the MoMLeT.

Macharia, P., Wathuo, M., & Mundia, C. (2015). Experts' Responses Comparison in a GIS-AHP Oil Pipeline Route Optimization: A Statistical Approach. American Journal of Geographic Information System, 4, 53-63. doi:10.5923/j.ajgis.20150402.02

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Paper presented at the In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, Calif.

Manap, N., Borhan, M. N., Yazid, M. R. M., Hambali, M. K. A., & Rohan, A. (2019). Determining spatial patterns of road accidents at expressway by applying Getis-Ord Gi* spatial statistic. Int. J. Recent Technol. Eng, 8(3S3), 345-350.

Manepalli, U. R., Bham, G. H., & Kandada, S. (2011). Evaluation of hotspots identification using kernel density estimation (K) and Getis-Ord (Gi*) on I-630. Paper presented at the 3rd International Conference on Road Safety and Simulation.

Mitra, S. (2009). Spatial Autocorrelation and Bayesian Spatial Statistical Method for Analyzing Intersections Prone to Injury Crashes. Transportation Research Record, 2136(1), 92-100.

Mohammed, A. F., & Baiee, W. R. (2020). The GIS based Criminal Hotspot Analysis using DBSCAN Technique. Paper presented at the IOP Conference Series: Materials Science and Engineering.

Mondal, S., Singh, D., & Kumar, R. (2022). Crime hotspot detection using statistical and geospatial methods: a case study of Pune City, Maharashtra, India. GeoJournal. Retrieved from https://doi.org/10.1007/s10708-022-10573-z

Morais, L. R. d. A., & Gomes, G. S. d. S. (2021). Applying Spatio-temporal Scan Statistics and Spatial Autocorrelation Statistics to identify Covid-19 clusters in the world - A Vaccination Strategy? Spatial and Spatio-temporal Epidemiology, 39, 100461. Retrieved from https://www.sciencedirect.com/science/article/pii/S1877584521000605

Moran. (1948). The Interpretation of Statistical Maps. Journal of the Royal Statistical Society: Series B (Methodological), 10(2), 243-251.

O'Sullivan, D., & Unwin, D. J. (2010). Geographic information analysis and spatial data. Geographic information analysis, 1-32.

Ord, J. K., & Getis, A. (1995). Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. Geographical Analysis, 27(4), 286-306. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1538-4632.1995.tb00912.x

Organization, Y. P. M. a. P. (2021). Statistical yearbook of Yazd province in 2021: Country Management and Planning Organization, Center for Scientific Documents and Publications.

Patil, J., Prabhu, M., Walavalkar, D., & Lobo, V. B. (2020, 16-18 Dec. 2020). Road Accident Analysis using Machine Learning. Paper presented at the 2020 IEEE Pune Section International Conference (PuneCon).

Pedrycz, W. (2005). Knowledge-based clustering: from data to information granules: John Wiley & Sons.

Prasannakumar, V., Vijith, H., Charutha, R., & Geetha, N. (2011). Spatio-Temporal Clustering of Road Accidents: GIS Based Analysis and Assessment. Procedia - Social and Behavioral Sciences, 21, 317-325. Retrieved from https://www.sciencedirect.com/science/article/pii/S1877042811013437

Ran, X., Zhou, X., Lei, M., Tepsan, W., & Deng, W. (2021). A Novel K-Means Clustering Algorithm with a Noise Algorithm for Capturing Urban Hotspots. Applied Sciences, 11(23), 11202. Retrieved from https://www.mdpi.com/2076-3417/11/23/11202

Rogerson, P. (2015). Statistical methods for geography : a student's guide.

Satria, R., & Castro, M. (2016). GIS Tools for Analyzing Accidents and Road Design: A Review. Transportation Research Procedia, 18, 242-247. doi:https://doi.org/10.1016/j.trpro.2016.12.033

Schabus, S., & Scholz, J. (2015). Geographic Information Science and technology as key approach to unveil the potential of Industry 4.0: How location and time can support smart manufacturing. Paper presented at the 2015 12th International Conference on Informatics in Control, Automation and Robotics (ICINCO).

Shen, L., Lu, J., Long, M., & Chen, T. (2019). Identification of accident blackspots on rural roads using grid clustering and principal component clustering. Mathematical Problems in Engineering, 2019.

Sinclair, C., & Das, S. (2021, 21-23 Jan. 2021). Traffic Accidents Analytics in UK Urban Areas using k-means Clustering for Geospatial Mapping. Paper presented at the 2021 International Conference on Sustainable Energy and Future Electric Transportation (SEFET).

Steenberghen, T., Dufays, T., Thomas, I., & Flahaut, B. (2004). Intra-urban location and clustering of road accidents using GIS: a Belgian example. International Journal of Geographical Information Science, 18(2), 169-181.

Sunori, S. K., Negi, P. B., Maurya, S., Juneja, P., Rana, A., & Bhawana. (2021, 20-22 Jan. 2021). K-Means Clustering of Ambient Air Quality Data of Uttarakhand, India during Lockdown Period of Covid-19 Pandemic. Paper presented at the 2021 6th International Conference on Inventive Computation Technologies (ICICT).

Teknomo, K. (2006). Analytical Hierarchy Process (AHP). In.

Tessa , K. A. (2009). Kernel density estimation and K-means clustering to profile road accident hotspots. Accident analysis and prevention, 3(46), 6.

Wang, Z., Liang, L., & Wang, X. (2021). Spatiotemporal evolution of PM2.5 concentrations in urban agglomerations of China. Journal of Geographical Sciences, 31(6), 878-898. Retrieved from https://doi.org/10.1007/s11442-021-1876-2

WHO. (2018). World health statistics 2018: monitoring health for the SDGs, sustainable development goals: World Health Organization.

Xie, Z., & Yan, J. (2013). Detecting traffic accident clusters with network kernel density estimation and local spatial statistics: an integrated approach. Journal of transport geography, 31, 64-71.

Yang, S., Lu, S., & Wu, Y.-J. (2013a). Gis-based economic cost estimation of traffic accidents in St. Louis, Missouri. Procedia-Social and Behavioral Sciences, 96,

2907-2915.

Yang, S., Lu, S., & Wu, Y.-J. (2013b). GIS-based Economic Cost Estimation of Traffic Accidents in St. Louis, Missouri. Procedia - Social and Behavioral Sciences, 96, 2907-2915. Retrieved from https://www.sciencedirect.com/science/article/pii/S1877042813024488

Zhang, X., Lauber, L., Liu, H., Shi, J., Wu, J., & Pan, Y. (2021). Research on the method of travel area clustering of urban public transport based on Sage-Husa adaptive filter and improved DBSCAN algorithm. PLOS ONE, 16(12), e0259472. Retrieved from https://doi.org/10.1371/journal.pone.0259472

Zhang, Y., Rashid, A., Guo, S., Jing, Y., Zeng, Q., Li, Y., . . . Sun, Q. (2022). Spatial autocorrelation and temporal variation of contaminants of emerging concern in a typical urbanizing river. Water Research, 212, 118120. Retrieved from https://www.sciencedirect.com/science/article/pii/S0043135422000835

Khosravi Y., Hosseinali F., Adresi M. (2022). Assessing Road Accidents in Spatial Context via Statistical and non-statistical Approaches to Detect Road Accident Hotspot Using GIS. Geodetski vestnik, 66 (3), 412-431.
DOI: https://doi.org/10.15292/geodetski-vestnik.2022.03.412-431

Yegane Khosravi, M.S
Department of Surveying Engineering, Faculty of Civil Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran.
e-mail: Yeganekhosraviii76@gmail.com

Assist. Prof. Farhad Hosseinali
Department of Surveying Engineering, Faculty of Civil Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran.
e-mail: f.hosseinali@sru.ac.ir

Assist. Prof. Mostafa Adresi
Department of Geotechnical and Water Engineering, Faculty of Civil Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran.
e-mail M.adresi@sru.ac.ir