https://doi.org/10.3311/PPtr.22392



**Periodica Polytechnica Transportation Engineering**, 52(1), pp. 33–39, 2024

# Exploring the Factors Influencing Traffic Accidents: An Analysis of Black Spots and Decision Tree for Injury Severity

Pires Abdullah[1,2*], Tibor Sipos[1,3]

[1] Department of Transport Technology and Economics, Faculty of Transportation Engineering and Vehicle Engineering, Budapest University of Technology and Economics, Műegyetem rkp. 3, H-1111 Budapest, Hungary
[2] Department of Spatial Planning, College of Spatial Planning, University of Duhok, Zakho Street 38, 1006 AJ Duhok, Kurdistan Region, Iraq
[3] Directorate for Strategy, Research & Development and Innovation, KTI Institute for Transport Sciences Non-profit Ltd., Than Károly u. 3-5, 1119 Budapest, Hungary
* Corresponding author, e-mail: pires.abdullah@edu.bme.hu

Received: 12 April 2023, Accepted: 10 October 2023, Published online: 31 October 2023

**Abstract**

This research aimed to examine the spatial distribution of road traffic accidents in Budapest, Hungary. The primary objective was to identify the factors associated with traffic accidents on the city's transportation network and to determine the locations of the most frequent accidents during peak and off-peak hours. A quantitative methodology was employed in this study, utilizing a dataset of recent accidents that occurred between 2019 and 2021, classified into peak and off-peak incidents. The data was analyzed using Python software and Quantum Geographic Information System (QGIS) tools for big data analytics. These programs enabled the creation of spatial maps of the study area and the identification of accident spots based on latitude and longitude information. A decision tree classification approach was used in the machine-learning method implemented with Python software. Additionally, the dataset file was uploaded to QGIS, which applied the heatmap (Kernel Density Estimation) algorithm to identify accident hotspots. The study findings revealed that the city center was the most common location for accidents overall, with peak and off-peak times, lanes, and days of the week investigated as potential contributing factors. The target variable was the number of accidents involving serious and minor injuries, which were found to be significantly associated with the identified accidents in this study.

**Keywords**

traffic accident analysis, road safety, decision tree, black spots, machine learning

## 1 Introduction

The issue of road traffic accidents has been a persistent challenge globally, prompting countries to seek effective solutions. Despite the integral role road transport plays in facilitating economic growth, it imposes significant socio-economic costs (Zefreh and Torok, 2021), accounting for approximately 2% of the European Union's GDP annually. The annual loss of nearly 30,000 lives due to road accidents highlights the serious implications of this issue (Török, 2017). It is crucial to identify traffic-related issues to increase the likelihood of policy and strategy acceptance aimed at addressing these issues (Shatanawi et al., 2020). The impact of road accidents is widespread, affecting drivers, passengers, vulnerable groups, and both public and private property. Researchers have endeavored to examine the factors contributing to road accidents and suggest measures to mitigate their occurrence (Wisutwattanasak et al., 2023). Human errors, vehicle malfunctions, and defects in the road environment are among the multiple factors that could lead to road accidents (Turoń et al., 2019).

Currently, traffic and transport management systems are a common concern in many cities, and city officials require an efficient and uncomplicated approach to identifying significant issues (Kaparias et al., 2020). A study carried out in the Czech Republic examined the factors that contribute to traffic accidents, which are often attributed to human behavior. The study revealed that specific groups of drivers, such as young and elderly drivers and those who engage in risky driving, are more susceptible to accidents. Irrespective of age or gender, inattention was the primary

**Cite this article as:** Abdullah, P., Sipos, T. (2024) "Exploring the Factors Influencing Traffic Accidents: An Analysis of Black Spots and Decision Tree for Injury Severity", Periodica Polytechnica Transportation Engineering, 52(1), pp. 33–39. https://doi.org/10.3311/PPtr.22392

factor responsible for accidents. Women were more likely to have panic reactions, while men often misjudged road situations, leading to accidents (Bucsuházy et al., 2020).

The visualization of data is an essential aspect of analyzing intricate datasets as it provides multiple perspectives on the information under examination. Various computer software programs aid in the representation of data, including GIS-based file formats like shape files that allow the geographic representation of urban road networks and spatial events. An example of such visualization is presented in a study (Sobral et al., 2019) that employed GIS-based visualization to identify road segments with a higher potential for vehicle crashes based on ranking and estimation. The study utilized map-based visualization to display the accident locations and heat maps to represent their distribution. Pairwise relationships between variables, such as accidents associated with multiple properties like fatality or injury, were explored using scatterplots and heat map matrices (Fan et al., 2019).

In a recent research study (Dereli and Erdogan, 2017), a model that employed Geographic Information Systems (GIS) and statistical techniques was employed to determine traffic accident hotspots or "black spots" on 2408 state roads in Turkey with the objective of enhancing traffic safety and reducing accident rates. The study aimed to identify areas with a high incidence of traffic accidents, and GPS and GIS-compatible systems were utilized to gather coordinate information about the accident sites. Temporal and spatial analysis of traffic accidents occurring between 2005 and 2013 was performed to achieve this goal. Despite an increase in the number of registered vehicles and traffic intensity during this time, the analysis showed a decrease in traffic accident fatalities since 2008, although the overall number of traffic accidents and injuries had increased.

Two separate research studies, conducted by Sun et al. (2021) and Kurakina et al. (2020), focused on analyzing the rates of injuries and fatalities resulting from traffic collisions based on the primary function of different types of roads. The classification of road types included administrative roads, functional roads, urban expressways, and urban general roads. The research revealed that administrative roads had the highest death rate, followed by functional roads, among all road types. Moreover, the incidence of traffic accidents was found to be significantly (11.6 times) higher on urban general roads compared to urban expressways. The research findings provide crucial insights into the association between road types and the incidence of traffic accidents, which could facilitate transportation authorities in devising targeted interventions to reduce accidents and promote road safety.

In order to understand how traffic operates under different conditions throughout the day, separate studies were conducted during AM peak, Noon off-peak, and PM peak hours. The study revealed that variations in speed were associated with factors such as the number of lanes and access points, the presence of bus stops, and mean speed. Specifically, increases in these factors were found to be associated with increased speed variation. Conversely, traffic volume and traffic signal green times were associated with reduced speed variation. These findings have implications for road network planning and traffic management, as they can help engineers minimize speed differences and improve overall traffic flow. The study also highlights the importance of considering variations in traffic demand when designing road systems (Wang et al., 2016).

In order to identify the determinants of congestion, a generalized linear mixed-effects model was employed with spatiotemporal panel data, taking into account the variation of the impact of congestion across space and time. To investigate the determinants of congestion, a case study was conducted in Beijing using real data. The study found that the severity of congestion is mostly determined by the types of traffic accidents, types of vehicles involved, and occurrence time. More specifically, the level of congestion during peak hours, such as morning and afternoon peaks, was found to be 5.87% and 6.57% more severe than that during off-peak hours, respectively. These findings provide valuable insights for traffic management and transportation planning aimed at reducing congestion levels in urban areas (Zheng et al., 2020).

The use of data visualization in analyzing traffic safety has not yet been fully explored. To address this gap, researchers integrated visualization techniques into machine learning algorithms to investigate traffic accident data in the UK from 2017. Their proposed hybrid algorithm, Light Gradient Boosting Machine-Tree-structured Parzen Estimator (LightGBM-TPE), identified four risk factors – "Longitude", "Latitude", "Hour", and "Day_of_Week" – as strongly associated with accident severity. These findings were further verified through data visualization (Li et al., 2022). In another study, several other variables, such as angle accidents, injury severity, traffic volume and composition, morning or pre-dawn hours, and blockage of overtaking lanes, were found to have significant but uniform impacts on accident clearance time

(Zeng et al., 2023). The ability to improve traffic safety and manage traffic lane capacity is one advantage of the binary integer modeling approach found in a related study (Pauer and Török, 2021).

Traditional research methodologies face challenges when it comes to examining the behavioral aspects of traffic accidents due to their rare and unpredictable nature (Clarke et al., 1998). To overcome this challenge, decision tree (DT) analysis can be used to identify the factors that influence the outcome of accidents and construct a predictive model. A DT typically consists of a root, branches, nodes, and leaves, with the root representing the primary node and branches growing from it representing different values that can be expected from various factors such as people's perceptions, social groups, or specific characteristics. The leaves represent the final node, and no branches grow from them. Each node represents a variable, and the branches represent the set of values that are associated with that variable. Through DT analysis, researchers can gain a better understanding of the complex interplay of factors that contribute to traffic accidents and develop targeted interventions to prevent them (Bordarie, 2019).

In a research study conducted by Taamneh (2018), a decision tree algorithm was employed to identify the most significant variables affecting the understanding of different categories of traffic signs by drivers belonging to diverse socio-economic backgrounds. The decision tree algorithm aimed to segment the data and recognize homogeneous groups based on the dependent variable. The algorithm's iterative process continued until the stopping level was achieved. The decision tree was constructed based on two criteria, namely entropy and the Gini index (Figueira et al., 2017). The entropy was utilized as a measure of the heterogeneity of the input data for classification, and the decision tree was constructed with the aim of reducing entropy while being consistent with the dataset and having the minimum number of nodes possible. Meanwhile, the Gini index, which was developed by Conrado Gini in 1912, was used to assess the level of data heterogeneity and node impurity. A node was regarded as pure when the Gini index value was zero, whereas approaching one indicated impurity (Figueira et al., 2017).

To summarize, the literature suggests that traffic accidents have a considerable adverse impact on society, yet there is limited research exploring the diverse factors influencing accident profiling in different regions. Thus, the primary aim of this study is to undertake a spatial analysis of traffic accidents in the transportation network and identify the accident locations most frequently occurring in the city. Furthermore, this study aims to identify the factors linked to accidents involving serious and minor injuries as reported in the city. Through this research, a novel contribution to the existing literature on traffic accidents will be made.

## 2 Methods

In today's world, the amount of data produced is rapidly increasing, necessitating the use of data mining methods and software applications for analyzing large datasets. In this study, a Python language program was employed to analyze the dataset. The program offers a web-based application process that facilitates the creation of clear and succinct reports through its visualization capabilities, including various libraries with efficient visualization tools. The study considered a set of independent transport-related variables, while the dependent variable was defined as the occurrence of accidents resulting in minor and serious injuries, represented dichotomously as "0" and "1", respectively, over three consecutive years. Python programming was utilized to implement the decision tree process. The heat map was generated in QGIS using the kernel density estimation algorithm to identify the black spots.

Decision Tree (DT) is a supervised learning classification technique that uses a set of predictive features to predict the class variable based on a training set of observations. The accuracy of the model is evaluated using a separate test set of observations (Abellán et al., 2013). DT has the advantage of producing graphical output that is straightforward to interpret, and it can identify significant variables in the model quickly (Abdullah and Sipos, 2022). The research was conducted in Budapest, the capital city of Hungary, with a population of approximately 1.7 million and an area of 525 $km^2$ consisting of 23 districts (Jaber and Csonka, 2023). Accident data from 2019 to 2021 were utilized to investigate the current accident status and the relationships between the corresponding attributes. The dataset included accident locations across the city, severity of injuries, the number of fatalities, and the date and time of the accidents, which were used for spatial analysis of road traffic accidents.

## 3 Results

### 3.1 Road accidents

The study examined the proportion of traffic accidents that occurred over a period of three consecutive years (2019-2021) and categorized them based on peak and off-peak

hours. The total number of accidents recorded during this period was 8537, which was analyzed and presented graphically. According to Fig. 1, the highest percentage of accidents occurred in 2019, accounting for approximately 40% of the total accidents, whereas the lowest percentage of accidents was recorded in 2020, at around 30%.

Additionally, it was found that nearly 58% of the accidents took place during off-peak hours, while the remaining accidents occurred during peak hours, as depicted in Fig. 2.

In Fig. 3, the proportions of human fatalities, severely injured individuals, and slightly injured individuals involved in accidents are presented. It was observed that the majority of individuals involved in accidents suffered

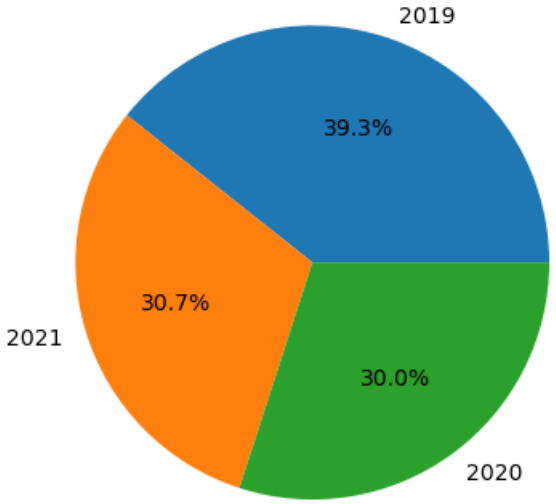


**Fig. 1** Percentage of accidents in the years

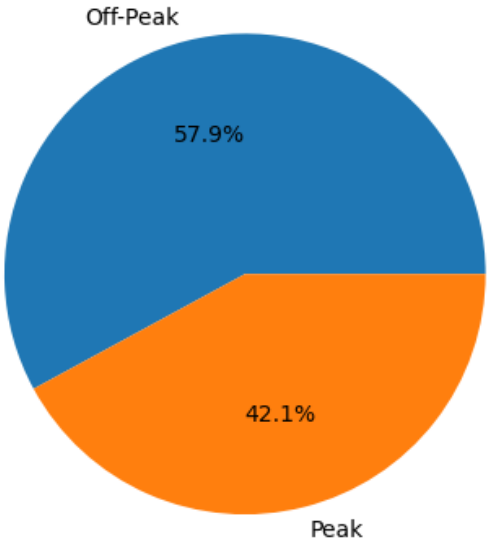


**Fig. 2** Accidents based on peak and off-peak

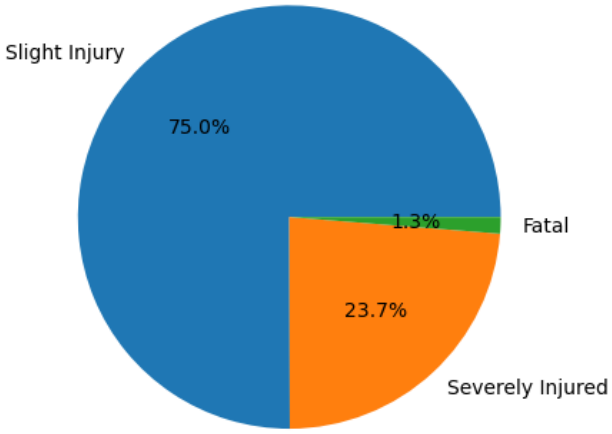


**Fig. 3** Percentage of different type of accidents

only slight injuries, constituting 75% of the total percentage, followed by individuals who were severely injured. However, the fatality rate was found to be relatively low, accounting for only 1.3% of the total accidents.

In this study, it was observed that traffic accidents occurred at various times of the day and on different days of the week, and on different lanes of the roads. The distribution of accidents with respect to the lanes of the roads is presented in Fig. 4, which indicates that around 78% of the accidents occurred on the right lanes while a lower percentage of accidents occurred on the left lanes of the roads.

Moreover, Fig. 5 illustrates the distribution of accidents with respect to weekdays and weekends, revealing that nearly a quarter of the accidents took place on weekends.

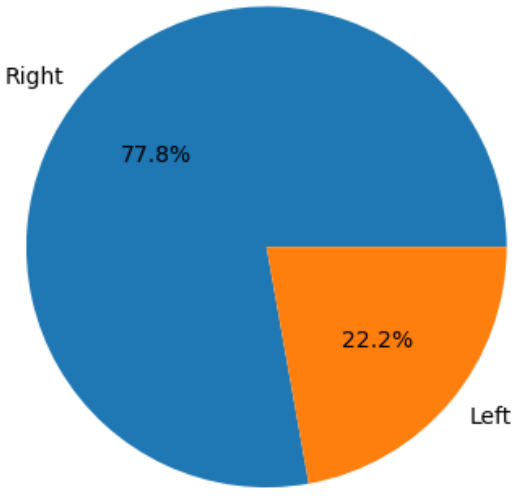


**Fig. 4** Accidents on different road lanes

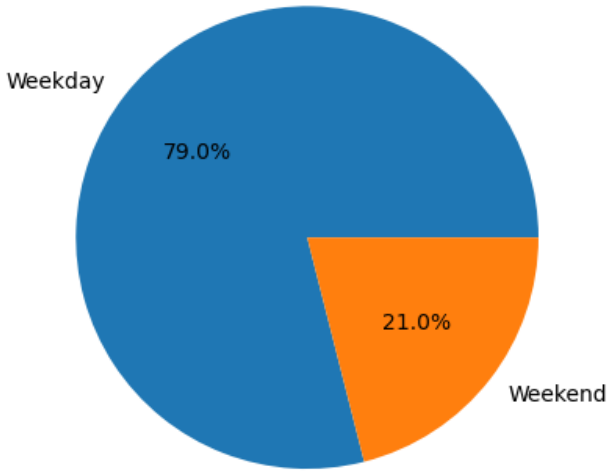


**Fig. 5** Accidents on weekdays and weekends

### 3.2 Spatial analysis

In Section 3.2, the spatial analysis was conducted using Quantum GIS software, which was equipped with the library to display the location of traffic accidents that occurred during peak and off-peak hours. To achieve this, the program was used to upload traffic accident locations in the form of georeferenced numbers that comprised coordinating longitudes and latitudes. Fig. 6 presents the spatial map of all traffic accidents that occurred in the city, while Fig. 7 displays the heat map package that highlighted the hotspots on the city map, indicating the areas where accidents occurred more frequently.

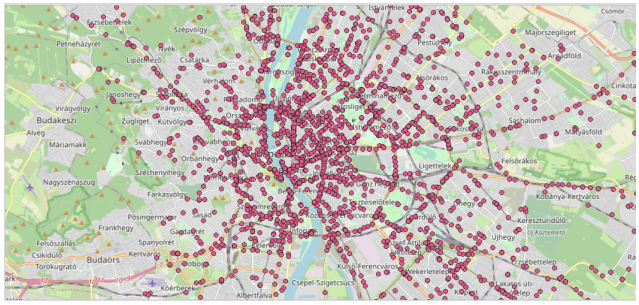

**Fig. 6** City map of Budapest with the locations of accidents

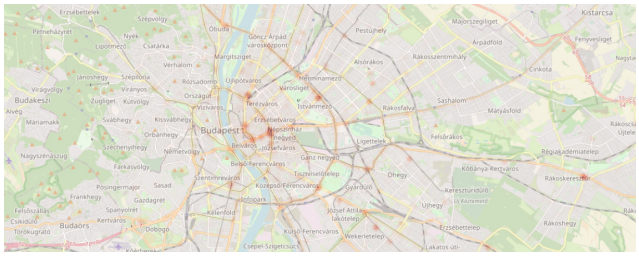

**Fig. 7** Heat map indicating the hot spots of the crashes in Budapest

### 3.3 The Decision Tree (DT) method

In Fig. 8, the DT method's Python output is displayed. Table 1 provides a list of the variables used for analysis. The decision tree shows the key factors contributing to both serious and minor injuries. The tree's hierarchical structure displays the most significant factors at the top and the least significant ones at the bottom. The DT method produced a 76% accuracy rate for the analysis conducted in this study.

## 4 Discussion

This paper explores the spatial analysis and the contributing factors of traffic accidents in a city. The study identified several traffic accident hotspots, known as "black spots", where at least two accidents occurred. The dataset analyzed in the study included 6405 minor injuries and 2025 serious injuries. The study tested three independent variables: road lanes (side), weekends, and "peak or off-peak" against the dependent variable, the accident severity. The study found that there was a gap in the literature in testing these variables against the target variable. The DT method analysis showed that the "weekend" variable was the most significant factor in the classification, as it was at the top of the decision tree. The dataset contained 8430 observations after removing fatal accidents, which accounted for only 1.3% of the total accidents over three years. Table 1 shows that number 0 represents weekdays, and number 1 represents weekends. The branches to the left of the tree represent accidents that occurred on weekdays, while the branches to the right represent accidents that occurred only on weekends.

The less significant variables in this study were peak times and road lanes. The branches to the right of the tree depicted that accidents occurring on weekends were more likely to happen during off-peak hours (over 65%) and on the right lane side of the road (over 80%). On the other hand, the branches to the left showed accidents that occurred on weekdays, with 23% of accidents taking place on the left side of the road and over 44% of them happening during peak hours. Additionally, more than half of the accidents during weekdays occurred during off-peak hours. The variable "peak or off-peak" was the second most influential factor for accidents that occurred on weekends (branches to the right), while the road lane was the third most influential factor.

The distribution of accidents throughout the city was analyzed and presented in Fig. 6, indicating that they occurred on different types of roads in various locations.

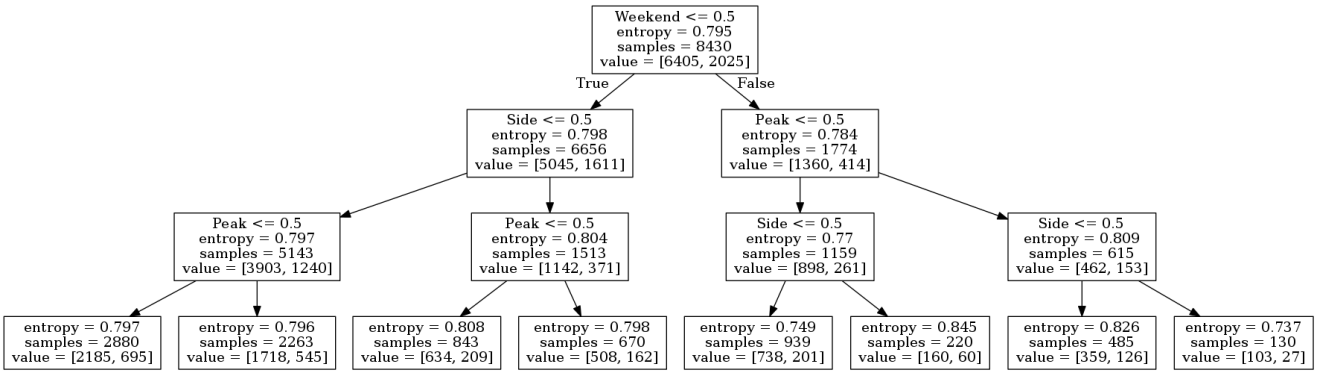


**Fig. 8** Python output showing the graph of DT method

**Table 1** The variables names and their numerical description

| Variables | Description |
|---|---|
| Side | Left-lane accidents = 1; Right-lane accidents = 0 |
| Weekend | Weekend = 1; Weekday = 0 |
| Peak | Peak = 1; Off-Peak = 0 |
| Target Variable | Description |
| Seriously_Injured_S | Seriously Injured = 1; Slightly-Injured = 0 |

However, the central area and the main administrative roads in the study area were identified as the most concentrated black spots. Fig. 7 illustrates the locations of accidents based on their hotspot level.

## 5 Conclusion

The study suggests that raising awareness is essential for ensuring traffic safety in the city, particularly during non-peak hours, which account for approximately 60% of all accidents. The analysis showed that around 35% of accidents occurring on weekends happened during peak hours, resulting in both slight and serious injuries, and the majority of crashes occurred on the right-side lanes, with 25% of them resulting in severe injuries. The study's model demonstrated 76% accuracy, as validated through the DT method. Using QGIS analysis, the study identified the city center as the location with the highest number of car accidents, with severe accidents being more common on administrative and arterial roads.

## Acknowledgement

The project presented in this article is supported by the Faculty of Transportation Engineering and Vehicle Engineering of the Budapest University of Technology and Economics, and the Department of Transport Technology and Economics supported this research.

## References

Abdullah, P., Sipos, T. (2022) "Drivers' Behavior and Traffic Accident Analysis Using Decision Tree Method", Sustainability, 14(18), 11339. https://doi.org/10.3390/su141811339

Abellán, J., López, G., de Oña, J. (2013) "Analysis of traffic accident severity using Decision Rules via Decision Trees", Expert Systems with Applications, 40(15), pp. 6047–6054. https://doi.org/10.1016/j.eswa.2013.05.027

Bordarie, J. (2019) "Predicting intentions to comply with speed limits using a 'decision tree' applied to an extended version of the theory of planned behaviour", Transportation Research Part F: Traffic Psychology and Behaviour, 63, pp. 174–185. https://doi.org/10.1016/j.trf.2019.04.005

Bucsuházy, K., Matuchová, E., Zůvala, R., Moravcová, P., Kostíková, M., Mikulec, R. (2020) "Human factors contributing to the road traffic accident occurrence", Transportation Research Procedia, 45, pp. 555–561. https://doi.org/10.1016/j.trpro.2020.03.057

Clarke, D. D., Forsyth, R., Wright, R. (1998) "Machine learning in road accident research: Decision trees describing road accidents during cross-flow turns", Ergonomics, 41(7), pp. 1060–1079. https://doi.org/10.1080/001401398186603

Dereli, M. A., Erdogan, S. (2017) "A new model for determining the traffic accident black spots using GIS-aided spatial statistical methods", Transportation Research Part A: Policy and Practice, 103, pp. 106–117. https://doi.org/10.1016/j.tra.2017.05.031

Fan, Z., Liu, C., Cai, D., Yue, S. (2019) "Research on black spot identification of safety in urban traffic accidents based on machine learning method", Safety Science, 118, pp. 607–616. https://doi.org/10.1016/j.ssci.2019.05.039

Figueira, A. C., Pitombo, C. S., de Oliveira, P. T. M. e S., Larocca, A. P. C. (2017) "Identification of rules induced through decision tree algorithm for detection of traffic accidents with victims: A study case from Brazil", Case Studies on Transport Policy, 5(2), pp. 200–207. https://doi.org/10.1016/j.cstp.2017.02.004

Jaber, A., Csonka, B. (2023) "How Do Land Use, Built Environment and Transportation Facilities Affect Bike-Sharing Trip Destinations?", Promet - Traffic&Transportation, 35(1), pp. 119–132. https://doi.org/10.7307/ptt.v35i1.67

Kaparias, I., Liu, P., Tsakarestos, A., Eden, N., Schmitz, P., Hoadley, S., Hauptmann, S. (2020) "Predictive road safety impact assessment of traffic management policies and measures", Case Studies on Transport Policy, 8(2), pp. 508–516. https://doi.org/10.1016/j.cstp.2019.11.004

Kurakina, E., Kravchenko, P., Brylev, I., Rajczyk, J. (2020) "Systemic approach to auditing road traffic accident black spots", Transportation Research Procedia, 50, pp. 330–336. https://doi.org/10.1016/j.trpro.2020.10.039

Li, K., Xu, H., Liu, X. (2022) "Analysis and visualization of accidents severity based on LightGBM-TPE", Chaos, Solitons & Fractals, 157, 111987. https://doi.org/10.1016/j.chaos.2022.111987

Pauer, G., Török, Á. (2021) "Binary integer modeling of the traffic flow optimization problem, in the case of an autonomous transportation system", Operations Research Letters, 49(1), pp. 136–143. https://doi.org/10.1016/j.orl.2020.12.004

Shatanawi, M., Abdelkhalek, F., Mészáros, F. (2020) "Urban Congestion Charging Acceptability: An International Comparative Study", Sustainability, 12(12), 5044. https://doi.org/10.3390/su12125044

Sobral, T., Galvão, T., Borges, J. (2019) "Visualization of Urban Mobility Data from Intelligent Transportation Systems", Sensors, 19(2), 332. https://doi.org/10.3390/s19020332

Sun, T.-J., Liu, S.-J., Xie, F.-K., Huang, X.-F., Tao, J.-X., Lu, Y.-L., Zhang, T.-X., Yu, A.-Y. (2021) "Influence of road types on road traffic accidents in northern Guizhou Province, China", Chinese Journal of Traumatology, 24(1), pp. 34–38. https://doi.org/10.1016/j.cjtee.2020.11.002

Taamneh, M. (2018) "Investigating the role of socio-economic factors in comprehension of traffic signs using decision tree algorithm", Journal of Safety Research, 66, pp. 121–129. https://doi.org/10.1016/j.jsr.2018.06.002

Török, Á. (2017) "Comparative Analysis between the Theories of Road Transport Safety and Emission", Transport, 32(2), pp. 192–197. https://doi.org/10.3846/16484142.2015.1062798

Turoń, K., Czech, P., Tóth, J. (2019) "Safety and Security Aspects in Shared Mobility Systems", Scientific Journal of Silesian University of Technology. Series Transport, 104, pp. 169–175. https://doi.org/10.20858/sjsutst.2019.104.15

Wang, X., Fan, T., Li, W., Yu, R., Bullock, D., Wu, B., Tremont, P. (2016) "Speed variation during peak and off-peak hours on urban arterials in Shanghai", Transportation Research Part C: Emerging Technologies, 67, pp. 84–94. https://doi.org/10.1016/j.trc.2016.02.005

Wisutwattanasak, P., Jomnonkwao, S., Se, C., Champahom, T., Ratanavaraha, V. (2023) "Correlated random parameters model with heterogeneity in means for analysis of factors affecting the perceived value of road accidents and travel time", Accident Analysis & Prevention, 183, 106992. https://doi.org/10.1016/j.aap.2023.106992

Zefreh, M. M., Torok, A. (2021) "Theoretical Comparison of the Effects of Different Traffic Conditions on Urban Road Environmental External Costs", Sustainability, 13(6), 3541. https://doi.org/10.3390/su13063541

Zeng, Q., Wang, F., Chen, T., Sze, N. N. (2023) "Incorporating real-time weather conditions into analyzing clearance time of freeway accidents: A grouped random parameters hazard-based duration model with time-varying covariates", Analytic Methods in Accident Research, 38, 100267. https://doi.org/10.1016/j.amar.2023.100267

Zheng, Z., Wang, Z., Zhu, L., Jiang, H. (2020) "Determinants of the congestion caused by a traffic accident in urban road networks", Accident Analysis & Prevention, 136, 105327. https://doi.org/10.1016/j.aap.2019.105327