Detecting crash hotspots using grid and density-based spatial clustering

Amin Ganjali Khosrowshahi MSc

Research Assistant, Faculty of Civil Engineering, Shahrood University of Technology, Shahrood, Iran

Iman Aghayan PhD

Assistant Professor, Faculty of Civil Engineering, Shahrood University of Technology, Shahrood, Iran (Orcid:0000-0001-9113-524X)
(corresponding author: iman.aghayan@shahroodut.ac.ir)

Mehmet Metin Kunt PhD

Associate Professor, Department of Civil Engineering, Eastern Mediterranean University, Gazimagusa KKTC, Mersin 10, Turkey
(Orcid:0000-0001-5339-3811)

Abdoul-Ahad Choupani PhD

Assistant Professor, Faculty of Civil Engineering, Shahrood University of Technology, Shahrood, Iran

Data mining techniques, specifically spatial clustering methods, are used to analyse crash data and find their spatial patterns. In the present study, a grid and density-based clustering algorithm called GridBSCAN was utilised for injury crash data. Other clustering methods such as nearest neighbour hierarchical and kernel density estimation were also applied to validate the results of the GridBSCAN algorithm. Crash points recorded for Gebze and Izmit (in Turkey) were clustered through these methods. The findings revealed that GridBSCAN had the highest value for hit rate. In addition, the GridBSCAN algorithm placed data points into a grid mesh to decrease the runtime and could estimate the clusters with a higher accuracy due to the recognition of the noise points. Furthermore, the proposed approach allowed the detection of unique crash factors for both cities. The factors contributing to injury crashes in both cities included collision and junction types, along with speed limit.

1. Introduction

According to a World Health Organization report (WHO, 2015), the rate of fatalities caused by traffic crashes increased to 13% around the world during 2000–2015. Almost 1.25 million people died and 50 million were injured by road crashes worldwide. Particularly in low- and middle-income countries, road accident fatalities and injuries are estimated to be the fourth main cause of healthy life years lost by the total population in 2030 (Aghayan *et al.*, 2012; Bliss and Breen, 2012; Khademi and Choupani, 2018; Kunt *et al.*, 2011; Tazik *et al.*, 2017). In Turkey, as an upper-middle-income country in the Middle East, the number of fatalities was reported to be 3685 people in 2015; the percentage of fatalities has declined over the past few years (WHO, 2015). Considering the crash statistics, it can be understood that traffic crashes have detrimental effects on the public health of society and impose economic costs on individuals. Identifying where the crashes occur and understanding whether the events are concentrated in certain locations, called hotspots, play an important role in reducing crashes by preventing the most influential factors in crashes. Thus, different techniques including K-means, nearest neighbour hierarchical (Nnh) and kernel density estimation (KDE) are employed to analyse the clustering of crashes in order to identify the locations of crashes throughout hotspot analysis. The geolocation of crashes can be a fundamental step for spatial analysis (Shafabakhsh *et al.*, 2017).

One valid method is density-based spatial clustering of applications with noise (DBSCAN), because it accounts for the density of spatial data. Moreover, the grid is also considered to be an acceptable method because it enhances the accuracy and speed of the computation of mass data, such as crash data. In order to apply a grid into clustering algorithms, it is

recommended to divide the data space into a finite number of cells where clustering is implanted based on the grid cells. Some studies have proposed grid-based algorithms such as CLIQUE (clustering in quest), STING (statistical information grid approach) and grid-clustering (Agrawal *et al.*, 1998; Schikuta, 1993; Wang *et al.*, 1997). In addition, the formation principles of density-based clustering algorithms including DBSCAN, DBCLASD (distribution-based clustering of large spatial databases) and DENCLUE (density-based clustering) are built on dense regions separated by low-density regions. The automatic detection of clusters based on arbitrary shapes has been a widely known algorithm to exploit local neighbourhood queries (Ester *et al.*, 1996; Hinneburg and Keim, 1998; Xu *et al.*, 1998). Analysing the most influential factors contributing to fatality in traffic crashes using XGBoost and grid-based analysis, Ma *et al.* (2019) found that drink-driving, the number of parties involved, rear-end crash, lighting condition, pedestrian involvement, motorcycle involvement, the day of the week and time of the day are the most influential factors. Agrawal *et al.* (2018) also used the DBSCAN algorithm to find the clusters of crash areas and reported that if the user is notified when entering crash areas, crash rates decline.

Overall, identifying contributory factors to road crashes and the location of their occurrence play a pivotal role in the reduction of crash injuries and fatalities. This issue has become a major problem for safety researchers and transport engineers – to seek methods of identifying accurate crash factors, in order to minimise their consequences, and the locations where crashes occur. This problem motivated the researchers of recent studies to use data mining methods and geographical analysis to investigate crash hotspots and accident-prone segments. However, studies related to determining

contributory factors have not been complete owing to the complexity of the environmental, human and vehicle factors contributing to road crashes. Therefore, in order to facilitate the solution of these problems and find road crash factors with high accuracy in comparison with other methods, in the present study crash points recorded for Gebze and İzmit (in Turkey) were clustered by three specific methods, namely, Nnh, KDE and GridBSCAN. The main aim of this study was to determine the crash hotspots using the GridBSCAN algorithm, a new algorithm to improve the performance of DBSCAN by considering grid partitioning, yielding, merging and high performance. Data mining methods, especially clustering and classification techniques, were utilised for reducing the heterogeneity of crash data and discovering the ambiguous patterns. After extracting the results from different clustering methods, various analyses were demonstrated in a geographic information system (GIS) regarding special geographical areas or events, including the crashes that occur in these areas, followed by calculation of the related clustered areas. Then, the clustered area corresponding to each method was calculated as well. Further, the predictive accuracy index (PAI) and hit rate (HR) were computed to compare the above-mentioned methods in terms of hotspot mapping techniques. Finally, GridBSCAN clusters were analysed by applying an analysis of variance (Anova) test to find the parameters leading to crashes resulting in injury.

2. Background

Erdogan *et al.* (2008) worked on KDE and repeatability analysis to determine the hotspots located in the highways. KDE along with K-means clustering was used to specify the road crash hotspots (Anderson, 2009). Steenberghen *et al.* (2004) applied GIS and KDE to evaluate the spatial patterns of injury-related road crashes in London, UK. Boroujerdian *et al.* (2010) developed a new model for prioritising road segments with high crash incidence based on the causes of accidents and found that high proneness to accidents is better determined by accident frequency from a particular cause and specific severity within a segment. Keskin *et al.* (2011) investigated the clusters of a crash using KDE, K-means clustering and Nnh clustering methods within the boundary of the Middle East Technical University campus considering the seasons, days and time periods. The results revealed that crashes mostly occurred between 12 a.m. and 7 p.m. and the highest frequency of crashes occurred on Mondays. Moreira *et al.* (2012) studied three different methods to specify hazardous road locations in the city of Vila Real, in Portugal, utilising different techniques such as the Nnh clustering algorithm, KDE and point density. Based on the results, high speed was considered as the main factor involved in single-vehicle crashes. Dai (2012) also used a spatiotemporal clustering technique to determine clusters of injured pedestrians in GIS and indicated that suburban high-activity corridors considerably increased injury risks in crashes compared to the other areas. Kaygisiz (2012) examined the correlation of land use and the occurrence of traffic crashes in

Eskişehir, Turkey using prediction models through the application of land use and traffic crash data.

Considering different types of vehicles involved in crashes within urban areas in Osmaniye, Turkey using K-function analysis, KDE and nearest neighbour distances (Nnd) and the chi-square test, Yalcin and Duzgun (2015) found that the black spots and violation patterns emerged with a high percentage in all crashes related to two-wheeled vehicles. Data mining techniques including K-nearest neighbours and K-means algorithms have been utilised to examine crash data and driving violation patterns in order to determine contributory accident factors (Kumar and Toshniwal, 2016; Mauro *et al.*, 2013; Shirmohammadi *et al.*, 2019; Thapa and Lee, 2016). Kaur and Kaur (2017) predicted the location of black spots as being on state highways and ordinary district roads using K-nearest neighbours. The results indicated that straight roads and intersections have the most influential effects on the risk of crashes.

Dereli and Erdogan (2017) also proposed a descriptive model to specify the black spots of 300 000 traffic crashes in 2408 different state roads in Turkey during 2005–2013, applying model-based spatial statistical methods. The results identified the Bayesian method as the most accurate among comparison methods. Zhang *et al.* (2018) used a fuzzy expert system to identify hazardous road sections and determine factors contributing to them. A clustering algorithm called DBSCAN, relying on a density-based concept of clusters, was introduced, leading to the identification of the arbitrary shape of the clusters (Ester *et al.*, 1996). The findings demonstrated that DBSCAN can be considered as a more effective method in finding clusters with an arbitrary shape when compared to the well-known algorithm CLARANS (clustering large applications based on randomized search). Ye *et al.* (2010) developed a hotspot system for analysing traffic crashes based on the GIS system, DBSCAN algorithm and thematic map analysis. The results indicated that crashes could be identified based on the designed framework. Szénási and Csiba (2014) utilised the DBSCAN method to identify black spots. The high level of performance observed in DBSCAN relies heavily on the choice of input. Likewise, DBSCAN mainly depends on the maximum radius (ϵ) and a minimum number of points (minPts) required to form the clusters. Doğru and Subaşı (2015) compared clustering techniques, including DBSCAN and hierarchical clustering of the crash points, and reported that DBSCAN and agglomerative hierarchical clustering algorithms showed a higher level of performance compared to expectation maximisation and K-means. When considering the clustering of vehicles in traffic, density- and connectivity-based algorithms showed a superior performance compared to centroid- or distribution-based algorithms.

Alotaibi (2018) also examined density-based clustering to analyse road crash data utilising different methods, such as data-mining algorithms, as well as DBSCAN and a parallel

frequent mining algorithm. On the other hand, Almjewail *et al.* (2018) investigated traffic crash records in Riyadh with the purpose of identifying high-risk crash locations in relation to K-means and DBSCAN methods; they found that the K-means and DBSCAN could be considered the two clustering techniques that were used to determine the most frequently occurring crash locations.

The review of the literature demonstrated that no study applied grid and density-based spatial clustering methods to detect crash hotspots in relation to traffic crash factors. In this regard, the novelty of the present study lies first in considering the mixture of grid and density-based spatial clustering methods, namely the GridDBSCAN algorithm, and clustering methods including Nnh, and KDE to identify crash spatial patterns. Second, the Anova test is used to find the parameters leading to crashes causing injury. Then, Nnh and KDE are used to validate the findings on clusters. Furthermore, two accuracy methods, namely, PAI and HR, are taken into consideration to determine the best clustering method among the Nnh, KDE and GridDBSCAN methods.

3. Case study

The current study separately investigated two cities in Kocaeli province, namely, İzmit and Gebze, which are located on the state highways of Turkey near the Marmara Sea. According to the official population census of Turkey (TÜİK, 2018), the population of both cities has increased: it rose from 313 964 to 363 416 and from 297 029 to 371 000 during the period from 2009 to 2018, respectively.

First, the city of Gebze was examined; it is situated 65 km (30 miles) south-east of Istanbul on the Gulf of İzmit in the eastern arm of the Marmara Sea. In terms of population, Gebze is considered to be the second largest district in Kocaeli province after İzmit. Gebze is located in the western part of Kocaeli province. Subsequently, the city of İzmit was evaluated; this is located in the eastern part of Kocaeli province on the main route of the D100 highway that connects east to west, alongside the Marmara Sea. The geographic location of the studied area is shown in Figure 1. According to the data collected by Kocaeli Metropolitan Municipality (2015) in coordination with the Turkish General Directorate of Security, the highest numbers of crashes within Kocaeli province, 1502 and 1463 out of 6689 crashes, respectively, occurred in Gebze and İzmit during 2013–2014 (Figure 2).

4. Data description

In total, 6689 crashes were recorded for Kocaeli province during 2013–2014. Considering the statistics, 5689 and 1000 crashes occurred in urban and rural areas, respectively. Thus, the data for each city were separated for clustering. In this study, 647 and 611 recorded urban crashes for Gebze and İzmit were used to apply and analyse the clustering methods. The Kocaeli province is displayed on the map in Figure 3.

Figures 3(a) and 3(b) depict the distribution of fatality crash points and the injury crash points, respectively. Only the data for injury crashes (98%) were considered because there was an insignificant number of fatal crashes reported. In addition, considering the high frequency of urban crashes (79% of raw data), only the urban data were used in this study. Figure 4 illustrates the distribution of crash points in Kocaeli province, in which the blue and red points indicate urban and rural crash areas, respectively.

According to the existing data, in total, 17 descriptive and two numerical parameters, as reported by Kocaeli Metropolitan Municipality (2015), were utilised in the present study. Furthermore, the numerical parameters were the speed limit and the number of vehicles involved in each injury crash. Moreover, the descriptive data, along with the crash frequency, are provided in Table 1. As shown, the crash data comprised various parameters, such as the physical conditions of the road (e.g. the existence of guardrails, the light conditions of the road, etc.), the geometrical conditions (e.g. grade or level roads, curved sections or straight road sections, etc.) and some crash characteristics, including collision type, and crash type.

Different types of crashes recorded in the existing data are depicted in Figure 5, among which ‘right-angle’ collisions are considered to be the most common type of crash in these cities because they mostly occurred at intersections. The time of the crash was taken into account in this study. According to the reports, most injury crashes in Gebze and İzmit occurred during the evening.

5. Method

Several analysis methods were used in different software programs to examine the density and clustering of the crashes, including Nnh clustering, KDE method and GridDBSCAN clustering in CrimeStat, version 4.02 (Levine, 2010), ArcGIS 10.5 and Elki release (version 0.7.5) software, respectively.

The Nnh clustering technique is an unsupervised method, which leads to a hierarchical structure in the clusters. Based on this method, the main difference is the minimum distance between the items inside each cluster. In addition, this method is used for recognising the crash points that are concentrated in a dense area. In Nnh clustering, the Euclidean distance between each pair of points is checked and utilised as a criterion for clustering as well. Thus, if a threshold distance (d) is selected, pairs with smaller distances are clustered together. If desired, a second criterion can be determined as the minimum number of points ($n_{\min}$) to be in a cluster. Then, data points that meet both criteria (i.e. distance and $n_{\min}$) can be labelled as a ‘cluster’. After calculating the first-order clusters, the second- and high-order clusters are similarly created until only one cluster is left, or the threshold criterion represents a failure. For this reason, Nnh clustering creates no cluster from all observations in the study area. (Kundakci, 2014).

Figure 1. Satellite images of the area studied and the crash points

Figure 2. Crash frequency in Kocaeli cities during 2013–2014

In other words, the distance selection process, through which the range of clustering analysis is defined, is highly important in the Nnh method. The users of CrimeStat should determine the search radius and the minimum number of points. The distance of 1000 m is considered as the starting point for detecting crash points, which is also applicable for the analysis method used for examining the crash points. First, the minimum number of points was considered as 20 ($n_{\min} = 20$) in order to take a large number of injury crashes into consideration. Then, $n_{\min} = 10$ was considered and the search distance was reduced to 500 and 200 m by observing the big clusters and not the accurate results. Finally, the best values were the search distance 200 m and $n_{\min} = 10$ in relation to the criteria of Nnh clustering. Different predefined distance thresholds can

Figure 3. Crash points in the entire data of (a) fatal crashes and (b) injury crashes

Figure 4. Study area and crash points in the urban and rural areas of Kocaeli province during 2013–2014

be investigated to obtain more robust results. Moreover, the criteria can be changed according to the scope of the analysis. The threshold distance and $n_{\min}$ could be changed because each city had its network characteristics.

On the other hand, using a smaller threshold distance may be appropriate for recognising specific crash locations. A criterion (the definition of $n_{\min}$) is also subjective and related to data features. Thus, the researcher should increase the number and test several numbers to ensure that the identified cluster represents a meaningful number of points. Like threshold distance determination, specifying the $n_{\min}$ value depends on the experience and the optimum value can be obtained after a trial period. There are two types of output for the Nnh method, namely, convex hull and ellipse.

The KDE is regarded as a non-parametric method, which is employed to estimate the probability density of a random variable. By evaluating the spatial pattern, this method estimates the crash risk at a spatial unit given the crash counts at the neighbouring spatial units. Accordingly, a symmetric surface is placed on the centre point of a spatial unit and the distances between the centre point, followed by calculation of the locations of the crashes within the surface (Fotheringham *et al.*, 2001). To compute the KDE map, the search radius (bandwidth size) and the cell size were also obtained by calculating the detected crash points applying the Nnh clustering method. Therefore, the total area of convex hulls, along with the maximum distance of the crash points in each cluster were taken into account. In this study, the search radius was also adopted based on the results of Nnh clustering.

DBSCAN is a well-known algorithm for density-based clustering. It is also effective; thus it can detect the arbitrary shaped clusters in dense regions. Mahran and Mahar (2008) introduced a new algorithm called GridDBSCAN to improve the performance of DBSCAN by considering grid partitioning, yielding, merging and high performance, with the advantage of a high degree of similarity. This proved to run much faster than the original DBSCAN. In addition, the clusters were determined by splitting a similarity graph of the data into connected components. Likewise, the density function was approximated on a sparse grid so as to make the method feasible in higher-dimensional settings and scalable with respect to the number of data points. In this algorithm, a grid was created, partitioning the surrounding space of the spatial data; therefore, data could be transferred to the cells. In addition, DBSCAN was separately applied on each partition and then the clusters of all partitions were integrated to obtain the data clustering.

Table 1. The descriptive data of urban injury crashes in Gebze and İzmit during 2013–2014

No.	Descriptive parameters	Subcategories	Gebze	İzmit
			Relative frequency: % (injury crash frequency)	Relative frequency: % (injury crash frequency)
1	Road type	Two-way divided	48.7 (315)	55.1 (337)
		One-way	6.5 (42)	9.3 (57)
		Two-way undivided	44.8 (290)	35.6 (217)
2	Road class	State highway	7.5 (49)	69.6 (425)
		Street	92.5 (598)	30.4 (186)
3	Guardrail	No	83.1 (538)	71.7 (438)
		Yes	16.9 (109)	28.3 (183)
4	Shoulder	No	85.5 (553)	75.1 (452)
		Yes	14.5 (94)	24.9 (159)
5	Road marking	No	33.4 (216)	25 (153)
		Yes	66.6 (431)	75 (458)
6	Traffic signs	No	55.73 (329)	48.5 (296)
		Yes	44.26 (318)	51.5 (315)
7	Lighted sign	No	92.5 (599)	85 (525)
		Yes	7.5 (48)	15 (86)
8	Junction type	Four-leg intersection	25.6 (166)	18.2 (111)
		No junction	35.1 (227)	57.8 (353)
		Roundabout	8.3 (54)	3 (18)
		T-shape intersection	27.2 (176)	15.8 (97)
		Y-shape intersection	4.88 (24)	5.2 (32)
9	Collision point	On the median	2.9 (19)	6.2 (38)
		On the road	87.9 (569)	81.8 (500)
		On the shoulder	1.6 (10)	2.4 (15)
		On the road side	4.3 (28)	5.5 (33)
		On pedestrian	3.3 (21)	4.1 (25)
10	Crash time	Day	0.748 (315)	55.3 (338)
		Night	51.3 (332)	44.7 (273)
11	Weather condition	Clear	86.2 (558)	83.7 (511)
		Rainy/snowy	0.813 (89)	16.3 (100)
12	Road surface	Dry	81.6 (528)	78.5 (480)
		Wet	18.4 (119)	21.5 (131)
13	Horizontal alignment	Curve	14.7 (99)	16.2 (99)
		Straight	85.3 (552)	83.8 (512)
14	Vertical alignment	Grade	35.5 (230)	25.3 (155)
		Level	64.5 (317)	74.7 (456)
15	Light condition	No	43.4 (281)	20.6 (126)
		Yes	56.6 (366)	79.4 (485)
16	Collision type	Head on	12.8 (83)	9.8 (60)
		Hitting an object	10.7 (69)	13.1 (80)
		Rear end	11.4 (74)	19.9 (120)
		Right angle	55.6 (360)	42.4 (259)
		Rollover	5.3 (34)	9.5 (58)
		Run-off-road collision	4.2 (27)	5.5 (34)
17	Crash type	With multiple vehicles	82.7 (535)	76 (646)
		With a vehicle	17.3 (112)	24 (147)

Various results are obtained in terms of size, shape and location by using different techniques for detecting hazardous crash points. Finally, other clustering methods, namely, Nnh and KDE were also applied and their results were compared with those of the GridBSCAN. However, no accurate method was determined in order to evaluate the accuracy of the most critical factors contributing to crash hotspots. Thus, it should be mentioned that the percentage points in the hotspots (crash points) in the total points of the area under study is known as HR. Higher HR implies that the hotspot

technique is more accurate, whereas a larger hotspot area represents the higher likelihood for a higher number of crashes in the future in that hotspot. Thus, HR fails to consider the area of the hotspot, which could make the results less meaningful to law enforcement agencies. The PAI was first introduced by Chainey *et al.* (2008) for addressing the probable problem resulting from the HR. The PAI takes into account the sizes of the hotspots and the study area. Furthermore, the PAI is applied in crime hotspot mapping in order to compare different capturing algorithms and the prediction ability of

Figure 5. Collision types in the urban area for Gebze and Izmit during 2013–2014

the crash points. This index is determined by considering the ratio of the points in the hotspots over the total points of the area under study to the detected crash points over the whole study area.

After clustering, the Anova test was used to determine whether the clustering results were statistically significant, followed by examining the effective criteria which were involved in clustering. In other words, the crashes were analysed to detect the crash points with high density and to determine their characteristics. When a variable has a nominal and ordinal scale, whereas another variable has a distance or ordinal scale, an index should be selected to predict one variable from the other. The correlation coefficient index is one of these indices, which is represented by the symbol η^2 .

6. Results and discussion

In the current study, different types of clustering methods were visualised in a GIS environment to detect and examine hotspots. Then, the results (related to Gebze and Izmit) of three clustering methods – Nnh, KDE and GridBSCAN – were discussed in detail. The best methods for hotspot mapping were determined by means of PAI and HR values. Finally, the significant parameters leading to injury crashes in both cities were derived from the Anova tests. A thematic map, displaying the severity levels of the crash points, was also used to illustrate the results of the clustering method. As shown in Figure 6, large clusters covering more than just a road or intersection were constructed when the Nnh clustering method was applied with the given criteria. It is more appropriate to detect smaller areas in order to recognise properly the correlation between the crashes and the components of the urban areas. Two types of output (i.e. convex hull and ellipse are clear) are depicted in Figure 6. The only difference between these outputs lies in the enclosing of the crash points.

Figure 6. Severity map of injury crashes with Nnh method ($n_{\min} = 20$ and $d = 1000$ m)

Given that the aim of the present study was mainly to recognise the high-density points of injury crashes considering the urban area, understanding the search distance (d) of the crash points is of great significance. As illustrated in Figure 6, the resulting clusters of the crash points with $d = 1000$ m and $n_{\min} = 20$ included more than a road or intersection. As a result, the distance reduced to 500 m while the detected clusters covered a large area for this case. This process was repeated to discover a proper distance threshold value for clustering. Eventually, the most appropriate distance for demonstrating the crash points of the urban areas was assumed as 200 m, which generated results that were much easier to interpret. Considering the constant $n_{\min}$ criterion, smaller distances ($d = 100$ and 50 m) can be taken into consideration to detect the crash points near intersections. The crash points at an intersection were properly demonstrated, considering the intended d and $n_{\min}$ values (Figure 7).

Figure 8 depicts the clusters which were extracted from the Nnh method with $n_{\min} = 10$ and $d = 200$ m in Izmit and Gebze. Similarly to the above case, the search radius in KDE was also considered as 200 m. Looking at the maps in Figure 9, the purple areas with a relative density of 80–100% were recognised as hazardous crash points. Considering the urban areas, most of the crashes occurred in the state road D100 in Izmit, Istanbul on both highway mainline and ramps (Figure 9(a)).

The values of ϵ , 'minPts' were set like the parameters of Nnh (d and $n_{\min}$) and KDE (search radius) in order to compare the findings with GridBSCAN results. Moreover, as mentioned earlier, the 'Gridwidth' is considered twice as much as that of ϵ . The input parameters of the GridBSCAN algorithm were selected as $\epsilon = 200$ m, minPts = 10, and Gridwidth = 400 m. The GridBSCAN algorithm clustered the areas with a higher level of crash density but recognised the scattered points as the outliers or the noise points. As shown in Figure 10,

Figure 7. One of the intersections of Izmit with $n_{min}=10$ and $d=100$ m

Figure 8. Nnh clustering for (a) Izmit and (b) Gebze with $n_{min}=10$ and $d=200$ m

the GridBSCAN algorithm was used for clustering the data points of Izmit and Gebze.

After examining the clusters, it was found that the most hazardous crash points were located near or at the intersections or the entrances/exits of the highways. Few crash points were located in suburban roads or outside the city centre. In Gebze, with narrow urban streets, rugged topography and encompassing areas with great elevation differences, the hazardous crash points were located at the junction of the local streets and suburban roads. These cities demonstrated different traffic behaviours considering special topography, as well as the

urban context of Gebze and the wide streets of Izmit. Based on the data in Table 2, the GridBSCAN clustering method had a lower runtime compared to the conventional DBSCAN algorithm. The main difference between conventional DBSCAN and GridBSCAN algorithms lies in the implementation speed of the algorithm. The GridBSCAN algorithm was used because it has a higher implementation speed compared to a conventional DBSCAN algorithm.

The HR and PAI were utilised to determine the best clustering method in this study. Tables 3 and 4 represent the average PAI values for three hotspot mapping techniques. These results

Figure 9. Kernel density map with 200 m search radius in (a) Izmit and (b) Gebze

Figure 10. GridBSCAN clustering for (a) Izmit and (b) Gebze ($\epsilon = 200$ m, $\text{minPts} = 10$, and $\text{Gridwidth} = 400$ m)

Table 2. The parameter values used in GridBSCAN and DBSCAN algorithms

Parameters		Gebze	Izmit
Input	ϵ : m	200	200
	minPts	10	10
	Gridwidth: m	400	400
Output	Number of clusters	12	13
	GridBSCAN runtime: ms	1	2
	DBSCAN runtime: ms	16	18

were generated to show whether the differences in the prediction abilities of the hotspot mapping techniques were consistent. The results revealed that HR is the best technique for the GridBSCAN clustering method in terms of recognising the

patterns of the crash points. Regarding the area ratio of the clusters, the Nnh method showed the highest value for the PAI (36.92) in Izmit.

The GridBSCAN clusters for Gebze and Izmit were evaluated and the significant parameters of the GridBSCAN clustering method were obtained using the Anova test. In other words, the Anova test was applied to understand the correlation between the parameters affecting the crashes and the number of injuries in the GridBSCAN method.

Cluster CL9, with the most number of crashes (41), is located in the south-west of Gebze. As indicated in Table 5, the junction type was the main cause of crashes in this cluster and nearly 75% of the crashes happened at a roundabout called

Table 3. The HR and PAI values of different clustering methods for Gebze

	GriDBSCAN	KDE	Nnh ellipse	Nnh convex hull
All points	647	647	647	647
Outliers	417	430	499	499
Points in cluster	230	217	148	148
All area: m ²	46 519 394	46 519 394	46 519 394	46 519 394
Cluster area: m ²	578 692	946 546	325 737	322 211
HR	35.55	33.54	22.87	22.87
PAI	28.82	16.48	32.67	33.02
Number of clusters	12	20	9	9

Note: HR, hit rate; PAI, predictive accuracy index; KDE, kernel density estimation; Nnh, nearest neighbour hierarchical

Table 4. The HR and PAI values of different clustering methods for Izmit

	GriDBSCAN	KDE	Nnh ellipse	Nnh convex hull
All points	611	611	611	611
Outliers	397	401	435	435
Points in cluster	214	210	176	176
All area: m ²	43 560 604	43 560 604	43 560 604	43 560 604
Cluster area: m ²	493 186	727 980	342 688	319 929
HR	35.02	34.80	27.12	27.12
PAI	30.93	20.82	34.47	36.92
Number of clusters	13	23	13	13

Note: HR, hit rate; PAI, predictive accuracy index; KDE, kernel density estimation; Nnh, nearest neighbour hierarchical

Table 5. The Anova results for the clusters of Gebze using the GriDBSCAN clustering technique

Cluster	No. of injury crashes in each cluster	Significant parameters	Sum of squares	Degrees of freedom	Mean square	F	Sig.	η	η^2
CL0	20	Vertical alignment	65.766	1	65.766	6.773	0.018	0.523	0.273
CL1	21	—	—	—	—	—	—	—	—
CL2	14	Collision type	3.381	4	0.845	4.149	0.036	0.805	0.648
		Road class	1.525	1	1.525	4.962	0.046	0.541	0.293
		Road type	3.648	2	1.824	12.805	0.001	0.836	0.7
CL3	14	Traffic signs	1.578	1	1.578	5.207	0.042	0.55	0.303
CL4	21	Road class	5.486	1	5.486	6.204	0.022	0.496	0.246
		Guardrail	5.486	1	5.486	6.204	0.022	0.496	0.246
CL5	10	—	—	—	—	—	—	—	—
CL6	16	Crash time	8.215	1	8.215	5.687	0.032	0.537	0.289
		Junction type	17.854	4	4.464	4.639	0.019	0.792	0.628
		Weather condition	78.019	1	78.019	39.443	0	0.867	0.752
CL7	15	Surface condition	47.426	1	47.426	10.949	0.006	0.676	0.457
		—	—	—	—	—	—	—	—
CL8	22	—	—	—	—	—	—	—	—
CL9	41	Junction type	19.462	4	3.564	6.854	0.005	0.250	0.625
CL10	10	Speed limit	23.733	2	11.867	5.664	0.034	0.786	0.618
CL11	26	Junction type	3.202	2	1.601	6.228	0.007	0.593	0.351

Istasyon Square. Table 5 also shows that CL11 constituted 26 crashes, and the junction type was known as a significant parameter that led to injury crashes. After a thorough investigation of this cluster in GIS, it was found that the crashes occurred at the intersections of residential streets (four-leg and three-leg junctions).

Izmit is divided into two sections: the urban area and the main axis of the Istanbul–Izmit highway. According to Table 6, CL1

and CL9 involved the highest number of injury crashes (29). Both of these clusters are located in the main axis of the D100 Istanbul–Izmit highway. For the CL1 cluster, the crash time was a significant parameter that led to injury crashes. In CL1, 70% of the crashes occurred in a day in which rear-end crashes were the common type of collision. As regards CL9, the crash type was recognised as a major contributing parameter to injury crashes and multiple vehicle crashes played an important role in the injury crashes, accounting for 80% in this cluster. In

Table 6. The Anova results for the clusters of Izmit using the GridBSCAN clustering technique

Cluster	No. of injury crashes in each cluster	Significant parameters	Sum of squares	Degrees of freedom	Mean square	F	Sig.	η	η^2
CL0	19	Speed limit	19.798	2	9.899	7.603	0.005	0.698	0.487
CL1	29	Crash time	13.435	1	13.435	4.36	0.046	0.373	0.139
CL2	12	Shoulder	7.202	1	7.202	9.336	0.012	0.695	0.483
CL3	25	—	—	—	—	—	—	—	—
CL4	11	Vertical alignment	2.327	1	2.327	8.727	0.016	0.702	0.492
CL5	10	—	—	—	—	—	—	—	—
CL6	12	—	—	—	—	—	—	—	—
CL7	16	—	—	—	—	—	—	—	—
CL8	11	Shoulder	3.157	1	3.157	20.455	0.001	0.833	0.694
		Collision type	2.962	2	1.481	7.483	0.015	0.807	0.652
CL9	29	Crash type	25.369	3	8.456	5.793	0.004	0.64	0.41
CL10	15	—	—	—	—	—	—	—	—
CL11	13	Road marking	11.856	1	11.856	21.78	0	0.791	0.626
CL12	10	—	—	—	—	—	—	—	—

addition, the speed limit was ranked as the next causative crash factor for CL0, CL9, and CL1, respectively.

Moreover, regarding the various traffic crash factors in Table 1 with high accuracy in comparison with Nnh and KDE methods, comparing the GridBSCAN method with other studies it was found that CL1 and CL9 contributed to crashes with 70 and 80% involvement rates, respectively. However, Agrawal *et al.* (2018) focused only on the DBSCAN method, based on the limited road crash factors and crash severity (critical, and non-critical), which simply determined the most major factors of road crashes without considering the percentage of factors involved in crashes or the spatial patterns based on GIS, and without using sensitivity parameters such as PAI and HR. Thus, the GridBSCAN method had a better performance in comparison with the other methods. Additionally, other works focused just on prioritising the contributory factors using regression techniques, without considering grid analysis and without the use of spatial analysis (Bedard *et al.*, 2002; Desai and Patel, 2011; Kim *et al.*, 1995; Murthy and Srinivasa, 2015; Xi *et al.*, 2014). However, this study not only takes into consideration grid analysis by way of clustering methods in relation to the GIS system for finding spatial patterns, but also uses the Anova test in order to characterise contributory factors with the highest number of injury crashes. Therefore, the GridBSCAN method found clusters contributing to road traffic crashes with a high accuracy in comparison with the Nnh and KDE methods. The results of this study are likely to be helpful for safety researchers and organisations in recognising the most influential contributory factors in road crashes.

7. Conclusions

Crash frequency and severity are influenced by several factors, and the identification of these factors may contribute to a reduction in traffic crashes. Considering the number of vehicle crashes and the existence of multiple factors, clustering the

crashes with GridBSCAN can be an effective approach to obtaining these influential factors. The present study sought to recognise the pattern of crash points using the GridBSCAN algorithm. In order to validate the results, KDE and Nnh methods were applied to the crash data. Crash points recorded for Gebze and Izmit (in Turkey) were clustered by these methods. Based on the results obtained, the GridBSCAN algorithm yielded the best accuracy given the HR in comparison with the other methods. Moreover, the GridBSCAN clustering algorithm showed a lower runtime compared to the conventional DBSCAN algorithm. After interpreting the clusters of both cities, the factors affecting the crashes in Gebze and Izmit (e.g. physical and geometrical conditions) were detected by Anova test. In addition, the crash points in Izmit were mainly observed in the main axis of the D100 Izmit–Istanbul highway, and the crash type, speed limit and time of crash were recognised as the major contributing parameters to injury crashes. Finally, with its urban narrow streets with high densities and diverse topography, the injury crashes occurred at most intersections and roundabouts in Gebze. Overall, the GridBSCAN algorithm is suggested for identifying the crash hotspots. It is expected that the procedure developed can be applied to any city to identify the specific set of factors causing traffic crashes.

Acknowledgements

We thank our colleagues from Kocaeli Metropolitan Municipality and the Kocaeli Provincial Police Department in Turkey for collecting and providing the traffic accident data.

REFERENCES

- Aghayan I, Noii N and Kunt MM (2012) Extended traffic crash modelling through precision and response time using fuzzy clustering algorithms compared with multi-layer perceptron. *Promet-Traffic and Transportation* **24**(6): 455–467.
- Agrawal R, Gehrke J, Gunopulos D and Raghavan P (1998) Automatic subspace clustering of high dimensional data for data mining

- applications. *ACM SIGMOD Record* **27(2)**: 94–105, <https://doi.org/10.1145/276304.276314>.
- Agrawal K, Ruth VM, Nandini Y and 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.
- Almjewail A, Almjewail A, Alsenaydi S, ALSudairy H and Al-Turaiki I (2018) Analysis of traffic accident in Riyadh using clustering algorithms. In *Proceedings of the 5th International Symposium on Data Mining Applications* (Alenezi M and Qureshi B (eds)). Springer, Cham, Switzerland, pp. 12–25.
- Alotaibi AS (2018) Density-based clustering for road accident data analysis. *International Journal of Advanced and Applied Sciences* **5(8)**: 113–121.
- Anderson TK (2009) Kernel density estimation and K-means clustering to profile road accident hotspots. *Accident Analysis & Prevention* **41(3)**: 359–364.
- Bedard M, Guyatt GH, Stones MJ and Hirdes JP (2002) The independent contribution of driver, crash, and vehicle characteristics to driver fatalities. *Accident Analysis & Prevention* **34(6)**: 717–727.
- Bliss T and Breen J (2012) Meeting the management challenges of the decade of action for road safety. *IATSS Research* **35(2)**: 48–55.
- Boroujerdian AM, Saffarzadeh M and Abolhasannejad V (2010) Developing a model for prioritising high crash road segments. *Proceedings of the Institution of Civil Engineers – Transport* **163(1)**: 19–29, <https://doi.org/10.1680/tran.2010.163.1.19>.
- Chainey S, Tompson L and Uhlig S (2008) The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal* **21(1–2)**: 4–28.
- Dai D (2012) Identifying clusters and risk factors of injuries in pedestrian–vehicle crashes in a GIS environment. *Journal of Transport Geography* **24**: 206–214.
- Dereli MA and Erdogan S (2017) A new model for determining the traffic accident black spots using GIS-aided spatial statistical methods. *Transportation Research Part A: Policy and Practice* **103**: 106–117, <https://doi.org/10.1016/j.tra.2017.05.031>.
- Desai MM and Patel AK (2011) Road accidents study based on regression model: a case study of Ahmedabad City. *National Conference on Recent Trends in Engineering & Technology, Anand, India*.
- Doğru N and Subaşı A (2015) Comparison of clustering techniques for traffic accident detection. *Turkish Journal of Electrical Engineering and Computer Sciences* **23(1)**: 2124–2137.
- Erdogan S, Yilmaz I, Baybura T and Gullu M (2008) Geographical information systems aided traffic accident analysis system case study: city of Afyonkarahisar. *Accident Analysis & Prevention* **40(1)**: 174–181.
- Ester M, Kriegel HP, Sander J and Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD'96: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining* (Simoudis E, Han J and Fayyad U (eds)). AAAI Press, Palo Alto, CA, USA, vol. 96, no. 34, pp. 226–231.
- Fotheringham AS, Brunson C and Charlton M (2001) Quantitative geography: perspectives on spatial data analysis. *Geographical Analysis* **33(4)**: 370–372.
- Hinneburg A and Keim DA (1998) An efficient approach to clustering in large multimedia databases with noise. *Proceeding 4th International Conference on Knowledge Discovery and Data Mining, New York, NY, USA*, pp. 58–65. See http://www.tuik.gov.tr/PreIstatistikTablo.do?istab_id=2305 (accessed 27/05/2019).
- Kaur G and Kaur H (2017) Black spot and accidental attributes identification on state highways and ordinary district roads using data mining techniques. *International Journal of Advanced Research in Computer Science* **8(5)**: 2312–2317.
- Keskin F, Yenilmez F, Çolak M, Yavuzer I and Düzgün HS (2011) Analysis of traffic incidents in METU campus. *Procedia – Social and Behavioral Sciences* **19**: 61–70.
- Kaygisiz Ö (2012) *Kentsel Arazi Kullanımı Trafik Kazas İlişkisi: Eskişehir Örneği*. PhD thesis, Gazi University, Ankara, Turkey (in Turkish).
- Khademi N and Choupani AA (2018) Investigating the road safety management capacity: toward a lead agency reform. *IATSS Research* **42(3)**: 48–55.
- Kim K, Nitz L, Richardson J and Li L (1995) Personal and behavioral predictors of automobile crash and injury severity. *Accident Analysis & Prevention* **27(4)**: 469–481.
- Kocaeli Metropolitan Municipality (2015) *Traffic Accident Data for Years 2013–14, Kocaeli, Turkey* (in Turkish). See <https://www.kocaeli.bel.tr/en/> (accessed 26/01/2021).
- Kumar S and Toshniwal D (2016) A data mining approach to characterize road accident locations. *Journal of Modern Transportation* **24(1)**: 62–72.
- Kundakci E (2014) *Identification of Urban Traffic Accident Hot Spots and their Characteristics by Using GIS*. Master's thesis, Orta Doğu Teknik Üniversitesi, Ankara, Turkey.
- Kunt MM, Aghayan I and Noii N (2011) Prediction for traffic accident severity: comparing the artificial neural network, genetic algorithm, combined genetic algorithm and pattern search methods. *Transport* **26(4)**: 353–366.
- Levine N (2010) *CrimeStat III: A Spatial Statistics Program for the Analysis of Crime Incident Locations (Version 3.3)*. Ned Levine & Associates, National Institute of TX, Houston, TX, USA.
- Ma J, Ding Y, Cheng JC et al. (2019) Analyzing the leading causes of traffic fatalities using XGBoost and grid-based analysis: a city management perspective. *IEEE Access* **7**: 148059–148072.
- Mahran S and Mahar K (2008) Using grid for accelerating density-based clustering. In *2008 8th IEEE International Conference on Computer and Information Technology*. IEEE, Piscataway, NJ, USA, pp. 35–40.
- Mauro R, De Luca M and Dell'Acqua G (2013) Using a k-means clustering algorithm to examine patterns of vehicle crashes in before-after analysis. *Modern Applied Science* **7(10)**: 11–19.
- Moreira I, Gonçalves A and Shrubbsall S (2012) Planning on safer urban mobility through spatial analysis: application to Vila Real, Portugal. In *AESOP 26th Annual Congress*. METU, Ankara, Turkey.
- Murthy NSSR and Srinivasa RR (2015) Development of model for road accidents based on intersection parameters using regression models. *International Journal of Scientific and Research Publications* **5(1)**: 1–8.
- Schikuta E (1993) *Grid-clustering: A Fast Hierarchical Clustering Method for Very Large Data Sets*. Center for Research on Parallel Computation, Rice University, Houston, TX, USA, CRPC-TR93358.
- Shafabakhsh GA, Famili A and Bahadori MS (2017) GIS-based spatial analysis of urban traffic accidents: case study in Mashhad, Iran. *Journal of Traffic and Transportation Engineering (English Edition)* **4(3)**: 290–299.
- Shirmohammadi H, Hadadi F and Saeedian M (2019) Clustering analysis of drivers based on behavioral characteristics regarding road safety. *International Journal of Civil Engineering* **17(8)**: 1327–1340.
- Steenberghen T, Dufays T, Thomas I and 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.
- Szénási S and Csiba P (2014) Clustering algorithm in order to find accident black spots identified by GPS coordinates. *Proceedings of the 14th International Multidisciplinary Scientific GeoConference SGEM 2014, Albena, Bulgaria*, pp. 497–504.

- Tazik M, Aghayan I and Sadeghi M (2017) Application of multiple logistic regression, Bayesian logistic and classification tree to identify the significant factors influencing crash severity. *Journal of Engineering Science and Technology* **12**(11): 3044–3056.
- Thapa S and Lee J (2016) *Data Mining Techniques on Traffic Violations*. University of Bridgeport, Bridgeport, CT, USA.
- TÜİK (2018) *Population of Provinces by Years, 2000–2019*. TÜİK, Ankara, Turkey. See https://tuikweb.tuik.gov.tr/PreIstatistikTablo.do?istab_id=1590 (accessed 26/01/2021).
- Wang W, Yang J and Muntz RR (1997) STING: a statistical information grid approach to spatial data mining. *Proceeding of the 23rd International Conference on Very Large Data Bases, Athens, Greece*, pp. 186–195.
- WHO (World Health Organization) (2015) *Global Status Report on Road Safety 2015*. World Health Organization, Geneva, Switzerland.
- Xi JF, Liu HZ, Cheng W, Zhao ZH and Ding TQ (2014) The model of severity prediction of traffic crash on the curve. *Mathematical Problems in Engineering* **2014**: article ID 832723, <https://doi.org/10.1155/2014/832723>.
- Xu X, Ester M, Kriegel HP and Sander J (1998) A distribution-based clustering algorithm for mining in large spatial databases. *Proceedings 14th International Conference on Data Engineering, Orlando, FL, USA*, pp. 324–331.
- Yalcin G and Duzgun HS (2015) Spatial analysis of two-wheeled vehicles traffic crashes: Osmaniye in Turkey. *KSCE Journal of Civil Engineering* **19**(7): 2225–2232, <https://doi.org/10.1007/s12205-015-0661-0>.
- Ye J, Zhou Y, Li M and Wang C (2010) Research and implement of traffic accident analysis system based on accident black spot. In *2010 5th International Conference on Computer Science & Education*. IEEE, Piscataway, NJ, USA, pp. 1805–1809.
- Zhang C, Ma X, Yan X et al. (2018) Identifying hazardous road sections using a fuzzy expert system. *Proceedings of the Institution of Civil Engineers – Transport* **173**(3): 147–158, <https://doi.org/10.1680/jtran.16.00119>.

How can you contribute?

To discuss this paper, please email up to 500 words to the editor at journals@ice.org.uk. Your contribution will be forwarded to the author(s) for a reply and, if considered appropriate by the editorial board, it will be published as discussion in a future issue of the journal.

Proceedings journals rely entirely on contributions from the civil engineering profession (and allied disciplines). Information about how to submit your paper online is available at www.icevirtuallibrary.com/page/authors, where you will also find detailed author guidelines.