Traffic & Transportation

Identifying and Analysing Traffic Accident Hotspots – A Holistic Approach Combining Spatial and Data Mining Techniques

Omer Faruk CANSIZ[1], Mehmet Fatih CAN[2], Kevser UNSALAN[3]

[1] ofaruk.cansiz@iste.edu.tr, Department of Civil Engineering, Iskenderun Technical University, Iskenderun, Turkey
[2] mfatih.can@iste.edu.tr, Department of Water Resources Management, Iskenderun Technical University, Iskenderun, Turkey
[3] Corresponding author, kevser.keskin@iste.edu.tr, Department of Civil Engineering, Iskenderun Technical University, Iskenderun, Turkey

Original Scientific Paper
Submitted: 31 Dec 2024
Accepted: 17 June 2025
Published: 30 Mar 2026

Publisher:
Faculty of Transport and Traffic Sciences, University of Zagreb

ABSTRACT

This study presents a holistic approach to traffic accident analysis by integrating hierarchical clustering, variogram modelling and association rule mining. The analysis identified four critical accident-prone zones: dense residential areas, roads near city centres, multi-curved roads, and dispersed residential and agricultural areas. The Gaussian variogram model revealed significant spatial dependencies, indicating that accidents are concentrated in specific hotspots rather than evenly distributed. Association rule mining revealed key factors contributing to accidents, including dry road surfaces, fair weather conditions and the absence of public transportation vehicles. Additionally, road geometry, particularly overlapping horizontal and vertical curves, significantly contributed to accident frequency in multi-curved regions. The study's findings align with existing literature, offering deeper insights through a unified framework. Recommendations include the implementation of advanced traffic monitoring systems, improvements in road infrastructure and targeted driver education, contributing to more effective traffic safety strategies and accident prevention.

KEYWORDS

traffic accident analysis; hotspot detection; clustering; variogram modelling; association rule mining.

1. INTRODUCTION

It is estimated that approximately 1.2 million fatalities and 20 million injuries occur annually due to traffic accidents [1]. According to the World Health Organisation (WHO), the number of injuries resulting from traffic accidents is projected to increase by 50% by 2030, compared to 2004 levels [2]. The growth in urban population and vehicle ownership is observed to be directly proportional to the rise in traffic accidents. Although technological advancements have been developed to prevent accidents and reduce injuries, the local characteristics of traffic accidents necessitate region-specific analyses. Consequently, traffic accident case studies have maintained their relevance over time. In existing literature, statistical methods, machine learning techniques and GIS-based approaches are commonly applied, with most studies emphasising regression, classification and spatial analysis methodologies.

Within the regression analysis category, various approaches have been employed. For instance, Özen (2020) analysed the relationship between traffic volume, intersection geometry and traffic control features in urban traffic accidents using Poisson and negative binomial regression methods [3]. Similarly, Codur et al. (2013) investigated the influence of road geometry, seasonal factors, annual average daily traffic (AADT) and heavy vehicle presence through generalised linear regression [4]. In a complementary study, Wang et al. (2019) examined both temporal and spatial dynamics of traffic accidents using spatial delay models (SLM), spatial error models (SEM) and time-effects error models (T-FEEM) [5]. Additionally, the human factor has been addressed in several studies. For example, Gümüs et al. (2013) utilised hierarchical regression to assess the psychological conditions of drivers [6], while binary logistic regression has been applied to understand injury severity among pedestrians [7].

Beyond regression techniques, classification methods are also widely applied in traffic accident analysis due to the multivariate nature of accident data. Krishnaveni and Hemalatha (2011) compared the performance of five classification algorithms, including the naive Bayes Bayesian classifier, AdaBoostM1 meta classifier, PART rule classifier, J48 decision tree classifier and random forest tree classifier, in predicting injury severity [8]. In another study, clustering techniques were used alongside classification to enhance performance [9]. Additionally, Chang and Wang (2006) applied the classification and regression tree (CART) method to reveal the impact of vehicle type on injury severity [10]. Manga and Murat (2009) further explored spatial accident data using factor analysis, particularly focusing on hotspot areas [11].

In the context of GIS-based studies, significant contributions have been made towards the spatial and temporal analysis of traffic accidents. For example, Özlü et al. (2021) utilised ArcGIS to determine peak accident times and black spot locations [12]. Xie and Yan (2008) refined the kernel density estimation (KDE) method to enhance the identification of black spots in traffic networks [13]. Le et al. (2022) expanded upon this by employing Moran's I statistics to prioritise accident hotspots [14]. Kang et al. (2018) examined spatial-temporal patterns of traffic accidents involving the elderly population in Seoul, recommending time- and location-specific improvements [15]. Moreover, hotspot analysis using the Getis-Ord Gi* statistic has been employed to determine the significance levels of various accident locations [16].

In modern urban environments, traffic accidents continue to represent a critical social challenge, necessitating comprehensive analyses for effective mitigation. Although a variety of analytical methods exist, including data mining, statistical modelling and spatial analysis, each method typically provides partial insights. To address these limitations, this study proposes a holistic approach that combines clustering, variogram analysis and association rule mining. By integrating these methodologies, the study aims to examine accident patterns more comprehensively by not only analysing the spatial and temporal dimensions but also investigating the adverse factors causing accidents. The proposed framework is demonstrated through a case study, showcasing how this combination of methods can yield deeper insights and contribute to the development of effective road safety strategies.

2. STUDY AREA AND METHODOLOGY

2.1 Definition of traffic accident hotspot

To effectively reduce traffic accidents, it is essential that frequently occurring accident locations are accurately identified. Although a universal definition of an accident hotspot is absent in the literature, a location is typically classified as a hotspot when accidents are concentrated in that region ([17]; [18]). The intensity threshold for defining hotspots varies according to the minimum number of accidents per kilometre ([19]; [18]).

The accurate and reliable identification of accident hotspots is considered crucial for the development of policies aimed at the prevention and reduction of traffic accidents. Therefore, the spatial and infrastructural profile of the study area must be systematically analysed to ensure effective policy formulation and implementation.

2.2 Study area

The province of Hatay, located in southern Turkey along the eastern shores of the İskenderun Gulf, has been selected as the study region. It is bordered by the Mediterranean Sea to the west, Syria to the east, Adana to the northwest, Osmaniye to the north and Gaziantep to the northeast. The province encompasses 15 districts, namely: Antakya, Altınözü, Arsuz, Belen, Defne, Dörtyol, Erzin, Hassa, İskenderun, Kırıkhan, Kumlu, Payas, Reyhanlı, Samandağ and Yayladağı. The total land area of Hatay, excluding lakes, is 5,524 km².

Hatay's geopolitical significance is complemented by its industrial and agricultural prominence at both local and global scales. The region ranks second in Turkey's iron and steel production. Agriculturally, Amik Plain, with its high yield, plays a critical role, supporting 25% of Turkey's fresh fruit and vegetable exports. Hatay ranks first in the production of parsley, plum, persimmon and lettuce. Additionally, the province contributes 10.71% to Turkey's olive production and 27.62% to its citrus output [20]. The high levels of road, rail and sea traffic, coupled with dense border gate operations, underscore the complexity of traffic dynamics in the region.

2.3 Data acquisition and pre-processing

The dataset utilised in this study consists of traffic accident reports documenting fatal and injury-related accidents, sourced from the Hatay Serinyol Traffic Branch Office and covering the period 2017–2020. The analysis is limited to main road accidents occurring within the districts of Antakya, Belen, Samandağ, Yayladağı, Kırıkhan and Hassa.

Initially, the accident locations were identified using descriptions in the reports, such as intersection names, highway mileage and landmarks like stores, shops and hospitals. These descriptions were geocoded using Google Earth, with X, Y coordinates standardised to the WGS84 format. A total of 1,477 cases from the initial 1,574 records were considered after excluding secondary roads from the dataset. For uniformity, the roads were segmented into 5-kilometre sections to ensure consistent accident frequency accumulation per section.

Following the identification of hotspots, accident reports were further analysed to explore the underlying causes. The factors investigated, along with their sub-indices, include temporal and spatial conditions, climatic factors, road characteristics, collision types and vehicle categories involved. All factors used in association rule mining analysis are comprehensively detailed in *Table 1*, facilitating a systematic exploration of accident patterns within the hotspots.

Table 1 – Factors in the reports kept by the traffic police after the accident

Factor	Sub-indices
Weekday	Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday
Month	January, February, March, April, May, June, July, August, September, October, November, December
Season	Winter, spring, summer, autumn
Road type	Divided road, one-way, two-way
Daylight	Day, night, twilight
Weather	Rainy, fair, fog, sandstorm, snowy
Road surface	Dry, iced, puddle, snowy, wet, other
Horizontal curve	Alignment, hairpin curve, horizontal curve
Vertical curve	Alignment, hilltop, steep hill, vertical curve
Intersection	Bridge intersection, circular, four-way, T shape, Y shape, other, no
Vehicle number	Multiple, single, two
Occurrence form of accident	Animal crash, crashing stationary vehicle, crash obstacle, head-on, multiple crash, object drop from vehicle, overturn skidding, pedestrian crash, pile up, rear-end, run-off-road, side by side, side impact
First collision	Alignment, roadside (outside shoulder), shoulder, sidewalk, intersection, other, not detected
Are there any bicycles or motorcycles involved in the accident?	Yes, no
Is there a vehicle involved in the accident?	Yes, no
Are there any minibuses or buses involved in the accident?	Yes, no
Is there any heavy vehicle involved in the accident?	Yes, no
Is there any construction equipment involved in the accident?	Yes, no

2.4 Methodology

The similarities of accident locations based on accident frequency were identified using the hierarchical clustering method. This approach generated a dendrogram, which visually represents how the dataset is clustered. The unweighted pair-group method with arithmetic mean (UPGMA), in conjunction with singlelinkage and Ward's method, was applied in the cluster analysis. The UPGMA algorithm classifies locations by calculating the average distance between accident spots [21].

Additionally, a GIS-based variogram model integrated with PAST software was employed to determine accident hotspots and assess the degree of spatial dependency in the study area. The variogram modelling process involved several stages to select the model that best represents the spatial variability of the data. Initially, empirical variograms were computed by evaluating the pairwise differences between spatial data points at multiple distances and directions. Subsequently, theoretical models including the spherical, exponential, Gaussian and cubic models (*Equations 1–4*) were fitted to the empirical variograms using least squares fitting and weighted least squares fitting techniques. The key parameters of the selected model, namely the scale, nugget and range, were carefully estimated and adjusted to minimise discrepancies between the empirical and theoretical variograms.

Model selection was guided by both visual inspection and statistical criteria, such as the residual sum of squares. Upon determining the optimal variogram model, it was utilised as a fundamental component in geostatistical interpolation techniques, specifically kriging, to predict unknown values at unsampled locations. This process ensured that spatial dependencies were accurately captured, thereby enhancing the reliability of traffic accident hotspot predictions within the study region.

Spherical: $$\gamma(h) = f(x) = \begin{Bmatrix} \text{nugget} + \text{scale}\left(\frac{3h}{2} - \frac{1}{2}h^3\right), & h < 0 \\ \text{nugget} + \text{scale}, & h \geq 0 \end{Bmatrix} \tag{1}$$

Exponential: $$\gamma(h) = \text{nugget} + \text{scale}(1 - e^{-h}) \tag{2}$$

Gaussian: $$\gamma(h) = \text{nugget} + \text{scale}(1 - e^{-h^2}) \tag{3}$$

Cubic: $$\gamma(h) = f(x) = \begin{Bmatrix} \text{nugget} + \text{scale}(7h^2 - 8.75h^3 + 3.5h^2 - 0.75h^7), & h < 0 \\ \text{nugget} + \text{scale}, & h \geq 0 \end{Bmatrix} \tag{4}$$

In the literature concerning traffic accident hotspots, two-dimensional (2D) analyses are predominantly employed to determine spatial dependence by utilising the geographical coordinates of accident locations. These studies have primarily focused on the application of the kernel density estimation (KDE) method ([24]; [25]). A comparative study conducted by Aziz and Ram (2022) examined the Moran's I, Getis-Ord (Gi), and kernel density estimation (KDE) methods, which have been utilised by researchers over time for the identification of accident hotspots [26]. Additionally, Haybat et al. (2022) performed a temporal and spatial analysis of traffic accidents [27]. In this analysis, ArcGIS software tools such as spot density, total case, Anselin local Moran I, hotspot analysis and visualise space-time cube were employed. The spot density tool was used to calculate the densities of accident locations, with the output presented as raster data derived from vector data inputs. Clusters of accident spots were weighted according to their density through the total case tool. The Anselin local Moran I tool was applied to classify traffic accidents by implementing spatial autocorrelation techniques. The hotspot analysis tool was utilised to identify the cluster types of accidents. Finally, the 2D visualise space-time cube was employed to provide a comprehensive spatial-temporal visualisation of accident patterns.

In this study, the paleontological statistics (PAST) software was employed to examine the relationship between accident frequencies and spatial dependence. To achieve this, the gridding method was utilised, which generates a density map based on the frequency of data aligned with a two-dimensional (2D) coordinate system. The gridding method incorporates four distinct interpolation algorithms: inverse distance weighting, thin-plate spline, multiquadratic and kriging [28]. Among these, the multiquadratic algorithm was selected due to its suitability for land modelling applications [29]. To construct the semivariogram, four theoretical models – spherical, exponential, Gaussian and cubic (*Equations 1–4*) were evaluated. In these models:

- The nugget parameter represents a constant added to account for non-zero variance at zero distance, enabling the surface to deviate from passing precisely through the data points.
- The spacing parameter governs the horizontal extent of the curve along the distance axis.
- The normalised distance value (h) corresponds to the distance-to-range ratio, defining how spatial relationships change over increasing distances.
- The scale parameter influences the vertical extent of the curve along the variance axis, controlling how variance values respond to distance changes.

Among the evaluated models, the Gaussian function was found to produce the least standard error, indicating it is the best fit for the semivariogram analysis. The reduced sum-of-squares of the residuals (RSS) criterion was employed for model selection, ensuring the minimisation of discrepancies between the empirical and theoretical variograms. Additionally, the presence of spatial dependency between the data points was assessed using the spatial continuity ratio (SCR), following established methodologies [30, 31] (*Equation 5*). This evaluation ensured that the selected variogram model accurately captured the spatial structure inherent in the accident frequency data, thereby enhancing the reliability of subsequent geostatistical interpretations.

$$\mathrm{SCR} = \frac{\mathrm{scale}}{\mathrm{nugget+scale}} \tag{5}$$

As the final method within the holistic analytical framework, association rule mining was employed to analyse traffic accident data, with the objective of uncovering hidden patterns and identifying relationships among various factors contributing to accidents. When compared to other machine learning techniques, association rule mining demonstrates greater flexibility by allowing the utilisation of specific functions and dependent variables. This characteristic enhances its capacity to reveal meaningful patterns that may not be easily detected through traditional analytical methods. Furthermore, the insights derived from these association rules can provide policymakers with rapid, data-driven solutions for the prevention of traffic accidents, thereby supporting the development of targeted interventions aimed at improving road safety [32, 33].

In determining interesting rules within the association rule mining process, three key parameters are considered: support, confidence and lift value.

- Support indicates the rate of accidents, reflecting how frequently a particular rule appears within the dataset.
- Confidence represents the probability of an event occurring, given the presence of a related antecedent condition.
- The lift value measures the frequency of co-occurrence between the antecedent and consequent, providing insights into how much more likely these factors occur together than would be expected by chance.

These parameters are formulated using standardised equations, as outlined in the literature [32] (*Equations 6–10*). The combined interpretation of these values enables the identification of significant patterns, offering valuable insights for traffic accident prevention strategies and policy development.

$$\mathrm{Supp(X)}=\|\{t \in D | X \subseteq t\}\| / \|t \in D\| \tag{6}$$

$$\mathrm{Conf}(X \Rightarrow Y) = \mathrm{Supp}(X \cup Y)/\mathrm{Supp}(X) \tag{7}$$

$$\mathrm{lift}(X \Rightarrow Y) = \frac{\mathrm{Supp}(X \cup Y)}{\mathrm{Supp}(X)\mathrm{Supp}(Y)} \tag{8}$$

$$\mathrm{Supp}(X \cup Y) \geq \sigma \tag{9}$$

$$\mathrm{Conf}(X \cup Y) \geq \delta \tag{10}$$

σ and δ represent minimum support and confidence, respectively. In the study, meaningful rules were derived by considering the highest support values, as they balance both the confidence and lift values, while also reflecting common patterns [34].

3. RESULTS AND DISCUSSION

In this study, the data obtained from accident reports are analysed using a holistic approach that integrates variogram, clustering and rule mining techniques. Based on the results of the cluster analysis, four clusters are identified in terms of accident frequency. These clusters are subsequently designated as follows: Group-I: “high”, Group-II: “high-medium”, Group-III: “low-medium” and Group-IV: “low” roads (*Figure 1*).

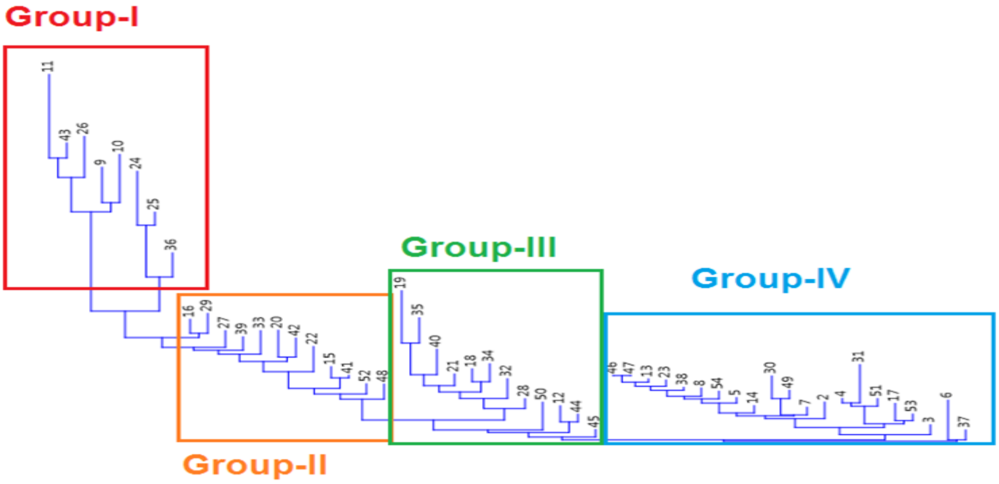


Figure 1 – Grouping of accident zones based on clustering analysis

Note: The numbers presented in the graph represent the road sections, which are divided into segments of 5 km each.

The results of the variogram analysis are presented in *Figure 2*. Based on the colour distribution, it is noteworthy that spots 9, 10, 11, 24, 25, 26, 36 and 43 are identified as hotspots. These hotspots can be categorised into the following regions: “dense residential area along the road”, “road near the city centre”, “multi-curved road” and “dispersed residential and agricultural area along the road.” The descriptions of these regions are provided below.

— **Dense residential area along the road (9-10-11):** This critical section covers a 15 km road segment located between Defne-Samandağ and Antakya. The number of accidents per kilometre in this section is 16.53 (*Figure 3a*).

— **Road near the city centre (24-25-26):** This critical section extends over a total of 15 km, with an accident frequency of 11.27 per kilometre. The region is characterised by the presence of high-density residential buildings and critical infrastructures, including universities, industrial zones, airports and hospitals (*Figure 3b*).

— **Dispersed residential and agricultural area along the road (36):** The total length of this road section is 5 km, with an accident frequency of 10 per kilometre. The settlement pattern in this region is highly dispersed, with agricultural lands predominantly occupying the area (*Figure 3d*).

— **Multi-curved road (43):** This critical section spans 5 km, with an accident frequency of 15.8 per kilometre. The section is notable for containing a total of 15 horizontal curves along the road (*Figure 3c*).

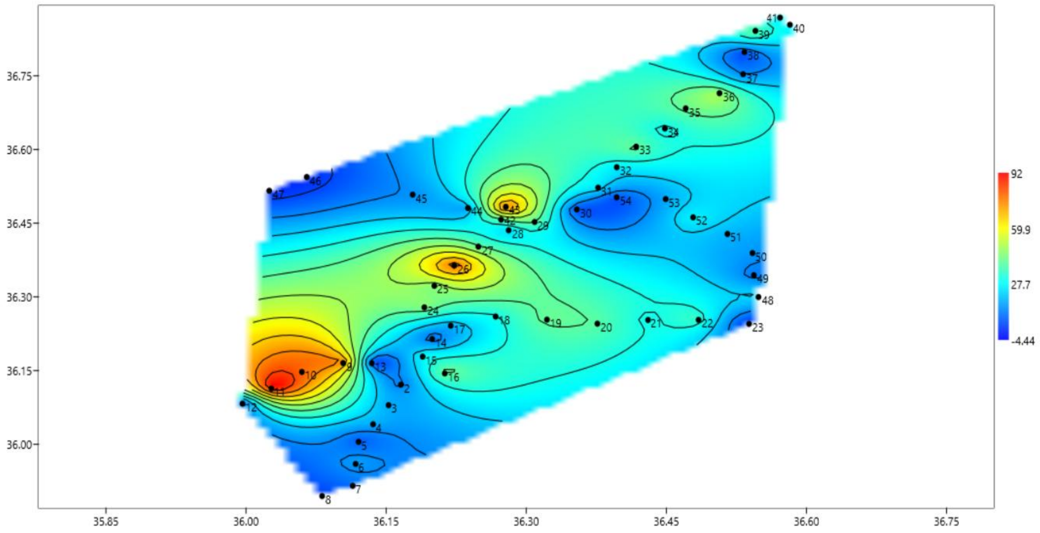
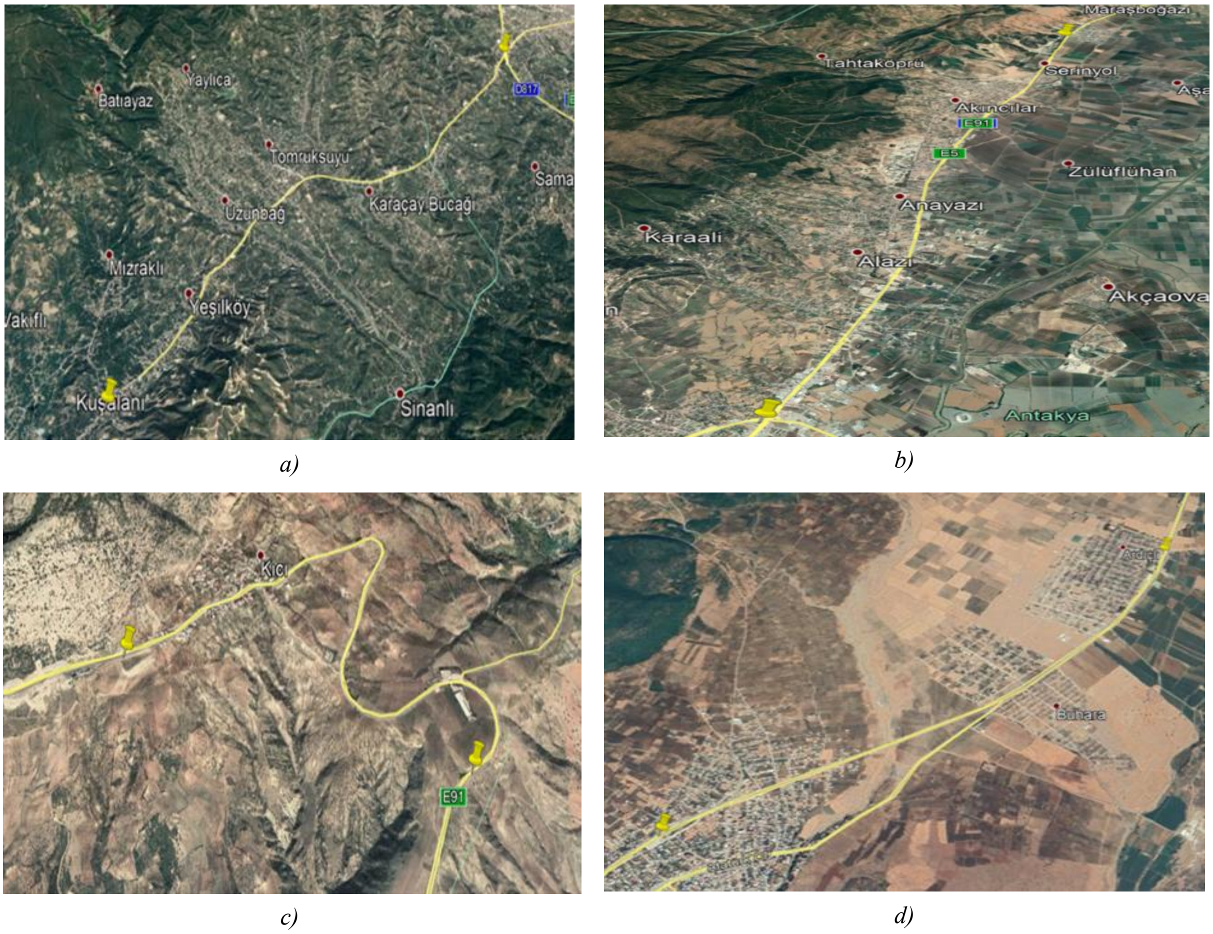


Figure 2 – Spatial distribution of traffic accidents

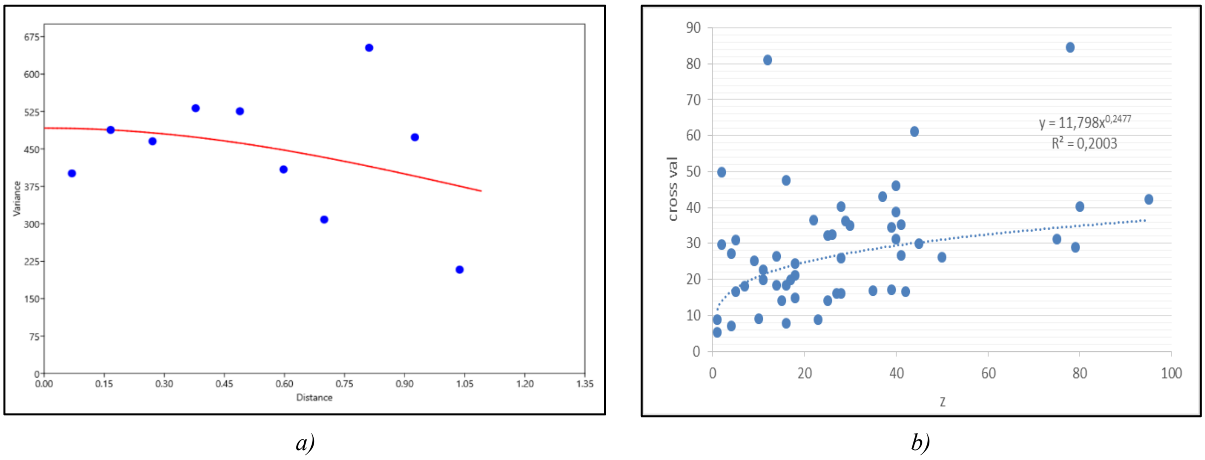
The numbers presented in the graph represent the road sections, which are divided into segments of 5 km each.



a) *b)*

c) *d)*

Figure 3 – Hot locations from Google maps: a) Defne-Samandağ (9-10-11); b) Antakya-Topboğazı (24-25-26); c) Topboğazı-Belen (43), d) Kırıkhan-Hassa (36)



a) *b)*

Figure 4 – a) Theoretical variogram fitted for traffic accident frequency; b) cross validation between actual traffic accident frequencies and variogram model outputs

The Gaussian function yielded the lowest residual sum of squares (RSS) value of 1.23×10^5 in the semivariogram model. The nugget, range and scale values were determined as 491, 7470, 1.6353 and -351.1000, respectively. Based on these values, the SCR value was calculated as -249%. The R-squared value obtained from cross-validation was found to be 0.2003, indicating that spatial dependency was weak and traffic accidents did not occur continuously along the road route. Instead, accidents were concentrated in specific locations, commonly referred to as hotspots (*Figure 4*). Consequently, it was observed that the results of clustering and spatial modelling overlapped.

To generate interesting rules at accident hotspots, the "Rattle" package provided by the R software was utilised. In the Rattle package, the Apriori algorithm was employed to obtain the solutions. The minimum support values selected for the entire region, Defne-Samandağ, Antakya-Topboğazı, Kırıkhan-Hassa and Topboğazı-Belen were 0.6, 0.6, 0.55, 0.65 and 0.25, respectively. Correspondingly, the minimum confidence values chosen for the entire region, Defne-Samandağ, Antakya-Topboğazı, Kırıkhan-Hassa and Topboğazı-Belen were 0.8, 0.8, 0.8, 0.8 and 0.6, respectively. Furthermore, during the interesting rule mining process, a total of 35, 28, 39, 39 and 91 rules were derived for the entire region and the Defne-Samandağ, Antakya-Topboğazı, Kırıkhan-Hassa and Topboğazı-Belen hotspots, respectively. *Table 2* presents five interesting rules identified for each hotspot.

When all hotspots were analysed, it was determined that 79% of the 546 accidents across all regions occurred under fair weather conditions, with dry road surfaces and no involvement of public transport vehicles. Additionally, in 99% of the accidents where the road surface was dry and no public transport vehicles were involved, the weather was reported to be fair. Furthermore, the road surface was found to be dry in 88% of the accidents that occurred in areas without horizontal curves and where public transport vehicles were not involved. Among all accidents, 63% took place under conditions where there were no horizontal or vertical curves, and public transportation vehicles were not present. Overall, the majority of accidents across all regions occurred under conditions characterised by fair weather, dry road surfaces, the absence of horizontal and vertical curves and no involvement of public transportation vehicles.

A total of 248 datasets were analysed for the Defne-Samandağ hotspot. It was found that in 74.6% of the accidents occurring under fair weather conditions in this region, the road surface was dry, no public transportation vehicles were involved, and there were no horizontal curves. Additionally, in 60.9% of the cases where the road surface was dry, the weather was clear, no horizontal curve was present, the first collision occurred at the alignment, and no public transport vehicles were involved. Moreover, in 64.9% of the accidents where the first collision occurred at the alignment, the weather was clear, the road surface was dry, and no horizontal curve was present. According to the accident patterns observed in this critical region, the accidents predominantly occurred under normal daytime and environmental conditions. It was also determined that road geometry had no significant impact on the occurrence of these accidents.

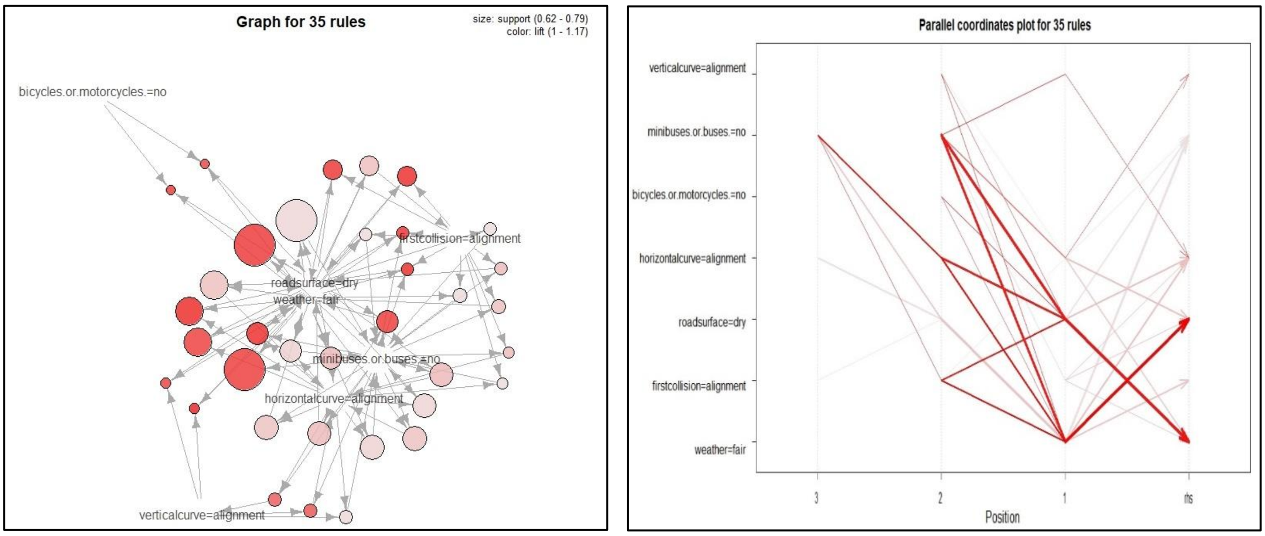
A total of 79 accident datasets were analysed for the critical Antakya-Topboğazı region. It was determined that in 62% of the accidents occurring under fair weather conditions, the road surface was dry, and no public transportation vehicles were involved. Furthermore, passenger vehicles were involved in 93.9% of the accidents where bicycles, motorcycles and heavy vehicles were not present. Additionally, 55.7% of the accidents that did not occur at intersections involved a single vehicle and no public transportation vehicles. In the same proportion (55.7%), accidents occurred on road sections with vertical curves and without the involvement of public transportation vehicles. Similarly, 55.7% of these accidents involved passenger vehicles, with no bicycles, motorcycles or public transportation vehicles present. Moreover, it was observed that 93.6% of the single-vehicle accidents without public transportation vehicle involvement did not occur at intersections.

In the Kırıkhan-Hassa critical region, 50 accident data records were analysed. It was found that in 80% of the accidents where no vertical curve was present, the weather was fair, the road surface was dry, and no public transportation vehicles were involved. Similarly, the same conditions were observed in 76% of the accidents that occurred in areas without horizontal curves. Additionally, 76% of the accidents that took place under fair weather conditions involved dry road surfaces and occurred in sections without horizontal or vertical curves. Furthermore, 80.5% of the accidents that occurred under fair weather conditions, with dry road surfaces and no vertical curves, took place at the first collision point.

For the Topboğazı-Belen critical region, 79 accident data records were subjected to analysis. An examination of the interesting rules for this region revealed that road geometry had a significant influence on the occurrence of accidents. Specifically, 93.8% of the accidents involving a single vehicle on a horizontal curve also occurred on a vertical curve. Additionally, the road surface was found to be dry in 79% of the accidents involving both horizontal and vertical curves and passenger cars. This indicates that weather conditions did not contribute to a reduction in the road friction coefficient during the occurrence of these accidents. Moreover, in 60.9% of the accidents involving a single vehicle on a vertical curve, the vehicle ran off the road. It was also observed that 74% of the accidents involving passenger vehicles and dry road surfaces occurred on vertical curves during daytime hours.

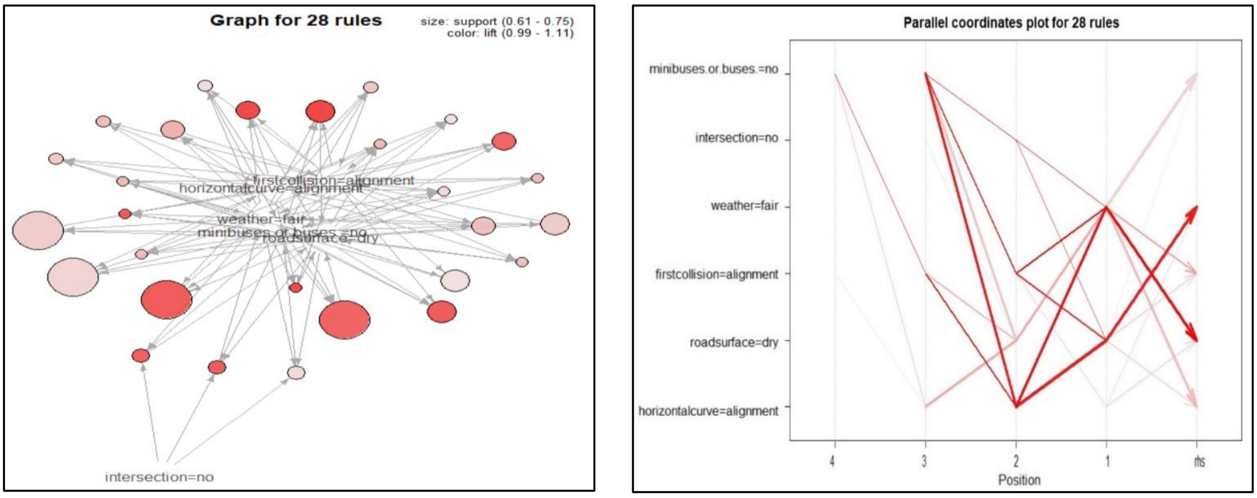
Table 2 – Interesting rules with high lift

Zone	Rules	LHS	RHS	Support	Confidence	Lift
ALL REGION	1	road surface=dry, minibuses or buses=no	weather= fair	0.791	0.993	1.156
	2	horizontal curve=alignment, minibuses or buses=no	road surface=dry	0.691	0.881	1.043
	3	road surface=dry, horizontal curve=alignment, minibuses or buses=no	weather=fair	0.685	0.992	1.155
	4	weather=fair, horizontal curve=alignment, minibuses or buses=no	road surface=dry	0.685	0.984	1.166
	5	horizontal curve=alignment, minibuses or buses=no	vertical curve=alignment	0.634	0.808	1.129
DEFNE-SAMANDAĞ	1	road surface=dry, horizontal curve=alignment, minibuses or buses=no	weather=fair	0.746	1.000	1.102
	2	weather=fair, first collision=alignment, minibuses or buses=no	road surface=dry	0.669	0.988	1.114
	3	weather=fair, road surface=dry, horizontal curve=alignment	first collision=alignment	0.649	0.826	1.045
	4	road surface=dry, intersection=no, minibuses or buses =no	weather=fair	0.629	1.000	1.102
	5	weather=fair, horizontal curve=alignment, first collision=alignment, minibuses or buses=no	road surface=dry	0.609	0.987	1.113
ANTAKYA-TOPBOĞAZI	1	road surface=dry, minibuses or buses=no	weather=fair	0.620	0.980	1.358
	2	bicycles or motorcycles=no, heavy vehicle=no	vehicle=yes	0.582	0.939	1.348
	3	vehicle number=single, minibuses or buses =no	intersection=no	0.557	0.936	1.088
	4	vertical curve=vertical curved, minibuses or buses=no	intersection=no	0.557	0.863	1.002
	5	bicycles or motorcycles=no, vehicle=yes, minibuses or buses=no	intersection=no	0.557	0.917	1.065
KIRIKHAN-HASSA	1	weather=fair, road surface=dry, minibuses or buses=no	vertical curve=alignment	0.800	0.976	0.996
	2	weather=fair, road surface=dry, minibuses or buses=no	horizontal curve=alignment	0.760	0.927	0.986
	3	road surface=dry, horizontal curve=alignment, vertical curve=alignment	weather=fair	0.760	0.974	1.160
	4	weather=fair, horizontal curve=alignment, vertical curve=alignment, minibuses or buses=no	road surface=dry	0.740	1.000	1.163
	5	weather=fair, road surface=dry, vertical curve=alignment	first collision=alignment	0.660	0.805	1.032
TOPBOĞAZI-BELEN	1	vertical curve=vertical curved, vehicle number=two	first collision=alignment	0.273	0.882	1.471
	2	horizontal curve=horizontal curved, vehicle number=single	vertical curve=vertical curved	0.273	0.938	1.258
	3	horizontal curve=horizontal curved, vertical curve=vertical curved, vehicle=yes	road surface=dry	0.273	0.790	1.113
	4	vertical curve=vertical curved, vehicle number=single	occurrence form=run-off-road	0.255	0.609	2.092
	5	daylight=day, road surface=dry, vehicle=yes	vertical curve=vertical curved	0.255	0.737	0.989



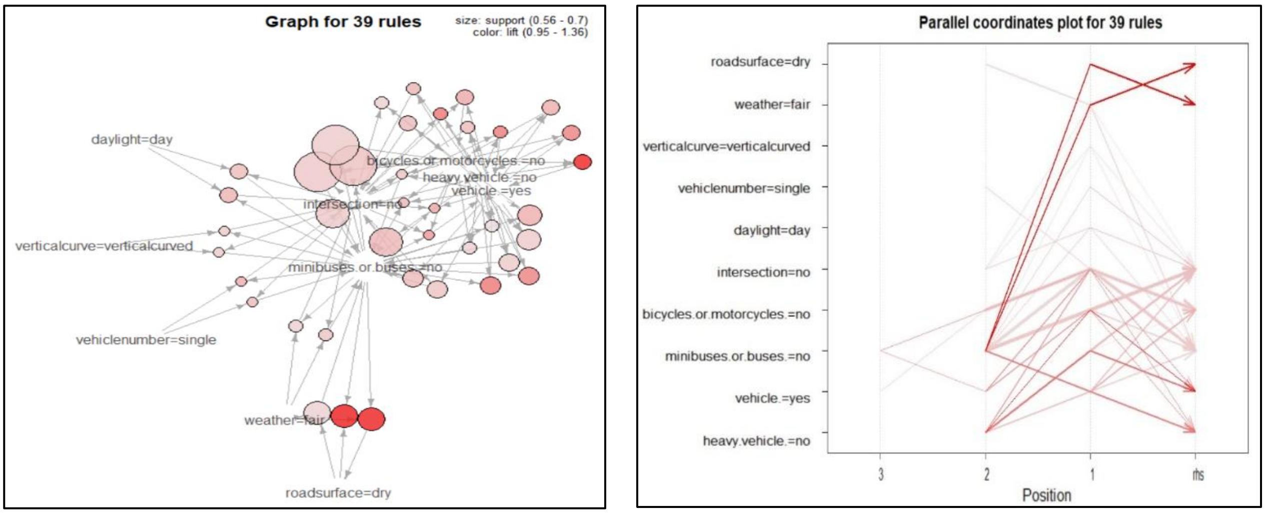
a) *b)*

Figure 5 – Association rule mining graphs for all regions: a) network plots; b) paracord graph



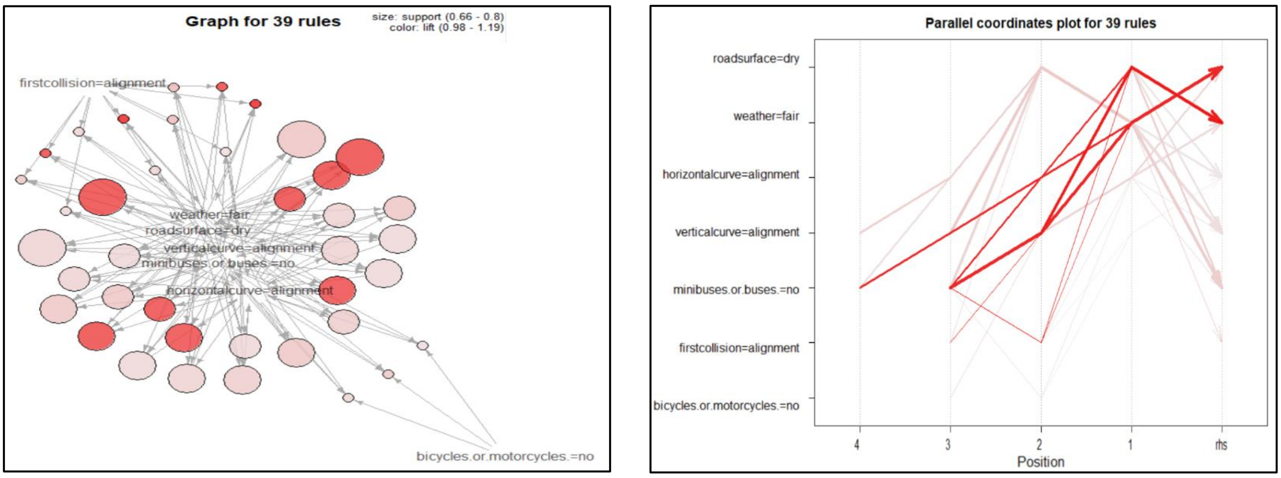
a) *b)*

Figure 6 – Association rule mining graphs Defne-Samandağ: a) network plots; b) paracord graph



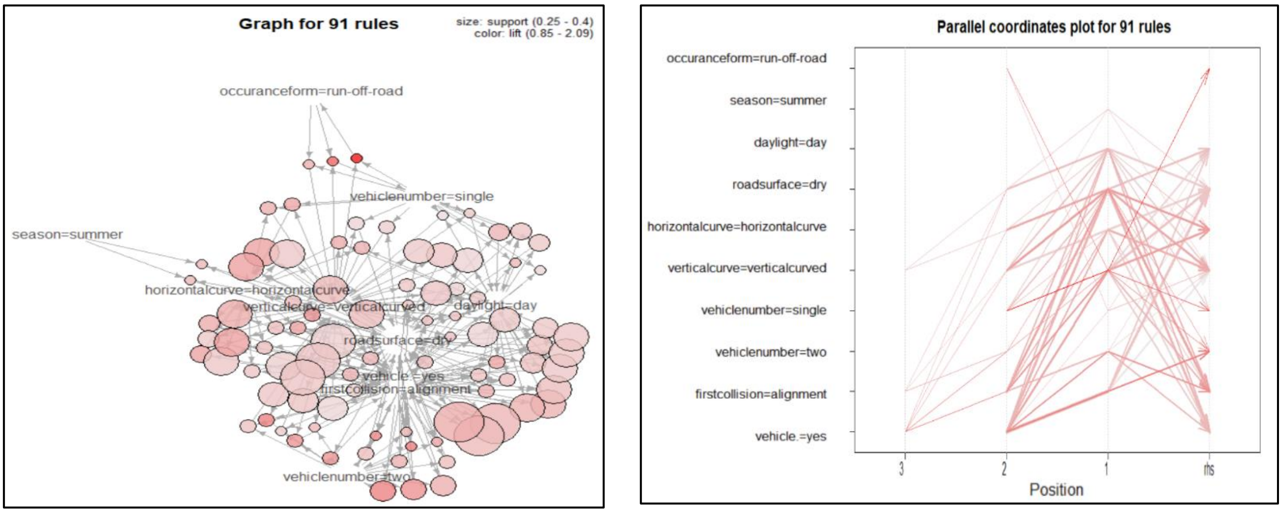
a) *b)*

Figure 7 – Association rule mining graphs Antakya-Topboğazı: a) network plots; b) paracord graph



a) *b)*

Figure 8 – Association rule mining graphs Kırıkhan-Hassa: a) network plots; b) paracord graph



a) *b)*

Figure 9 – Association rule mining graphs Topboğazı-Belen: a) network plots; b) paracord graph

In the final stage of the association rule mining analysis, network visualisations and paracord graphs were generated based on the rules identified for each critical region (*Figures 5–9*). An examination of the network plot and paracord graph encompassing all regions revealed that dry road surfaces and fair-weather conditions play a central role in the occurrence of accidents (*Figure 5*). Moreover, the absence of vertical and horizontal curves, the initial collision occurring at the alignment, and the lack of public transportation and two-wheeled vehicles were identified as key contributing factors to traffic accidents. As demonstrated in the paracord graph, these factors are shown to interact simultaneously in the occurrence of accidents. Consequently, it is expected that improvements made to one or more of these variables would have a corresponding impact on the other factors, thereby contributing to accident prevention. In the critical areas of Defne-Samandağ, Antakya-Topboğazı and Kırıkhan-Hassa (*Figures 6–8*), dry road surfaces and fair-weather conditions were again observed to play a central role. This finding suggests that accidents in these regions predominantly occur under normal daytime conditions. Conversely, in the Topboğazı-Belen region (*Figure 9*), vertical and horizontal curves were identified as the primary factors influencing accident occurrence. Additionally, the dry condition of the road surface, the involvement of two vehicles and the first collision occurring at the alignment emerged as decisive factors contributing to accidents in this critical area.

According to the results of the clustering analysis, the cluster with the highest accident frequencies, designated as Group 1, included the hotspots identified through the variogram analysis. Although the traffic accident regions differ, these four regions were classified within the riskiest cluster due to the high intensity of accidents observed.

Based on the variogram analysis, four hotspots were identified. In terms of accident frequency, the characteristics of these hotspots, ranked from highest to lowest, are as follows: "dense residential area along the road", "road near the city centre", "multi-curved road" and "dispersed residential and agricultural area along the road."

According to the results of interesting rule mining, in the road section categorised as "multi-curved road", both vertical and horizontal curves were found to have a significant influence on accident occurrences. This suggests an increased density of single-vehicle and run-off-road accidents in this region. In contrast, this pattern was not observed in the other critical sections.

Table 3 – An overview of methods used in the literature and our current study to understand accident analysis phenomena

References	Techniques	Findings
[3]	Poisson and negative binomial regression methods	Examined the interaction of traffic volume, intersection geometry and traffic control features in urban traffic accidents.
[4]	generalised linear regression	Heavy vehicles, summer season and annual average daily traffic (AADT) are determined as the most significant variables on the probability of accident.
[5]	linear regression, spatial lag model (SLM), spatial error model (SEM) and time-fixed effects error model (T-FEEM)	They divided the study area into traffic accident zones and revealed the effect of high-density population buildings, such as hospitals and schools, in these regions, on the frequency of accidents.
[6]	hierarchical regression	Anxiety in drivers increases the risk of having an accident, as well as it has been determined that men are more inclined to seek excitement in traffic than women.
[7]	binary logistic regression	Involvement of pedestrians aged 55 and over, involvement of male pedestrians, faulty pedestrians in the accident, speed defect in the accident and the presence of a horizontal curve at the accident point increased the injury severity of pedestrian accidents at a significance level of 0.05.
[8]	naive Bayes Bayesian classifier, AdaBoostM1 meta classifier, PART rule classifier, J48 decision tree classifier and random forest tree classifier	Random forest outperforms the other four algorithms in the classification of injury severity in traffic accidents.
[9]	used classifier techniques; decision tree, lazy classifier and multilayer perceptron classifier	Compared to an unclustered classified dataset, the accuracy level has increased to some extent by using clustering techniques on the dataset.
[10]	classification and regression tree (CART)	The type of vehicle has an effect on the severity of injury
[11]	factor analysis	Factors causing traffic accidents; platform features, environmental factors, driver learning status, pavement and road geometric features, time factor, safe stop and visibility distance.
[12]	GIS (Arc GIS)	By analysing the accident data temporally and spatially, it determined the peak times of the accident as month, day and hour and the locations where the accident peaked as the accident black spot.
[13]	kernel density	Developed the method of kernel density to obtain clearer results in the determination of traffic accident black spots with basic linear units of equal network length.
[14]	kernel density, Moran I	After analysing traffic accidents with the kernel density method, they determined the importance levels of the spots with Moran I statistics in order to determine the priority improvement of traffic accident hotspots.
[15]	hotspot analysis, space-time kernel density estimation (STKDE) and emerging hotspot analysis	How the spatial-temporal characteristics of traffic accidents involving the elderly population in Seoul changed according to the time period, and suggested improvements in the hours and locations where the elderly population is victimised.
[16]	kernel density, hotspot analysis (Getis-Ord Gi *)	After analysing traffic accidents with the kernel density method, they determined the importance levels of the spots with hotspot analysis (Getis-Ord Gi *).
Current study	variogram model (gridding method), hierarchical clustering method and association rule mining	A kind of useful hierarchical approach is suggested to produce specific findings from general to clarify the accident patterns of diverse road characteristics.

When comparing the existing literature with the holistic approach adopted in this study, which integrates clustering, variogram analysis and association rule mining, it becomes evident that previous studies primarily emphasise more limited and basic findings (*Table 3*). While clustering techniques in earlier research have predominantly been utilised for data segmentation [3, 4, 6–11], the integration of variogram analysis in the present approach provides a more comprehensive understanding of spatial dependencies, a dimension that remains largely underexplored in conventional clustering applications [5, 12–16]. Additionally, the incorporation of association rule mining differentiates this study by offering explicit and interpretable associations between variables, an aspect that is often absent in previous research. Although some studies have successfully employed these techniques individually, their combination within a unified analytical framework, as demonstrated in the current study, exhibits a superior capability in uncovering complex patterns and relationships within large and multidimensional datasets.

3.1 Recommendations for the identified accident hotspots based on contributing factors

Based on the factors influencing accidents at the accident hotspots identified in the study, the following recommendations are proposed for the four critical regions.

- **Dense residential area along the road (9-10-11):** Among the four regions, this area has been identified as the most critical location, with the highest frequency of accidents. Considering its geographical and strategic importance, and based on the analysis of significant association rules, it has been observed that the majority of accidents occur during the daytime and under dry road surface conditions. Moreover, as this region experiences a high concentration of side-impact accidents, it is inferred that these accidents may have been caused by vehicles overtaking one another or colliding with vehicles entering from side roads. To mitigate accidents resulting from these activities, it is recommended that additional traffic control measures be implemented. These may include the installation of overtaking restriction barriers, dedicated lanes or warning systems to regulate traffic flow and prevent unsafe overtaking manoeuvres.
- **Road near the city centre (24-25-26):** In this section of the road, accidents predominantly occur during the summer months, daytime hours, and at circular intersections. Notably, side and rear-end collisions are frequent in this area. The region's geographical location, characterised by high-density business and school traffic, suggests that accidents may be attributed to drivers failing to maintain safe following distances, inappropriate overtaking, and disregard for right-of-way rules at intersections. To address these issues, it is recommended that driver behaviour be improved through targeted awareness campaigns and enforcement strategies. The installation of an Electronic Monitoring System (EMS) is strongly advised, as it would enable real-time traffic surveillance and assist in reducing accident rates. Furthermore, the placement of additional traffic signals at critical points is expected to enhance traffic discipline and intersection safety.
- **Dispersed residential and agricultural area along the road (36):** Accidents in this region mostly occur during the day and under dry road surface conditions. The majority of accidents involve passenger cars, with a significant number resulting from vehicles running off the road and side-impact collisions. To prevent such accidents, it is recommended that speed limit signs be strategically placed along this road section. Additionally, regular monitoring and enforcement should be conducted by traffic control officers to ensure compliance. The installation of an EMS for speed monitoring would further contribute to enhanced road safety by discouraging excessive speeding.
- **Multi-curved road (43):** The geometric characteristics of this road section necessitate heightened driver caution, particularly given that most accidents occur at night. This pattern highlights the need for measures aimed at improving nighttime visibility for drivers. Furthermore, although speed limit signs have been placed, there is uncertainty regarding the extent of driver compliance with warning signs, especially given the curved nature of the road. It is recommended that an Electronic Detection System (EDS) be installed in this region to enable real-time traffic control. Such a system would encourage drivers to adopt safer driving behaviours through immediate feedback and enforcement. Additionally, improved road lighting and the use of reflective warning signs would contribute to better night visibility, thereby reducing the likelihood of accidents.

4. CONCLUSION

Traffic accidents can be interpreted as the negative outcomes resulting from the complex interaction of various factors, including driver behaviour, vehicle characteristics, road geometry, environmental conditions and climate. Therefore, the systematic analysis of studies aimed at reducing traffic accidents is of critical importance. In the present study, a systematic approach was undertaken to extract the spatial patterns of traffic accidents and to investigate the underlying causes of accidents across different patterns. The analysis revealed that the association rules derived from the study, alongside geometric features of the road such as horizontal and vertical curves, as well as factors including daytime conditions, working population density and structures surrounding the route, have significant effects on traffic accident occurrences. The findings obtained were found to be consistent with existing literature, further validating the results of this research.

To enhance road safety, it is crucial for policymakers to prioritise the mitigation of accident hotspots. The following measures are recommended:

- Implementation of advanced traffic monitoring systems: The integration of Electronic Detection Systems (EDS) and Electronic Monitoring Systems (EMS) to enable real-time surveillance and immediate enforcement of traffic regulations.
- Improvement of road infrastructure: Redesigning critical road sections with dangerous curves or inadequate signage to reduce accident risks. Enhancing nighttime visibility through improved road lighting and reflective warning signs.
- Strengthening law enforcement strategies: Conducting regular traffic patrols and speed enforcement in critical areas. Applying stricter penalties for violations related to overtaking, following distances and intersection rules.
- Public awareness campaigns and driver education: Promoting safe driving behaviours through targeted awareness programs. Educating drivers on the risks associated with specific road conditions and traffic patterns identified in the study.
- Collaborative, data-driven approach: Ensuring collaboration between local authorities, traffic engineers and the community to implement sustainable safety strategies. Utilising data analytics to continuously monitor accident trends and adjust safety measures accordingly.

As the focus of traffic accident analysis may vary based on regional characteristics, the results of this study are specifically applicable to the examined region. For future research, it is recommended that:

- Other regions with high accident rates should be analysed to investigate region-specific causes of accidents.
- Comparative studies should be conducted across different geographical areas to develop universal safety measures.
- The use of advanced data mining techniques should be expanded to uncover hidden patterns in large-scale accident datasets.

Ultimately, the proactive management of accident black spots will play a pivotal role in enhancing public safety. By adopting a multi-faceted, data-driven approach that involves technological advancements, infrastructure improvements and community engagement, it will be possible to create safer road networks and significantly reduce traffic-related fatalities and injuries.

ACKNOWLEDGEMENTS

The authors would like to express their sincere gratitude to the General Directorate of Security of the Ministry of Internal Affairs of the Republic of Turkey for providing access to the data utilised in this study. The provision of data constitutes the sole external support received; no financial, institutional or academic assistance was granted during the conception, execution or dissemination of the research. All components of the study, including the formulation of research questions, methodological design, data analysis and manuscript preparation, were carried out solely by the authors, independent of any external influence.

REFERENCES

[1] Rajasekaran RB, Rajasekaran S, Vaishya R. The role of social advocacy in reducing road traffic accidents in India. *Journal of Clinical Orthopaedics and Trauma*. 2021;12(1):2-3. DOI: 10.1016/j.jcot.2020.12.021.

[2] Global status report on road safety: time for action. 2009. https://iris.who.int/bitstream/handle/10665/44122/9789241563840_eng.pdf

[3] Özen M, 2020. Dört kollu sinyalize kentsel kavşaklarda trafik kazalarının sıklığını etkileyen faktörlerin incelenmesi. *Teknik Dergi*. 2020;31(3):10033-10053. DOI: 10.18400/tekderg.509128.

[4] Çodur MY, Tortum A. Erzurum-Pasinler road traffic accident prediction model. *Ordu Univ. J. Sci. Tech*. 2013;3(2):39-49. https://dergipark.org.tr/en/download/article-file/113903

[5] Wang W, et al. Factors influencing traffic accident frequencies on urban roads: A spatial panel time-fixed effects error model. *PLoS One*. 2019;14(4):e0214539. DOI: 10.1371/journal.pone.0214539.

[6] Gümüş G, Öztürk İ, Tekeş B. Psikolojik semptomların trafikte heyecan arama ile ilişkisinin incelenmesi. *Trafik ve Ulaşım Araştırmaları Dergisi*. 2020;3(2):109-120. DOI: 10.38002/tuad.773877.

[7] Özen M. Yaya kazalarının yaralanma şiddetinin incelenmesi: Ikili lojistik regresyon modeli uygulaması. *Teknik Dergi*. 2021;32(3):10859-10883. DOI: 10.18400/tekderg.670811.

[8] Krishnaveni, S, Hemalatha M. A perspective analysis of traffic accident using data mining techniques. *International Journal of Computer Applications*. 2011;23(7):40-48. DOI: 10.5120/2896-3788.

Twari P, Dao H, Nguyen GN. Performance evaluation of lazy, decision tree classifier and multilayer perceptron on traffic accident analysis. *Informatica*. 201;741(1): 39-46. https://informatica.si/index.php/informatica/article/view/1595

[9] Chang LY, Wang HW. Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accident Analysis & Prevention*. 2006;38(5):1019-1027. DOI: 10.1016/j.aap.2006.04.009.

[10] Manga AO, Murat YŞ. Trafik kazalarının faktör analizi yöntemiyle incelenmesi. *İzmir Ulaşım Sempozyumu 2009, 8-9 Dec. İzmir, Türkiye.* 2009. p. 1-10.

[11] Özlü T, Haybat H, Zerenoğlu H. Trafik kazalarının zamansal ve mekânsal incelenmesi: Eskişehir örneği. *lnternational Journal of Geography and Geography Education*. 2021;43:136-158. DOI: 10.32003/igge.746447.

[12] Xie Z, Yan J. Kernel density estimation of traffic accidents in a network space. *Computers, Environment and Urban Systems*, 2008;32(5):396-406. DOI: 10.1016/j.compenvurbsys.2008.05.001.

[13] Le KG, Liu P, Lin LT. Traffic accident hotspot identification by integrating kernel density estimation and spatial autocorrelation analysis: a case study. *International Journal of Crashworthiness*. 2022;27(2):543-553. DOI: 10.1080/13588265.2020.1826800.

[14] Kang Y, Cho N, Son S. Spatiotemporal characteristics of elderly population's traffic accidents in Seoul using space-time cube and space-time kernel density estimation. *PLoS One*. 2018;13(5):e0196845. DOI: 10.1371/journal.pone.0196845.

[15] Srikanth L, Srikanth I. A case study on kernel density estimation and hotspot analysis methods in traffic safety management. *International Conference on Communication Systems & Networks (COMSNETS) 2020, 7-11 Jan. Bengalaru, India*. 2020. p. 99-104. DOI: 10.1109/COMSNETS48256.2020.9027448.

[16] Anderson TK. Kernel density estimation and K-means clustering to profile road accident hotspots. *Accident Analysis & Prevention*. 2009;41(3):359-364. DOI: 10.1016/j.aap.2008.12.014.

[17] Karaşahin M, Terzi S. Determination of hazardous locations on Isparta-Antalya-Burdur highways through GIS. *Pamukkale University Engineering College Journal of Engineering Science*. 2003;9(3):305-311. https://dergipark.org.tr/tr/pub/pajes/issue/20531/218703.

[18] Homburger WS, Kell JH. 1981. *Fundamentals of traffic engineering.* California: U.S. Department of Energy Office of Scientific and Technical Information; 1981. https://www.osti.gov/biblio/5744665

[19] Hatay valiliği çevre, şehircilik ve iklim değişikliği il müdürlüğü. *İlimiz hakkında.* https://hatay.csb.gov.tr/ilimiz-hakkinda-i-2645/ [Accessed 15th October 2022].

[20] Øyvind H. *Paleontological Statistics Version 3.20 Reference Manuel*. Oslo: Norway. Natural history museum University of Oslo; 2018.

[21] Krige DG. *A statistical approach to some mine valuations and allied problems at the Witwatersrand*. Master's thesis. University of Witwatersrand; 1951.

[22] Matheron G. Principles of geostatistics. *Economic Geology*. 1963;58:1246-1266. DOI: 10.2113/gsecongeo.58.8.1246.

[23] Benedek J, Ciobanu SM, Man TC. Hotspots and social background of urban traffic crashes: A case study in Cluj-Napoca (Romania). *Accident Analysis & Prevention*. 2016;87:117-126. DOI: 10.1016/j.aap.2015.11.026.

[24] Liu J, et al.. How big data serves for freight safety management at highway-rail grade crossings? A spatial approach fused with path analysis. *Neurocomputing*. 2016;181:38-52. DOI: 10.1016/j.neucom.2015.08.098.

[25] Aziz S, Ram S. A Review of the spatial analysis techniques for the identification of road accident black spots and its application in context to India. *In International road federation world meeting & exhibition 2022, 7-10 Nov. Dubai,* United Arab Emirates. p. 511-524. DOI: 10.1007/978-3-030-79801-7_37.

[26] Haybat H, Zerenoğlu H, Özlü T. Temporal and spatial analysis of traffic accidents: the case of Bursa city. *lnternational Journal of Geography and Geography Education.* 2022;45:404-423. DOI: 10.32003/igge.1016204.

[27] Smith MJ, Goodchild MF, Longley P. *Geospatial analysis: A comprehensive guide to principles, techniques and software tools*. Harborough: Troubador publishing; 2007.

[28] Davis JC, Sampson RJ. *Statistics and data analysis in geology*. New York: Wiley; 1986.

[29] Can MF, Yılmaz AB, Mazlum Y. A Meta-analysis: Geo-statistical approach on the THQ values of Cu, Zn, and Fe accumulation in some fish species from Iskenderun Bay, North-Eastern Mediterranean. *V. In International Congress on Natural and Health Sciences (ICNHS) 2019, 13-15 Dec. Adana*, Türkiye. p. 69-89.

[30] Li Y, Li CK, Tao JJ, Wang LD. Study on spatial distribution of soil heavy metals in Huizhou city based on BP--ANN modeling and GIS. *Procedia Environmental Sciences*. 2011;10:1953-1960. DOI: 10.1016/j.proenv.2011.09.306.

[31] Feng M, Zheng J, Ren J, Xi Y. Association rule mining for road traffic accident analysis: a case study from UK. *In International Conference on Brain Inspired Cognitive Systems (BICS) 2020, 18-20 Dec. Hefei, China.* p. 520-529. DOI: 10.1007/978-3-030-39431-8_50.

[32] Xu R, Luo F. Risk prediction and early warning for air traffic controllers' unsafe acts using association rule mining and random forest. *Safety Science*. 2021;135:105125. DOI: 10.1016/j.ssci.2020.105125.

[33] Gajowniczek K, Ząbkowski T. Data mining techniques for detecting household characteristics based on smart meter data. *Energies*. 2015;8(7):7407-7427. DOI: 10.3390/en8077407.