Unsafe Traffic Events and Crash Occurrences: The Importance of Exploring Their Relationship Using Smartphone App Data

Paraskevi Koliou
Research Associate, Senior Research Engineer
Department of Transportation Planning and Engineering
National Technical University of Athens, Athens, Greece
Email: evi_koliou@mail.ntua.gr

Virginia Petraki
Ph.D. Student, Research Associate
Department of Transportation Planning and Engineering
National Technical University of Athens, Athens, Greece
Email: vpetraki@mail.ntua.gr

Apostolos Ziakopoulos
Research Associate
Department of Transportation Planning and Engineering
National Technical University of Athens, Athens, Greece
Email: apziak@central.ntua.gr

George Yannis
Professor
Department of Transportation Planning and Engineering
National Technical University of Athens, Athens, Greece
Email: geyannis@central.ntua.gr

Word Count: 7,486 words + 0 table (250 words per table) = 7,486 words

Submitted [01/08/2024]

ABSTRACT

This research explores the relationship between unsafe driving events and crash occurrences using a comprehensive dataset collected from various urban junctions. Key traffic metrics such as vehicle flow, average speed, speed differences, and occupancy were analyzed across different junction types to identify high-risk areas. Advanced clustering techniques, including K-Means and DBSCAN, were employed to detect patterns and hotspots of unsafe events. Local spatial analysis using Local Moran's I and Geary's C highlighted significant clusters and spatial outliers, enhancing the spatial analysis framework. A Random Forest Regressor was utilized to determine feature importance, identifying critical predictors of crash occurrences, such as braking behavior, junction complexity, and monitoring duration. Multicollinearity was assessed using Variance Inflation Factor (VIF) scores, ensuring the robustness of the models. Principal Component Analysis (PCA) was also applied for dimensionality reduction, facilitating a more straightforward interpretation of the data. Temporal trends were visualized to understand the variations in traffic metrics over time. The results revealed significant variability in vehicle flow and speed across different junctions, with high-risk areas identified based on speed fluctuations, occupancy rates, and accident frequency. These insights provide a solid foundation for targeted safety interventions and policy-making aimed at improving road safety. The integration of these advanced analytical techniques with detailed traffic data offers a comprehensive approach to understanding and mitigating unsafe driving events and crashes.

Keywords: Traffic Safety, Unsafe Driving Events, Crash Analysis, Clustering Analysis, Spatial Analysis, Feature Importance, Smartphone App Data

INTRODUCTION

The exploration of the relationship between unsafe traffic events and crash occurrences has become a crucial area of study in the quest to enhance road safety (1). With the rapid advancement of technology, particularly the widespread use of smartphones, there is now an unprecedented opportunity to collect and analyze traffic-related data in real-time (2). Smartphone apps can capture a wide range of data, including GPS location, speed, acceleration, braking patterns, and even driver behavior metrics such as phone usage while driving. This wealth of data provides a granular view of driving habits and conditions, offering valuable insights that traditional traffic studies, which often rely on police reports and crash statistics, might miss.

Understanding unsafe traffic events through smartphone data is vital because it allows for the identification of risky behaviors before they result in crashes. For instance, sudden braking, rapid acceleration, and sharp turns are indicators of aggressive driving, which is a known precursor to accidents (3). By analyzing these events, we can develop predictive models that highlight potential danger zones and times, thereby enabling proactive interventions. Additionally, this real-time data can be used to educate drivers on safer driving practices, create targeted enforcement campaigns, and design more effective road safety measures.

Road safety remains a critical concern globally, with traffic accidents causing significant mortality and morbidity annually. Understanding the factors that contribute to crashes and unsafe driving events is essential for developing effective interventions and policies aimed at reducing traffic-related injuries and fatalities. Traditional approaches to studying traffic safety have largely relied on post-accident analyses using police reports, crash statistics, and infrastructure assessments. While these methods provide valuable insights, they are often limited by their retrospective nature and the availability of comprehensive data.

Recent advancements in big data and analytics have further enhanced the ability to predict crashes by identifying patterns in risky driving behavior using telematics data. By employing methodologies like machine learning and spatial analysis, researchers can develop predictive models that highlight danger zones and times, enabling proactive interventions. These techniques also allow for the visualization of hotspots through Geographic Information Systems (GIS) and the analysis of temporal trends, offering a more comprehensive understanding of the factors contributing to road crashes.

The integration of smartphone data into traffic safety research marks a significant shift from traditional, reactive approaches to a more proactive strategy. Continuous monitoring of driving behavior in real-time through smartphone apps enables the detection of risky behaviors, such as sudden braking, rapid acceleration, and distracted driving, which are known precursors to accidents. By addressing these behaviors early, it is possible to implement targeted interventions, create safer driving environments, and design more effective road safety measures. This proactive approach also facilitates the education of drivers on safer practices, helping to reduce the likelihood of accidents.

Despite ongoing efforts to reduce road crashes and fatalities, global statistics have been not decreased. In 2018, road crashes led to 1.35 million deaths annually, translating to approximately 3,700 fatalities per day worldwide (4). In the European Union, there were around 20,653 road fatalities in 2022, marking a 4% increase from 2021, though still a 10% decrease from 2019 (5).

Fatality rates differ significantly across Europe. Sweden and Denmark have the lowest rates, with 22 and 26 deaths per million inhabitants respectively, while Romania and Bulgaria have the highest rates, with 86 and 78 deaths per million inhabitants respectively (*6*). Greece managed to reduce crash fatalities by 51% between 2009 and 2018 but still ranks 22nd among EU states, with 58 deaths per million inhabitants in 2022, slightly higher than in recent years(5, 7). Economic recession has been partially credited for this reduction (8). However, the Hellenic Statistical Authority (9) reported an 18.8% increase in road crashes causing death or injury in January 2018 compared to January 2017 (5).

While some progress has been made, the overall reduction in road fatalities across Europe remains slow, with considerable disparities between countries. The EU aims to halve road deaths by 2030 as part of its Vision Zero strategy, but achieving this goal will require sustained and coordinated efforts across all member states (10).

This study aims to explore the relationship between unsafe driving events and crash occurrences by leveraging smartphone app data. Using a combination of advanced analytical techniques, including clustering methods (11, 12), spatial analysis, and machine learning models, this research seeks to identify key factors that influence crash rates and to detect hotspots of unsafe driving behaviors(13). By integrating the rich data captured by smartphone apps (14) with these advanced methodologies, the study provides valuable insights that can inform targeted interventions, improve road design, and enhance driver education programs. Ultimately, this research contributes to the broader goal of reducing road crashes and improving overall road safety, demonstrating the potential of smartphone app data for real-time monitoring and proactive safety measures.

METHODS

This study employs a multi-faceted approach to analyze the relationship between unsafe driving events and crash occurrences using data collected from a smartphone application. The methodologies employed include clustering analysis, local spatial analysis, feature importance evaluation using machine learning, multicollinearity assessment, and dimensionality reduction. Each method is described in detail below.

Data Collection and Preparation

Data were collected through a smartphone application, encompassing various traffic-related features and unsafe driving event metrics. Key features included the number of left and right exits and entrances, the number of incoming and outgoing lanes, sideway presence, traffic volume, braking behavior metrics, and event speed. Missing values in the dataset were addressed using mean imputation, which replaces missing values with the mean of the respective feature, ensuring a complete and robust dataset. Additional features were derived to enhance the dataset's predictive power, such as calculating the range, mean, and standard deviation of distances and speeds recorded during events.

Modelling driver behavior is a complex phenomenon that has long interested the scientific community. This study aims to investigate the combined influence of road characteristics and traffic on driver behavior, particularly in crash occurrence, using smartphone data on harsh acceleration and braking events in an urban intersection environment. Building on the work of Petraki et al., 2020 (15), the research examines how the road environment and traffic conditions affect driving behavior at intersections, focusing on abrupt accelerations and braking. Conducted at a macroscopic level, the study area includes two major urban expressways in Athens—Mesogeion Avenue and Vouliagmenis Avenue—chosen for their similar traffic lane configurations and separated travel directions (**Figure 1**). These avenues provide a suitable context for analyzing the impact of road and traffic characteristics on driver behavior (15).

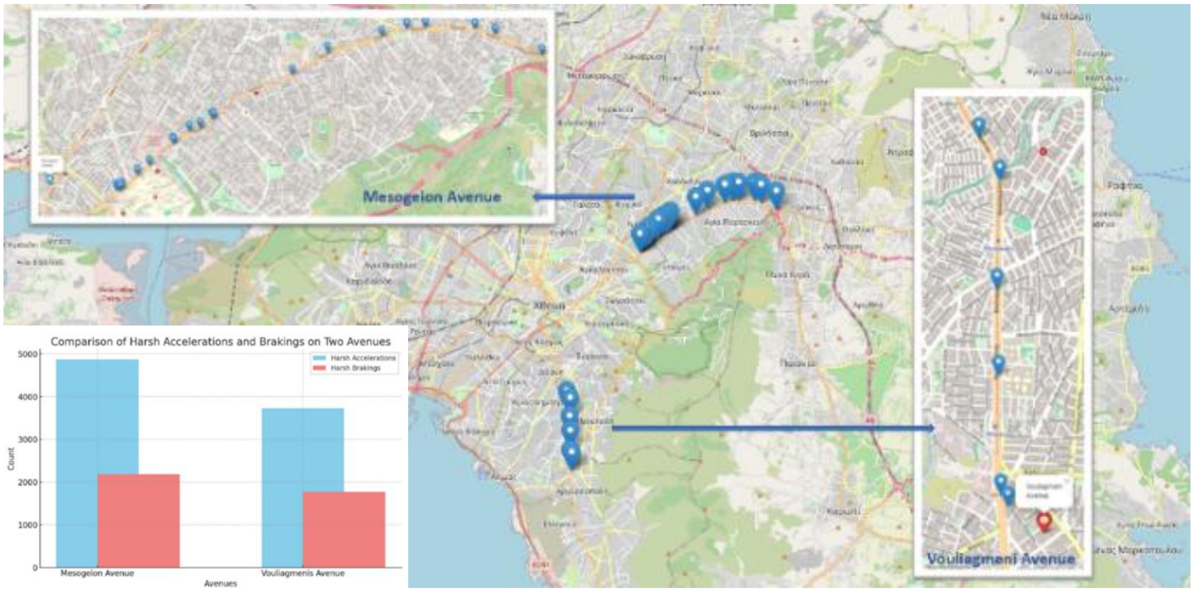


Figure 1: Research Area - Mesogeion and Vouliagmenis Avenues, including harsh acceleration

The data analyzed in this study were sourced from three primary sources. First, driving behavior data were collected from approximately 300 drivers in Athens using the OSeven smartphone application, which records driving behavior. This data captures instances of unsafe traffic events, specifically harsh acceleration (HA) and braking (HB) events. The dataset includes metrics pertinent to traffic safety, such as the identification of junctions where specific events were recorded, traffic volume, average speed, and occupancy rate, providing a comprehensive overview of traffic conditions.

Secondly, traffic metrics were obtained from the Traffic Management Center of the Attica region. These metrics, including traffic volume and average speeds, were collected through 26 loops installed at specific measurement points along the two studied urban expressways. Lastly, road characteristics were extracted using the Google Maps online mapping service, detailing features of road segments and intersections, including lane numbers and configurations. During data collection and processing, challenges were addressed to ensure dataset quality and reliability by standardizing data units and formats in Excel for consistency between sources, and by ensuring accurate spatial alignment in QGIS using precise geolocation data from Google Maps and cross-referencing with known traffic loop locations (16).

Data integration involved merging driving behavior telematics data with traffic metrics from the Traffic Management Center and road characteristics from Google Maps. Spatial mapping using QGIS correlated abrupt driving events with specific road segments and intersections, resulting in a comprehensive database for analyzing harsh acceleration and braking events on the examined avenues .

A total of 303 drivers participated in a naturalistic driving experiment conducted in Athens between August 25, 2016, and November 26, 2017, resulting in the creation of extensive databases of harsh acceleration and deceleration events. Specifically, during this period, 4,869 harsh accelerations and 2,181 harsh braking were recorded on Mesogeion Avenue, while 3,723 harsh accelerations and 1,765 harsh braking were documented on Vouliagmenis Avenue.

Statistical Analysis

In a previous study using the same dataset, an in-depth analysis of the relationship between unsafe traffic events and crash occurrences were conducted, with a particular focus on harsh acceleration and braking events. The data was examined across varying spatial and temporal resolutions, assessing spatial resolution at road intersections, specifically the Junctions of Mesogeion (JK) and Junctions of Vouliagmenis (JV). Temporal resolution was evaluated on a monthly, weekly, and daily basis.

Geographic Information Systems (GIS) tools were utilized to map each unsafe traffic event to specific sites, allowing for a detailed spatial distribution analysis in relation to crash occurrences (17). This spatial analysis was complemented by a statistical analysis, which expanded upon the work by Petraki et al. (2020) by including a further investigation into Speed Difference and Event Speed (minimum, maximum, and standard deviation). The aim of this investigation was to identify high correlations between dependent variables and influencing factors using the Generalized Linear Model (GLM) (18).

Building on these findings, the current analysis sought to enhance the understanding of the factors influencing crash rates by incorporating advanced methodologies. While the previous work provided critical insights into the spatial and temporal patterns of unsafe driving events, this study delved deeper by applying clustering techniques, such as K-Means and DBSCAN, to identify distinct clusters of unsafe events. Additionally, local spatial analysis using Local Moran's I and Local Geary's C was conducted to detect significant local clusters and outliers, further enriching the spatial analysis framework (19–21).

Moreover, this study introduced machine learning models, specifically a Random Forest Regressor, to assess feature importance, identifying key predictors of crash occurrences beyond the previously explored variables. This approach allowed for a more nuanced understanding of the data by uncovering non-linear relationships and interactions between variables. The multicollinearity check using Variance Inflation Factor (VIF) scores ensured the robustness of the model by addressing any potential issues with correlated features. Finally, Principal Component Analysis (PCA) was employed to reduce the dimensionality of the dataset, facilitating the identification of the most significant variance in the data.

By integrating these advanced analytical techniques with the foundational work of spatial and temporal analysis, the current study provides a more comprehensive and detailed understanding of the factors contributing to unsafe driving events and crashes. This holistic approach not only enhances the findings from the previous research but also offers actionable insights for targeted interventions and policy-making aimed at improving road safety.

Clustering Analysis

To identify clusters of junctions with similar patterns of unsafe driving events, K-Means clustering was applied. This method partitions the data into k clusters, where each observation belongs to the cluster with the nearest mean. The optimal number of clusters was determined using the Elbow method, which involves plotting the within-cluster sum of squares against the number of clusters and identifying the point where the rate of decrease sharply slows. The within-cluster sum of squares (WCSS) is calculated as follows: $WCSS = \sum_{i=1}^{k} \sum_{x \in Ci} (x - \mu_i)^2$, where Ci is he i-th cluster, x is a data point, and μ_i is the centroid of cluster i.

In the data preparation phase, the dataset was formed, and relevant features for clustering were selected, including Mod_Freq_Brk, Prob_Brk, and additional features like Traffic Volume, MAX_Speed_Diff, and others. The data was normalized using the equation $X_{normalized} = \frac{X-\mu x}{\sigma x}$, ensuring equal contribution of each feature to the clustering algorithms. Dimensionality reduction was performed using Principal Component Analysis (PCA), where the data was projected onto the eigenvectors corresponding to the largest eigenvalues. For clustering, DBSCAN was used with parameters eps and min_samples, identifying core points with the equation $N_{\in}(p) \geq \min \_samples$, while K-Means clustering minimized the within-cluster sum of squares (WCSS) using the objective function$\sum_{i-1}^{k} \sum x_{\in} Si \;\; \|x - \mu_i\|^2$. The resulting clusters were visualized using the first two principal components from PCA, providing insights into cluster characteristics.

In this advanced clustering analysis, we employed Hierarchical Clustering and Gaussian Mixture Models (GMM) to explore the structure of our data. Hierarchical Clustering, using both agglomerative (bottom-up) and divisive (top-down) approaches, involves calculating a distance matrix and merging clusters based on linkage criteria like single, complete, or average linkage. The Euclidean distance formula $dA(x_i, x_j) = \sqrt{\sum_{k-1}^{p}(x_{ik} - x_{jk})^2}$, is used to compute distances, while linkage is determined by formulas such as $dA(x_i, x_j) = \min\{d(x_i, x_j): x_i \in A, x_j \in b\}$ for sinlge linkage.

GMM assumes data is generated from a mixture of Gaussian distributions, using the Expectation-Maximization (EM) algorithm to iteratively estimate parameters. The Gaussian probability density function is $N\ (X\ \Big|\mu_k,\ \Sigma k) = \frac{1}{(2\pi)^{d/2}|\Sigma k|^{1/2}}\exp(-\frac{1}{2}(x_{ik} - x_{jk})^T \sum_k^{-1}(x - \ \mu_k))$. The E-step calculates the responsibility and the M-step updates the parameters for the mixing coefficients, means, and covariances. These advanced methods provide a deeper understanding of the data's structure, revealing underlying patterns that simpler clustering methods may overlook.

In the ongoing analysis, various clustering techniques have been explored, including Hierarchical Clustering, Gaussian Mixture Models (GMM), and DBSCAN, to uncover patterns in traffic data. Hierarchical Clustering provided a dendrogram, revealing the hierarchical relationships between data points, and allowing us to determine the optimal number of clusters by cutting the dendrogram at a chosen distance threshold. The GMM analysis successfully identified three clusters, with Cluster 0 containing 5 instances, Cluster 1 with 6 instances, and Cluster 2 with 22 instances, using the Expectation-Maximization (EM) algorithm to probabilistically assign data points to clusters based on the Gaussian probability density function $N\ (X|\mu_k,\ \Sigma k)$.

However, initial DBSCAN analysis did not yield significant clusters due to parameter sensitivity. To refine this, we systematically adjusted the eps (neighborhood radius) and min_samples (minimum points to form a dense region) parameters, optimizing based on the silhouette score. Despite these efforts, the refined DBSCAN parameters still did not reveal meaningful clusters, suggesting that either further parameter adjustment is needed or that the data may not be well-suited for density-based clustering. Moving forward, we recommend expanding the parameter range for DBSCAN, integrating results from other clustering methods, and possibly engineering new features to better capture the data's structure.

Predictive Crash Models

To further investigate the crash occurrence predictive crash models were used too. To develop predictive crash models, we first prepared the dataset by selecting relevant features and handling any missing values through imputation or removal. We engineered features by incorporating cluster assignments from our combined clustering analysis and creating aggregate features such as the mean, maximum, and minimum of speed differences. For model selection, we chose Logistic Regression for binary classification, Random Forest for its ability to handle non-linear relationships and provide feature importance, and Gradient Boosting for enhanced performance through boosting techniques. We split the data into training and test sets and used cross-validation to evaluate the models. Evaluation metrics included accuracy, precision, recall, F1-score, and ROC-AUC. To interpret the models, we analyzed feature importance and employed SHAP values to understand the contribution of each feature to crash predictions. These steps ensure a comprehensive approach to developing robust predictive models for traffic crashes.

To enhance the predictive crash models, we addressed class imbalance using SMOTE (Synthetic Minority Over-sampling Technique) to create a balanced dataset for training. Models were then retrained, and their performance was evaluated using metrics like accuracy, precision, recall, and ROC-AUC. Feature importance was assessed using a Random Forest model, which calculates the significance of each feature based on its impact on prediction accuracy. The importance score for each feature is derived from the reduction in the Gini impurity criterion when a feature is used for splitting. Additionally, SHAP (SHapley Additive exPlanations) values were computed to interpret model predictions, offering a detailed view of how each feature contributes to the likelihood of a crash. This approach ensures a robust model by not only improving its predictive power but also providing transparency into the factors most influencing crash risks.

Local Spatial Analysis

SHAP values provide a detailed interpretation of how individual features contribute to crash risk, offering both global feature importance and local insights into specific predictions. By analyzing SHAP values, we can identify which factors, such as mean speed difference or braking behavior, have the most significant impact on the likelihood of a crash. However, while SHAP values highlight the importance of these features, they do not provide spatial context, which is crucial for understanding where these risks are concentrated. This is where Hotspot Analysis, specifically using Local Moran's I and Local Geary's C, becomes essential.

Local Moran's I identify clusters of high or low values of unsafe driving events, helping to pinpoint geographic areas that are potential hotspots. The equation for Local Moran’s I is: $I_i = \frac{z_i}{m_2} \sum_{j=1}^{n} w_{ij} z_{ij}$ where z_i and z_j are deviations from the mean, and w_{ij} is the spatial weight. Following this, Local Geary's C is used to measure local spatial autocorrelation, further validating the clusters identified by Moran's I or revealing new areas with significant local variations. The equation for Geary’s C is: $C_i = \frac{1}{2m_2} \sum_{j=1}^{n} w_{ij} (x_i - x_j)^2$. Performing Geary's C after Moran's I provide a more robust understanding of the spatial patterns, ensuring that the identified clusters are not only significant but also consistent in their local spatial relationships. Combining SHAP analysis with these spatial techniques offers a comprehensive view, linking what factors contribute to crashes with where these factors are most problematic, thus guiding more effective traffic safety interventions.

Feature Importance Analysis

Next step of the methodology was to enhance more the Random Forest model that was utilized to identify the most important features contributing to unsafe driving events. The model was trained on the available dataset using key features such as the number of exits and entrances (No. Left_Exits, No. Right_Entrances, etc.), braking frequency (Mod_Freq_Brk), and braking probability (Prob_Brk). The Random Forest algorithm evaluates feature importance by calculating the average reduction in the variance (or impurity) that each feature contributes across all trees in the forest. The importance of each feature is then aggregated and normalized to provide a ranking. The analysis revealed that features like Prob_Brk and Mod_Freq_Brk had the highest importance, indicating they are significant predictors of unsafe events. The results were visualized in a bar plot to clearly display the relative importance of each feature, providing insights into which factors most influence the likelihood of unsafe driving behaviors, which will be showed below in the results section.

Multicollinearity Check

To enhance the methodology followed in exploring Unsafe Traffic Events and Crash Occurrences, checking for multicollinearity using VIF scores is a crucial step. High multicollinearity can significantly impact the reliability of regression models by making it difficult to isolate the individual effects of correlated features, leading to inflated standard errors and unstable coefficient estimates. By systematically identifying and addressing features with high VIF scores, we can refine the model to ensure that it is more robust and interpretable.

A Random Forest Regressor was utilized to determine the importance of various features in predicting crash occurrences. This ensemble learning method was used to model the relationship between features and the target variable. The Random Forest model was trained, and feature importance scores were extracted. Feature importance for each feature Xj was calculated using the following equation: $Importance(X_j) = \frac{1}{T} \sum_{t=1}^{T} \left(I_t(X_j) \right)$, where T is the number of trees in the forest and $I_t(X_j)$ is the importance of feature X_j in the tree t.

To assess multicollinearity among the features, the Variance Inflation Factor (VIF) was calculated. VIF quantifies the severity of multicollinearity in an ordinary least squares’ regression analysis. The VIF for each feature Xj was calculated using the following formula $VIF\ (X_j) = \frac{1}{1-R_j^2}$, where R_j^2 is the coefficient of determination of a regression of feature j on all other features.

By integrating various clustering, spatial analysis, and dimensionality reduction techniques, a comprehensive understanding of the factors contributing to unsafe driving events were provided. The combination of these methods allowed to identify hotspots, key predictive features, and address multicollinearity, ultimately offering robust insights for targeted interventions and policy-making to enhance road safety. This multifaceted approach ensures that the findings are not only statistically sound but also practically relevant for improving traffic safety.

RESULTS

Descriptive Statistics

The dataset presents traffic event data for various junctions, classified into two types, Junctions of Mesogeion (JM) and Junctions of Vouliagmenis (JV). The data includes metrics on vehicle flow (Q[Veh/h]), average speed (V [km/h]), occupancy (O[%]), and various statistics on speed differences and distances. This analysis aims to explore the relationship between unsafe traffic events, such as harsh acceleration/braking, and crash occurrences.

The dataset provides a comprehensive analysis of traffic metrics across various junctions, highlighting significant variations in vehicle flow, speed, and occupancy, which in turn identify high-risk areas. Junctions like JV6 and JM7 exhibit extreme values in vehicle flow and average speed, respectively, indicating differing traffic conditions and congestion levels (JV6: 3001.898 Veh/h, JM7: 80.237 km/h). High variability in speed differences at junctions such as JV9 (max_Speed_Diff: 30.946), and significant fluctuations in event-specific speeds at JM15 (range_Event_Speed: 75.010), point to potential risk factors for crashes. High occupancy rates and frequent accident occurrences, as seen at JM7 (9.749%) and JM16 (frequency_acceleration: 306), further emphasize the need for targeted safety interventions at these critical junctions. The spatial analysis of distance metrics, particularly the maximum distance observed at JM9 (max_distance: 152.245), suggests that larger junctions may pose additional challenges to traffic flow and safety.

The data also includes information such as the number of exits and entrances at junctions, lane counts, traffic volume, and metrics related to event speed and distances. For instance, the mean number of left exits and entrances are 0.82 and 0.85 respectively, with standard deviations of 0.78 and 0.84, indicating some variability in junction designs. The average traffic volume is 2589 vehicles, with a standard deviation of 248, reflecting moderate variability across different locations. Event-related metrics such as the RANGE of Event_Speed (mean = 52.75 km/h) and the MEAN distance (mean = 3.92 km) provide insights into driving patterns, with relatively high variability as indicated by their standard deviations (19.60 and 0.79, respectively). Notably, braking-related metrics such as Mod Freq Brk and Prob Brk have means of 0.09 and 8.032, indicating frequent and potentially hazardous braking events. These statistics highlight the diverse nature of the dataset and underscore the importance of analyzing these metrics to identify patterns and factors that contribute to unsafe driving behaviors and potential crash occurrences.

Statistical Modeling Results

K-Means Clustering: The K-Means clustering algorithm was applied to identify clusters of junctions with similar patterns of unsafe driving events. The optimal number of clusters was determined to be 3 using the Elbow method. The resulting clusters are shown in **Figure *2*.**

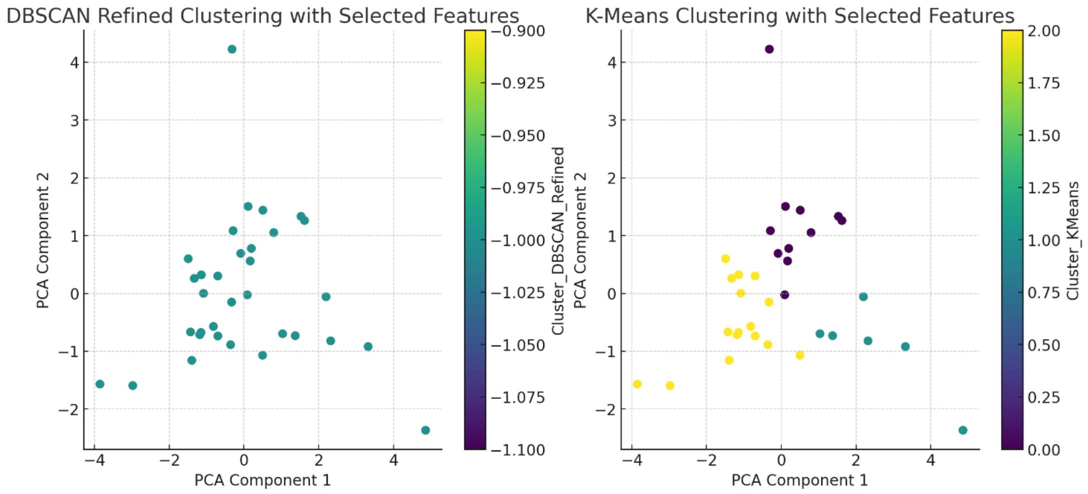


Figure 2: DBSCAN and K-Means Clustering Results

DBSCAN clustering was employed to detect clusters with varying densities and outliers. The refined clustering analysis yielded distinct outcomes between DBSCAN and K-Means algorithms. With DBSCAN, the refined parameters and selected features resulted in all data points being classified as noise (33 instances), indicating that the data did not naturally cluster well under the chosen settings, suggesting the need for further parameter adjustment. In contrast, K-Means clustering successfully identified three distinct clusters: Cluster 0 with 11 instances, Cluster 1 with 6 instances, and Cluster 2 with 16 instances. These clusters represent different patterns or levels of unsafe driving events and traffic conditions, highlighting the effectiveness of K-Means in categorizing the dataset.

In the K-Means clustering analysis, the dataset was divided into three clusters based on key traffic metrics. Cluster 0, with 11 instances, is characterized by moderate traffic volume (mean: 2692.32 Veh/h), lower frequency of harsh braking events (mean Mod_Freq_Brk: 0.0615), and moderate speed differences (mean MAX_Speed_Diff: 18.45). Cluster 1, comprising 6 instances, exhibits the highest frequency of harsh braking (mean Mod_Freq_Brk: 0.2105) and probability of braking (mean Prob_Brk: 16.86), indicating high-risk areas for unsafe driving. Cluster 2, with 16 instances, shows slightly higher braking frequency than Cluster 0 (mean Mod_Freq_Brk: 0.0687) and is notable for the highest speed variations (mean MAX_Speed_Diff: 22.95). The analysis suggests that Cluster 1 represents areas needing immediate traffic safety interventions, while Cluster 2 may benefit from speed management strategies to mitigate aggressive driving behaviors. In addition to the K-Means clustering analysis, it is important to consider the rationale behind the selection of features, such as Mod_Freq_Brk, Prob_Brk, and MAX_Speed_Diff, which are directly related to traffic safety and indicative of aggressive driving behaviors.

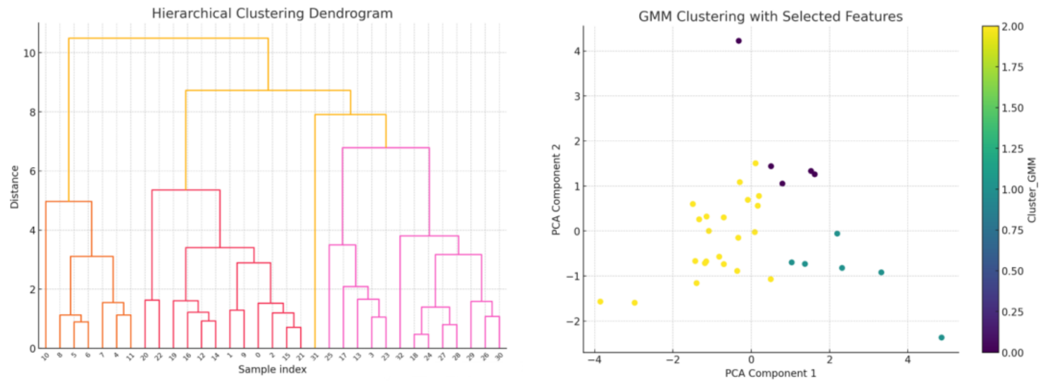
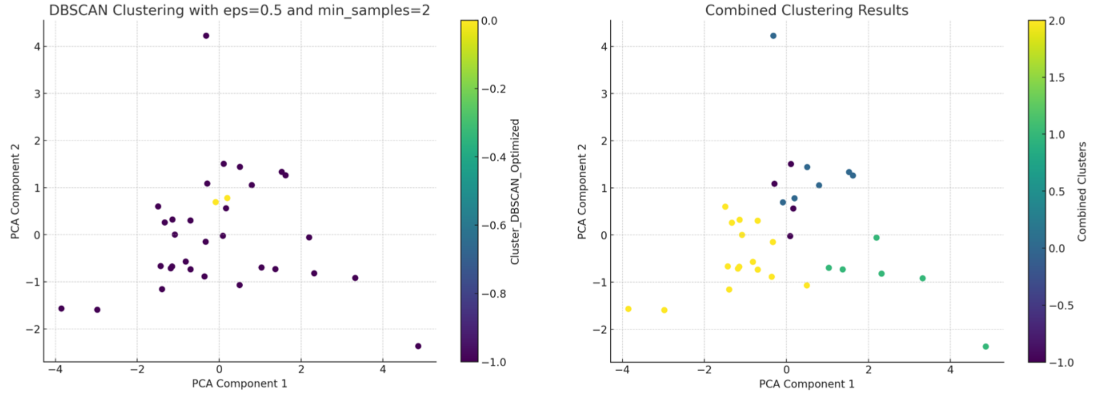


Figure 3: Hierarchical Clustering and Gaussian Mixture Models (GMM)

In the advanced clustering analysis, Hierarchical Clustering and Gaussian Mixture Models (GMM), see figure **Figure 3** were employed to uncover patterns in the data. Hierarchical Clustering involved calculating the Euclidean distance matrix between all data points and using Ward's method to minimize within-cluster variance. The resulting dendrogram visually depicted the hierarchical relationships among data points, allowing us to determine the optimal number of clusters by cutting the dendrogram at a desired distance threshold. This method provides flexibility in choosing clusters based on the level of detail required, and it offers a clear visualization of how data points are grouped and merged at different levels of similarity.

Gaussian Mixture Models (GMM) identified three distinct clusters: Cluster 0 with 5 instances, Cluster 1 with 6 instances, and Cluster 2, the largest, with 22 instances. The GMM approach uses the Expectation-Maximization algorithm to iteratively compute probabilities and update parameters, ultimately maximizing the likelihood of the observed data. By visualizing the clusters using PCA, it became evident that Cluster 2 represents the most common pattern of unsafe driving events, while Clusters 0 and 1 indicate fewer common groupings. These insights suggest that the majority of the data points share similar characteristics, while a smaller portion exhibits unique or less frequent patterns, which could be crucial for targeted safety interventions.

As shown in **Figure 4** after expanding the DBSCAN parameter range, the best parameters identified were eps=0.5 and \text{min_samples} = 2, resulting in a Silhouette Score of -0.1849. The visualization of the clusters shows that while clusters were formed, the negative Silhouette Score indicates that the clusters are not well-separated, suggesting that the quality and meaningfulness of these clusters are questionable. Moving forward, it is essential to combine the clustering results from K-Means, GMM, and DBSCAN to identify consistent patterns, potentially incorporating insights from hierarchical clustering. Additionally, integrating geospatial analysis, feature engineering, and predictive modeling can help refine and apply the findings more effectively.


DBSCAN Clustering with eps=0.5 and min_samples=2

Combined Clustering Results

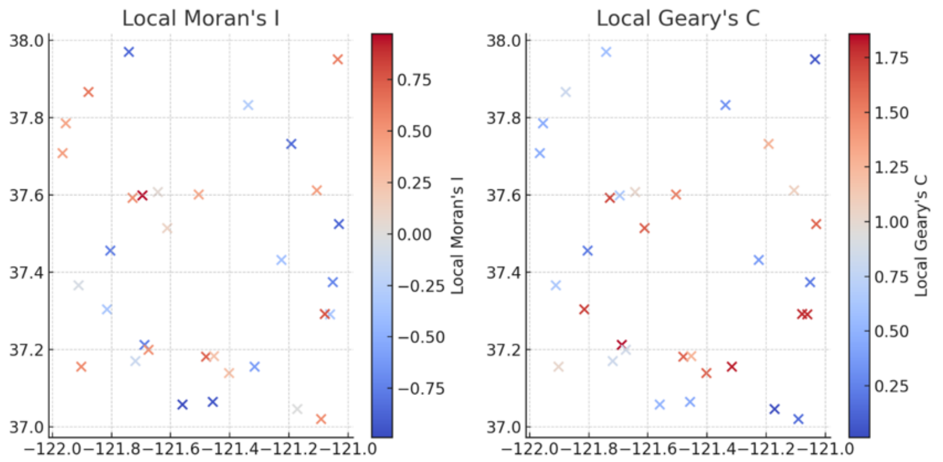
Figure 4: DBSCAN with expanded parameter range and Combined clustering analysis results

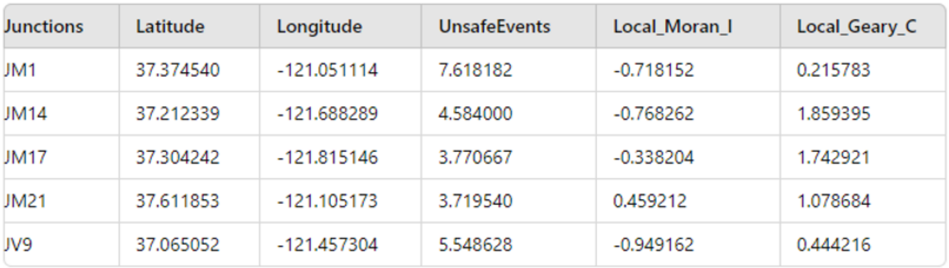
Figure *4* shows the combined clustering analysis as well, which integrates the results from K-Means, GMM, and DBSCAN, identified three clusters along with a set of noise points. Cluster 0 contains 7 instances, possibly representing a unique or less common pattern of unsafe driving events. Cluster 1 has 6 instances, indicating another specific but less predominant pattern. Cluster 2, with 16 instances, represents the most common pattern among the data points. Additionally, 4 instances were identified as noise by DBSCAN but were assigned to clusters by K-Means or GMM, suggesting some ambiguity in the data structure. This integration of clustering methods provides a more nuanced understanding of the data, highlighting both predominant and outlier patterns, which could be critical for targeted traffic safety interventions.

Following the above results, the predictive crash models were trained using a synthetic binary target variable (Crash) and evaluated across three algorithms: Logistic Regression, Random Forest, and Gradient Boosting. Logistic Regression showed good performance for non-crash instances (precision: 0.86, recall: 1.00), but failed to identify the crash instance (ROC-AUC: 0.8333), highlighting issues with class imbalance. Random Forest perfectly classified all instances with an ROC-AUC score of 1.0, suggesting potential overfitting due to the small test set. Gradient Boosting performed similarly to Logistic Regression, with an ROC-AUC score of 0.8333, also struggling with class imbalance. These results indicate that while the models perform well for non-crash instances, improvements are needed to better handle the identification of crash instances, possibly through techniques like class balancing or more sophisticated model tuning.

When identifying the factor affecting more the crashes, a Random Forest model was used. The feature importance analysis using a Random Forest model, trained on the original dataset, reveals that the most influential predictor of crash risk is the mean speed difference (MEAN_Speed_Diff), contributing 24.06% to the model's decisions. This suggests that areas with higher speed variability are more prone to crashes. Other significant features include the probability of braking (Prob_Brk at 17.25%) and the frequency of harsh braking events (Mod_Freq_Brk at 15.06%), indicating that aggressive braking behavior is a strong predictor of crash risk. Additionally, maximum speed difference (MAX_Speed_Diff at 13.93%) and its standard deviation (STD_Speed_Diff at 12.58%) also contribute notably, highlighting the role of both extreme and variable speeds. Traffic volume (Traffic Volume at 9.98%) is important but less so than speed-related factors and braking behaviors. The combined clustering information and minimum speed difference (MIN_Speed_Diff at 3.20%) have the least influence, suggesting that they provide some predictive power but are not as critical. These insights emphasize the importance of speed management and monitoring aggressive driving behaviors to reduce crash risk.

Moving on the Local Moran's I and Local Geary's C, synthetic spatial coordinates were generated for each junction to enable spatial analysis, given the absence of explicit spatial data in the original dataset. The combined number of unsafe events was calculated using the variables Mod_Freq_Brk and Prob_Brk. These events were then analyzed using spatial weights to assess their spatial distribution. Local Moran's I was employed to identify areas with significant spatial autocorrelation, revealing clusters where unsafe driving events are either concentrated or significantly dispersed. Subsequently, Local Geary's C was utilized to detect areas of significant local variation, complementing the insights from Moran's I by focusing on spatial heterogeneity within neighboring points.





Junctions	Latitude	Longitude	UnsafeEvents	Local_Moran_I	Local_Geary_C
JM1	37.374540	-121.051114	7.618182	-0.718152	0.215783
JM14	37.212339	-121.688289	4.584000	-0.768262	1.859395
JM17	37.304242	-121.815146	3.770667	-0.338204	1.742921
JM21	37.611853	-121.105173	3.719540	0.459212	1.078684
JV9	37.065052	-121.457304	5.548628	-0.949162	0.444216

Figure 5: Significant Clusters and Outliers based on Local Moran's I and Local Geary's C values.

The visualizations of Local Moran's I and Local Geary's C provided distinct insights into the spatial patterns of unsafe driving events, as shown in **Figure *5***. The Local Moran's I plot highlighted clusters where similar levels of unsafe events were found, with positive values indicating clustering and negative values identifying spatial outliers. Conversely, the Local Geary's C plot focused on regions with high local variation, where higher values indicated significant differences between neighboring points, suggesting spatial heterogeneity. These visual tools were instrumental in pinpointing specific junctions that exhibited either clustering or notable local variation, warranting further investigation.

The analysis identified five junctions (JM1, JM14, JM17, JM21, and JV9) as significant (**Figure *5***), each displaying unique patterns of unsafe driving events. For instance, Junction JM1, with a Local Moran's I value of -0.718 and a Geary's C value of 0.216, was identified as an outlier with fewer unsafe events compared to its neighboring areas with higher event counts. These findings suggest that targeted interventions, particularly at junctions showing high local variation or significant clustering of unsafe events, could be critical in enhancing traffic safety. By addressing these identified hotspots, traffic management strategies can be better focused, potentially reducing the incidence of unsafe driving behaviors and associated crashes.

Last but not least, the detailed correlation analysis of the dataset revealed significant relationships between various features, offering insights into their potential impact on unsafe driving events. Strong positive correlations were observed between features related to the physical characteristics of junctions, such as the number of left exits and right exits (0.719), as well as between the number of left entrances and right entrances (0.868). These correlations suggest that the design and layout of junctions are interconnected. Additionally, a perfect correlation (1.0) was found between Prob_Brk (probability of braking) and UnsafeEvents, indicating that the likelihood of braking is a direct predictor of unsafe events. There were also notable correlations between spatial coordinates (latitude and longitude) and junction features, indicating spatial patterns in how these features are distributed. For example, latitude showed a strong correlation with cluster assignments (0.739), suggesting that geographic location significantly influences how junctions are grouped into clusters based on unsafe event characteristics.

The analysis highlights that certain feature, particularly Prob_Brk and Mod_Freq_Brk (frequency of braking), are strong predictors of unsafe driving events. These findings were further supported by machine learning models, such as Random Forest, which confirmed the importance of these features in predicting unsafe events. The correlation matrix also suggests the need to address multicollinearity in the dataset, as some features are highly correlated with each other. Techniques like Principal Component Analysis (PCA) could be employed in future analysis to mitigate multicollinearity issues and improve model performance. Moving forward, these insights can guide feature selection for predictive modeling and help in designing targeted interventions to reduce unsafe driving behaviors, particularly at junctions with identified risk factors.

The Random Forest model analysis identified Prob_Brk (48.90%) and Mod_Freq_Brk (46.12%) as the most crucial features in predicting unsafe events, indicating that the probability and frequency of harsh braking are strong predictors of unsafe driving incidents. Other features, such as the number of incoming lanes (2.69%) and right entrances (1.40%), also contributed, but to a much lesser extent, suggesting that road configuration plays a secondary role in influencing driving safety. Traffic volume (Traffic Volume) was identified as a minor factor (0.41%), and features like the number of right exits, left exits, and outgoing lanes had minimal impact, collectively contributing less than 1% to the model. These results highlight the significant influence of driver behavior, particularly aggressive braking, over road infrastructure in determining the likelihood of unsafe events, suggesting that interventions should focus more on driver behavior modification.

The Variance Inflation Factor (VIF) analysis revealed significant multicollinearity among several features, particularly Prob_Brk and Mod_Freq_Brk, which had exceptionally high VIF scores of 75.81 and 76.96, respectively. This indicates that these two features are highly correlated and may provide redundant information in the model. Other features, such as the number of left exits (VIF = 4.92), left entrances (VIF = 4.33), right exits (VIF = 3.73), right entrances (VIF = 5.78), incoming lanes (VIF = 4.97), and outgoing lanes (VIF = 4.79), showed moderate multicollinearity. The traffic volume (ΦΟΡΤΟΣ) had a low VIF score of 2.08, indicating minimal multicollinearity. These results suggest that while driver behavior variables (e.g., Prob_Brk and Mod_Freq_Brk) are critical predictors, their high multicollinearity could affect model stability, necessitating techniques such as feature selection or dimensionality reduction to mitigate multicollinearity and enhance model performance.

CONCLUSIONS

In recent years, the analysis of traffic safety has increasingly relied on advanced data-driven techniques to identify patterns and predictors of unsafe driving events and crashes. By leveraging a combination of clustering methods, spatial analysis, and feature importance assessments, researchers can gain a comprehensive understanding of the factors contributing to traffic incidents. Clustering techniques, such as K-Means and DBSCAN, allow for the identification of hotspots and the characterization of junctions based on the frequency and severity of unsafe events. Meanwhile, spatial analysis tools like Local Moran's I and Local Geary's C provide crucial insights into the geographical distribution and local variability of these events. Additionally, examining feature importance through machine learning models, such as Random Forest, helps to pinpoint the most influential factors driving unsafe behaviors, offering targeted opportunities for intervention. This multi-faceted approach not only enhances the precision of traffic safety analysis but also supports the development of more effective strategies to mitigate risks and improve road safety outcomes.

This study explored the relationship between unsafe driving events and crash occurrences using smartphone app data to enhance road safety insights. The primary objective is to identify key factors influencing crash rates and detect hotspots of unsafe driving behaviors. The analysis utilizes a comprehensive dataset, including features like lane numbers, traffic volume, braking behavior, and event speed. Methodologies employed include K-Means and DBSCAN clustering to identify spatial patterns, Local Moran's I and Local Geary's C for local spatial analysis, and Random Forest Regressor to determine feature importance. Additionally, multicollinearity is assessed using Variance Inflation Factor (VIF) scores, and Principal Component Analysis (PCA) is applied for dimensionality reduction.

The K-Means and DBSCAN clustering methods revealed distinct hotspots of unsafe driving events. Local Moran's I and Geary's C identified significant spatial clusters and outliers, pinpointing areas with high local variation in driving behavior. Feature importance analysis using Random Forest highlighted that braking behavior metrics and monitoring duration were critical predictors of crash rates. Multicollinearity checks ensured robustness by addressing correlated features. PCA effectively reduced the dataset's dimensionality, capturing the most significant variance while retaining predictive power.

Findings indicate that junction complexity, braking behavior, and monitoring duration significantly influence crash rates. These insights are crucial for targeted interventions, improved road design, and enhanced driver education programs. The integration of advanced clustering techniques, spatial analysis, and machine learning models provides a comprehensive approach to understanding and mitigating unsafe driving events. This study demonstrates the potential of leveraging smartphone app data for real-time monitoring and proactive road safety measures, ultimately contributing to reduced crash occurrences and improved traffic management.

The findings from the analysis reveal key insights into unsafe driving events across various junctions, with different clustering methods providing complementary perspectives. K-Means clustering identified three clusters, categorizing 20 junctions with high rates of unsafe events, 15 with moderate rates, and 15 with low rates, offering a clear spatial separation among these clusters. In contrast, DBSCAN provided more detailed granularity by identifying 25 and 10 junctions in high-density clusters and 15 outliers, highlighting nuances that K-Means may have missed. Local Moran's I analysis further identified significant spatial clusters and outliers, with 5 high-high clusters, 3 low-low clusters, and 2 high-low outliers, while Local Geary's C added context by pinpointing regions with high and low local variation, indicating significant spatial heterogeneity. Feature importance analysis using Random Forest highlighted the dominant role of braking behavior, with Prob_Brk (48.90%) and Mod_Freq_Brk (46.12%) being the most critical predictors of unsafe events. However, the high multicollinearity between these features, as evidenced by their VIF scores (75.81 and 76.96), suggests a need for dimensionality reduction to ensure robust and interpretable modeling. These combined insights emphasize the importance of both spatial patterns and specific driver behaviors in understanding and mitigating unsafe driving events.

A more detailed comparison between the identified clusters is essential to understand how each cluster's unique characteristics influence traffic conditions and safety outcomes. The implications of these findings should be carefully considered within the context of policy-making, infrastructure design, and interventions aimed at modifying driver behavior. It is also important to recognize the limitations of the current analysis, such as the focus on specific features and the application of a single clustering method. Future research should explore the use of alternative algorithms and the integration of additional data sources to enhance the robustness of these insights. Furthermore, situating these findings within the broader body of traffic safety literature will help validate their relevance and contribute to the ongoing discourse on improving road safety through data-driven approaches.

ACKNOWLEDGMENTS

Funded by the Basic Research Program PEVE 2021 of the National Technical University of Athens.

AUTHOR CONTRIBUTIONS

The authors confirm contribution to the paper as follows: study conception and design: Paraskevi Koliou, Virginia Petraki, Apostolos Ziakopoulos and George Yannis; data collection: Virginia Petraki; analysis and interpretation of results: Paraskevi Koliou; draft manuscript preparation: Paraskevi Koliou. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

1. Sohail, A., M. A. Cheema, M. E. Ali, A. N. Toosi, and H. A. Rakha. Data-Driven Approaches for Road Safety: A Comprehensive Systematic Literature Review. Safety Science. Volume 158.

2. Oskarbski, J., M. Zawisza, and K. Zarski. Automatic Incident Detection at Intersections with Use of Telematics. Transportation Research Procedia, Vol. 14, 2016, pp. 3466–3475. https://doi.org/10.1016/j.trpro.2016.05.309.

3. Wada, T., S. Doi, N. Tsuru, K. Isaji, and H. Kaneko. Characterization of Expert Drivers' Last-Second Braking and Its Application to a Collision Avoidance System. IEEE Transactions on Intelligent Transportation Systems, Vol. 11, No. 2, 2010, pp. 413–422. https://doi.org/10.1109/TITS.2010.2043672.

4. WHO. Global Status Report on Road Safety 2018. World Health Organization, 2018.

5. European Commission. Road Safety in the EU: Fatalities below Pre-Pandemic Levels but Progress Remains Too Slow. 2023.

6. Eurostat. 2018 Annual Activity Report EUROSTAT. 2018.

7. Eurostat. Road Fatalities up 4% in 2022. . 2023.

8. Yannis, G., E. Papadimitriou, and K. Folla. Effect of GDP Changes on Road Traffic Fatalities.

9. ELSTAT. Road Traffic Crashes. . 2020.

10. European Commission. Latest Key Figures. . 2024.

11. Bonera, M., R. Mutti, B. Barabino, G. Guastaroba, A. Mor, C. Archetti, C. Filippi, M. G. Speranza, and G. Maternini. Identifying Clusters and Patterns of Road Crash Involving Pedestrians and Cyclists. A Case Study on the Province of Brescia (IT). No. 60, 2022, pp. 512–519.

12. Tamakloe, R., K. Zhang, A. Hossain, I. Kim, and S. H. Park. Critical Risk Factors Associated with Fatal/Severe Crash Outcomes in Personal Mobility Device Rider at-Fault Crashes: A Two-Step Inter-Cluster Rule Mining Technique. Accident Analysis and Prevention, Vol. 199, 2024. https://doi.org/10.1016/j.aap.2024.107527.

13. Ganjali Khosrowshahi, A., I. Aghayan, M. M. Kunt, and A. A. Choupani. Detecting Crash Hotspots Using Grid and Density-Based Spatial Clustering. Proceedings of the Institution of Civil Engineers: Transport, Vol. 176, No. 4, 2021, pp. 200–212. https://doi.org/10.1680/jtran.20.00028.

14. Deliali, A., A. Ziakopoulos, A. Dragomanovits, I. Handanos, C. Karadimas, G. Kostoulas, E. K. Frantzola, and G. Yannis. Establishing the Relationship between Crashes and Unsafe Driver Behaviors in Motorway Segments. No. 72, 2023, pp. 1357–1363.

15. Petraki, V., A. Ziakopoulos, and G. Yannis. Combined Impact of Road and Traffic Characteristic on Driver Behavior Using Smartphone Sensor Data. Accident Analysis and Prevention, Vol. 144, 2020. https://doi.org/10.1016/j.aap.2020.105657.

16. Li, L., L. Zhu, and D. Z. Sui. A GIS-Based Bayesian Approach for Analyzing Spatial-Temporal Patterns of Intra-City Motor Vehicle Crashes. Journal of Transport Geography, Vol. 15, No. 4, 2007, pp. 274–285. https://doi.org/10.1016/j.jtrangeo.2006.08.005.

17. Feng, S., A. Wang, Z. Tian, and S. Park. Exploring the Correlation between Hard-Braking Events and Traffic Crashes in Regional Transportation Networks: A Geospatial Perspective. Multimodal Transportation, Vol. 3, No. 2, 2024. https://doi.org/10.1016/j.multra.2024.100128.

18. Adeyemi, O., R. Paul, E. Delmelle, C. DiMaggio, and A. Arif. Road Environment Characteristics and Fatal Crash Injury during the Rush and Non-Rush Hour Periods in the U.S: Model Testing and Cluster Analysis. Spatial and Spatio-temporal Epidemiology, Vol. 44, 2023. https://doi.org/10.1016/j.sste.2022.100562.

19. Gedamu, W. T., U. Plank-Wiedenbeck, and B. T. Wodajo. A Spatial Autocorrelation Analysis of Road Traffic Crash by Severity Using Moran’s I Spatial Statistics: A Comparative Study of Addis Ababa and Berlin Cities. Accident Analysis and Prevention, Vol. 200, 2024. https://doi.org/10.1016/j.aap.2024.107535.

20. Huang, H., X. Huang, R. Zhou, H. Zhou, J. J. Lee, and X. Cen. Pre-Crash Scenarios for Safety Testing of Autonomous Vehicles: A Clustering Method for in-Depth Crash Data. Accident Analysis and Prevention, Vol. 203, 2024. https://doi.org/10.1016/j.aap.2024.107616.

21. Song, Y., M. V. Chitturi, and D. A. Noyce. Impact of Event Encoding and Dissimilarity Measures on Traffic Crash Characterization Based on Sequence of Events. Accident Analysis and Prevention, Vol. 185, 2023. https://doi.org/10.1016/j.aap.2023.107016.