Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 00 (2026) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

Road Safety and Simulation 2026 – RSS2026

Machine Learning–Based Analysis of Driving Behavior and Road Safety Using Connected Vehicle Data

Andreas Georgios Englezos[1]*, Armira Kontaxi[1], Eleonora Papadimitriou[1], George Yannis[1]

[1]*Department of Transportation Planning and Engineering, National Technical University of Athens, Athens, Greece*

Abstract

The aim of the present research is to investigate driving behavior through the analysis of data collected from connected vehicles. For the purposes of this research, variables related to speed, engine temperature, the anti-lock braking system (ABS), and other driving characteristics were examined. These data were collected over a three-month period to identify different route profiles in terms of driver behavior. The K-Means clustering method was applied to distinguish patterns of driving behavior, and the classification of trips into three clusters was found to provide satisfactory analytical results. Subsequently, the Random Forest algorithm was employed, using the anti-lock braking system as the dependent variable, to assess the importance of the independent variables. Results indicated that the variable with the highest importance value was the engine oil temperature. Finally, a Binary Logistic Regression was developed to examine the extent to which the independent variables affect the probability of ABS activation; revealing that engine activation is the most influential predictor. Overall, the findings highlight the effectiveness of combining connected vehicle data with machine learning techniques to support data-driven road safety analysis and decision-making.

Peer-review under responsibility of the scientific committee of the Road Safety and Simulation 2026 – RSS2026
Keywords: Driving Behavior, Data Analysis, Connected Vehicles, K-Means, Random Forest, Binary Logistic Regression

1. Introduction

Road safety remains a critical global challenge, strongly influenced by driver behavior and traffic conditions. In this context, the emergence of connected vehicle technologies provides new opportunities for monitoring, analysing, and ultimately improving driving behavior through real-time data collection and advanced analytics.

* Corresponding author. Tel.: +302107221575
E-mail address: andreas_englezos@mail.ntua.gr

Peer-review under responsibility of the scientific committee of the Road Safety and Simulation 2026 – RSS2026

A first approach to the term “Connected Vehicle” was presented by Uhlemann (2015), who defined it as a vehicle equipped with technologies that enable the exchange of data with other vehicles and with transport infrastructure. In this context, the operational framework of Cooperative Intelligent Transport Systems (C-ITS) and the core technologies on which they are based are also highlighted. Although connected vehicle technology is still at an early stage of development, existing studies provide valuable insights and contribute significantly to the extraction of useful conclusions.

More specifically, Zhou and Bridgeball (2020) noted that research in this field can generally be categorized into three main directions: (a) the analysis and evaluation of driving behavior through the processing of vehicle data, (b) the development of alternative road safety indicators, known as Surrogate Safety Measures (SSMs), which allow the assessment of road safety without relying exclusively on data from actual traffic accidents, and (c) the improvement of traffic signal control and intersection management. Overall, findings from the literature such as Sekadakis et al. 2021 and Oikonomou at al. (2023) indicate that the analysis of data generated by connected vehicles can play a substantial role in understanding driving behavior and, consequently, in improving road safety.

In addition, several studies have explored the application of advanced analytical techniques and machine learning methods to connected vehicle data. Commonly used algorithms include Fuzzy C-Means, K-Means, DBSCAN, Support Vector Machines (SVM), and Decision Trees, which are widely applied for clustering, classification, and pattern recognition in transportation research (Kontaxi et al. (2025); Petraki et al. (2025); Theofilatos et al. (2018)). Furthermore, studies by Mohammadnazar et al. (2021), Theodoraki et al. (2025) and Yang et al. (2022) focus on the classification of driving profiles using machine learning algorithms and datasets derived from connected vehicles. Similarly, Koliou et al. (2025) employed machine learning techniques to identify road segments in Athens with a high accident risk and to categorize them according to their safety characteristics. Furthermore, Ziakopoulos et al. (2025) analysed data collected from light commercial vehicles operating in London and demonstrated that harsh braking events can serve as effective indicators of road risk and surrogate safety measures (SSMs).

2. Data Collection

In the present research, the data originate from sensors installed in connected vehicles. They were generated within the framework of a research and development (R&D) project undertaken by OSeven Telematics titled 07Connected. The main objective of this project was the development of a new platform, similar to the existing 07Platform, capable of analysing driving behavior through the processing of vehicle-generated data.

The dataset was collected from two vehicles that were rented for a period of three months and driven by three different employees of the company. In total, 262 trips were recorded across both vehicles (152 from Vehicle 1 and 110 from Vehicle 2), comprising 420,855 observations. The analysis was conducted at the trip level, treating each trip as the unit of analysis. In general, the vehicles provided multiple streams of information, sufficient for the implementation of event detection algorithms. Beyond the execution of such algorithms, these data can also be utilized for a variety of analytical purposes related to driving behavior and vehicle operation.

The most important categories of collected data include:

- Geolocation: latitude, longitude, and heading, obtained from GPS and recorded every 6 seconds.
- Vehicle speed: measured by the vehicle’s own sensors at a frequency of 1 Hz (once per second).
- Odometer readings: distance travelled, recorded every 1 minute.
- Engine status: indicates whether the engine is running.
- Fuel data: consumption and levels provided by the vehicle.
- Collision indicators: front, rear, and side impact detection, with the type of collision; in this dataset all values were marked as repairs_not_needed.
- Driver-assistance systems: Lane Keep Assist (active/inactive for left and right lanes), parking assist sensors, Cruise Control, and the Anti-lock Braking System (ABS).
- Mechanical indicators: engine oil temperature and other operational parameters.

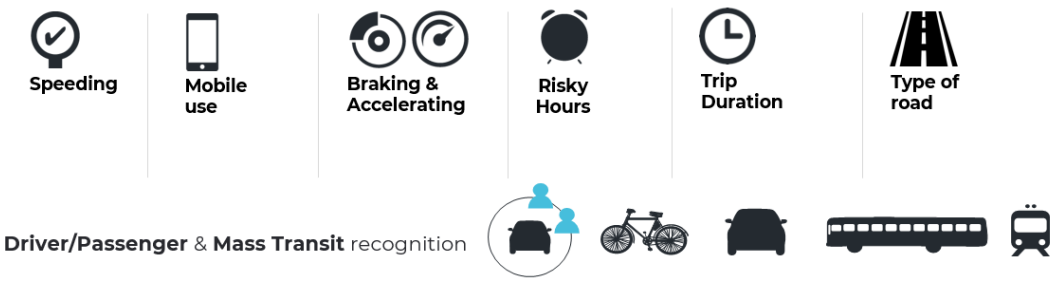

Fig. 1. The OSeven data flow system.

Overall, these data provide a comprehensive representation of vehicle dynamics, driver behavior, and the interaction between the driver and the vehicle's assistance systems. After completing the preprocessing steps and running the analysis code, exploratory visualizations were generated to provide an initial understanding of the dataset and driving behavior. Descriptive statistics were calculated for the main variables, and the term VIN refers to the Vehicle Identification Number of each vehicle.

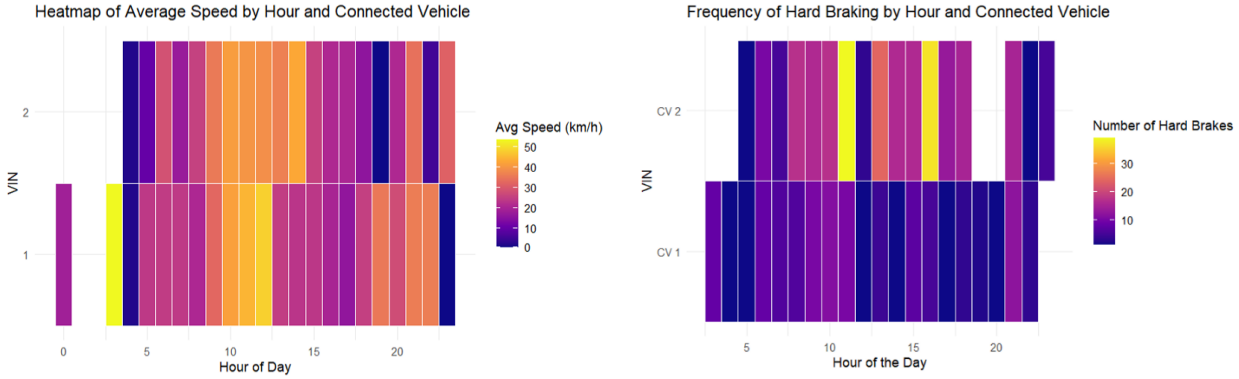

Fig. 2. Heatmap of the average speed of connected vehicles by hour of the day (left) and hourly frequency of harsh braking events per vehicle (right).

The visualizations provide valuable and insightful information regarding the driving behavior of the two vehicles. Specifically, Vehicle 1 primarily operates within a relatively narrow speed range of 20 to 65 km/h, with only occasional instances where speeds exceed 100 km/h. In contrast, Vehicle 2 demonstrates a much broader speed distribution, typically ranging from 20 to 100 km/h, while at certain times reaching higher peaks between 120 and 150 km/h .A closer examination of temporal patterns further reveals that average speeds tend to decrease during peak traffic hours, specifically between 07:00–09:00 in the morning and 15:00–19:00 in the afternoon, likely reflecting the influence of road congestion and traffic conditions. Conversely, higher speeds are generally observed during off-peak periods, namely between 10:00–13:00 and 21:00–24:00, indicating smoother and less restricted driving conditions.

Regarding speed variability, both vehicles exhibit notable fluctuations. Vehicle 1 shows a standard deviation of approximately 18 km/h, while Vehicle 2 demonstrates a slightly lower standard deviation of around 15 km/h. These values suggest frequent acceleration and deceleration events for both drivers, possibly influenced by traffic flow or road conditions. Additionally, the data indicate that harsh braking events occur more frequently for Vehicle 2, with a pronounced peak around 06:00 in the morning, and elevated frequencies also observed at 11:00 and 16:00. This pattern further highlights periods of more dynamic or demanding driving behavior throughout the day. Overall, these findings offer an initial yet comprehensive understanding of vehicle dynamics, driver behavior, and the impact of traffic conditions.

They also provide a solid foundation for more detailed analyses, which could uncover additional patterns or trends in driving behavior.

3. Results and Discussion

Prior to analysis, missing values were removed using listwise deletion. For the analysis of the data at the trip level, clustering was considered an appropriate initial approach in order to identify travel profiles based on driving behavior. After several trials, the K-Means algorithm was selected as the most suitable model for this purpose. Subsequently, in order to further investigate driving behavior, the analysis focused on identifying which variables, derived from connected vehicle data, influence the activation of the Anti-lock Braking System (ABS) and to what extent. For this purpose, the Random Forest machine learning method was employed to assess the importance of the variables, followed by the application of Binary Logistic Regression to further examine and model their effect.

Initially, in the present study, trip classification was performed for each vehicle using the K-Means algorithm, based on key features such as average and maximum speed, speed variability, Anti-lock Braking System (ABS) activation, and harsh braking events. Furthermore, using the Elbow and Silhouette methods, it was determined that classifying trips into three clusters (K = 3) provides a satisfactory analysis in the visualizations and their evaluation. Trip-level features including mean speed, maximum speed, speed standard deviation, average engine oil temperature, ABS activations per 100 km, and harsh braking events per 100 km were derived and standardized using z-score normalization. The K-Means algorithm was applied with 25 random initializations (nstart = 25) to ensure stable results.

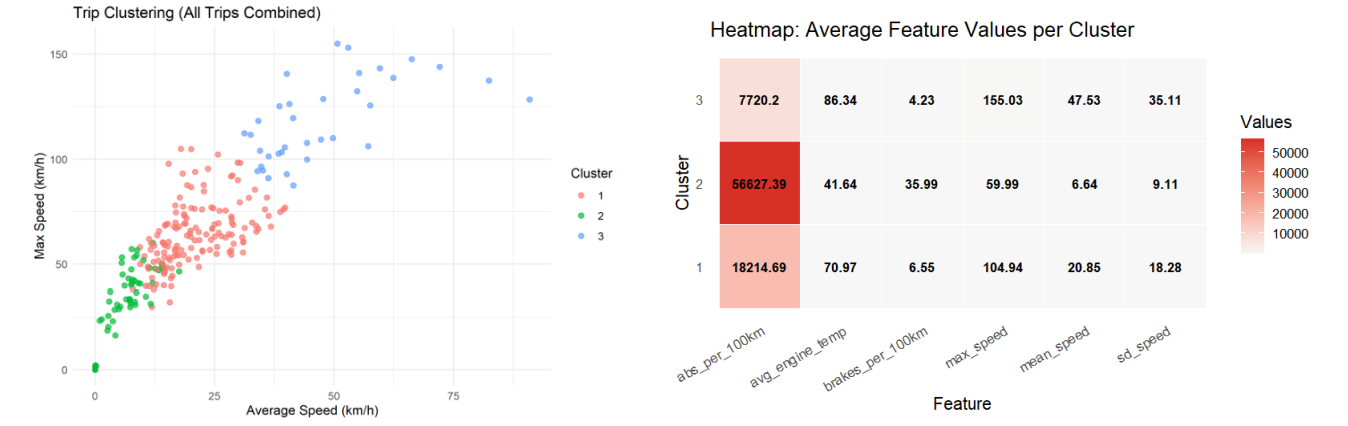

Fig. 3. Trip classification based on average and maximum speed (left) and heatmap of mean feature values per cluster (right).

Regarding the K-means plot, presented on Figure 3 (left), Cluster 1 (red) appears with higher frequency compared to the other two clusters. This cluster, characterized by intermediate speed values and moderate variability, represents a typical driving profile. Cluster 2 exhibits low speeds with minimal fluctuations, reflecting a cautious driving profile. Finally, Cluster 3 is distinguished by high average and maximum speeds, corresponding to an aggressive driving profile. These observations are further supported by the developed heatmap, (Figure 3 (right)), which presents the mean values of features per cluster. Specifically, Cluster 2, representing conservative driving, shows the highest activation levels of the Anti-lock Braking System (ABS), followed by Clusters 1 and 3.

Subsequently, the Random Forest algorithm was employed in the present study to evaluate the importance of the variables. The activation of the Anti-lock Braking System (ABS) was selected as the dependent variable, as ABS activation serves as a proxy for safety-critical braking events, consistent with the established use of harsh braking as a Surrogate Safety Measure (SSM) in the literature. Harsh braking events are generated by drivers as reactions to potentially hazardous situations and are widely used as road safety indicators in naturalistic driving studies (Ziakopoulos et al., 2022; Ziakopoulos et al., 2025). It is acknowledged, however, that ABS activation may also be influenced by external factors such as road conditions and vehicle dynamics, which represents a limitation of this approach. It should be noted that this is a binary variable, taking values of “active” or “inactive”. The dataset was split into an 80% training set and a 20% test

set. Given the observation-level nature of the dataset, which comprises 420,855 records across 262 trips, the split was performed at the observation level to ensure sufficient representation of both classes in training and testing. The model was trained using 300 trees, permutation-based variable importance, and inverse class-frequency weights to address class imbalance in the ABS variable. It is acknowledged that future studies with larger datasets could benefit from trip-level splitting to further assess model generalizability. The model achieved a high accuracy of 98%, indicating satisfactory overall performance. Additionally, with a sensitivity of 98.5%, the model effectively identifies instances when the Anti-lock Braking System (ABS) is activated. Finally, after a series of decision threshold tests, the specificity reached moderate to high values (65.4%), meaning that the model successfully recognizes a substantial portion of the non-activation events of the ABS. The variable importance diagram is presented below:

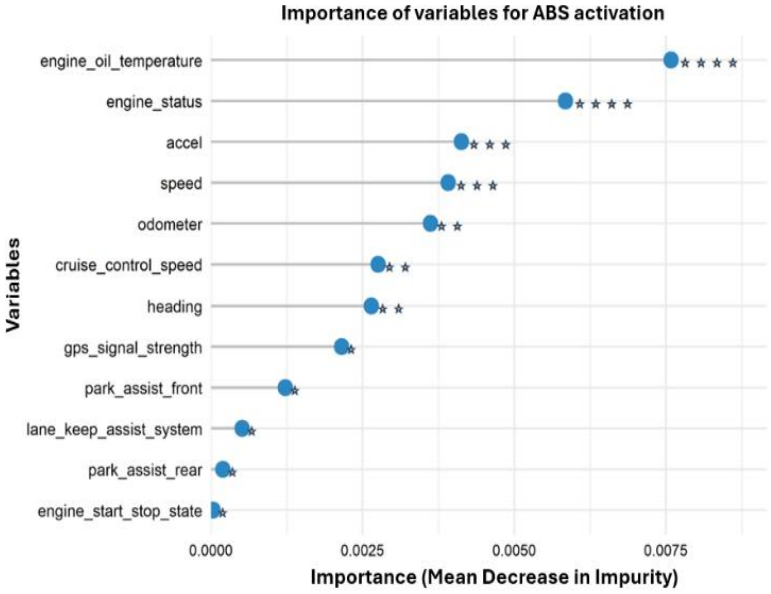

Fig. 4. Variable importance diagram – Random Forest.

The importance feature results indicate that the variables related to engine oil temperature, engine status, acceleration, and speed contribute most significantly, in that order, to the prediction of the dependent variable. Following these, variables with smaller yet meaningful influence include the odometer, cruise control speed, and vehicle heading. Finally, variables such as front and rear parking assistance and the lane-keeping system showed minimal impact on the model.

As noted above, the Anti-lock Braking System (ABS) variable is binary, taking values of "active" or "inactive". Thus, the Binary Logistic Regression method was selected for further analysis. The primary objective is to predict the probability of the dependent variable Y taking the value 1, based on the values of the independent variables X. The utility function (U) for the probability (P) is presented below:

$$Y_i = \beta_O + \beta_1 * X_{1i} + \beta_2 * X_{2i} + \cdots . + \beta_v * X_{vi} + e_i \quad (1)$$

$$P = \frac{e^U}{e^U + 1} \quad (2)$$

Table 1 presents the key variables along with their estimated coefficients, p-values, and Odds Ratios. A significance level of 95% was applied, with variables having p-values greater than 0.05 considered statistically insignificant. The Odds Ratio, calculated as e^{estimate}, indicates the change in odds for a one-unit increase in the independent variable. Multicollinearity was assessed using the VIF, with values ranging from 1 to 2, suggesting no multicollinearity concerns. Model performance was further evaluated using Pseudo-R^2 metrics, with McFadden $R^2 = 0.154$ and Cox & Snell $R^2 = 0.163$, indicating satisfactory model fit.

Table 1. Overall results of the Binary Logistic Regression model.

Independent Variables	estimate	p.value	Odds Ratio	Importance	VIF
(Intercept)	2,526	<0,001	12,499	<0,001	-
engine_oil_temperature	0,001	0,204	1,001	-	1,960
engine_status (on)	1,300	<0,001	3,671	<0,001	2,033
accel	-0,049	0,199	0,952	-	1,005
speed	0,056	<0,001	1,058	<0,001	1,183
odometer	0,000	0,078	1,000	-	1,006
cruise_control_speed	0,002	0,007	1,002	0,007	1,008
heading_sin	-0,076	<0,001	0,927	<0,001	1,003
heading_cos	-0,101	<0,001	0,904	<0,001	1,009
Pseudo-R^2	McFadden=0,154		Cox & Snell = 0,163		

From the analysis of the above table, the following conclusions can be drawn:

- The intercept of the function is positive (2.526), indicating that even when all variables take the value of 0, there remains a probability of ABS activation. This suggests that ABS engagement depends not only on the measured variables but also on unobserved factors, such as weather conditions during data collection or minor events occurring during trips.
- Engine activation is associated with a 3.7 times higher probability of ABS engagement. This may suggest that when the engine is running, braking is assisted and all sensors and system components operate fully. Modern connected vehicles may also deactivate the engine in specific situations, such as downhill driving, to save energy, which can lead to high speeds without ABS activation. However, this hypothesis could not be validated in the present study due to the absence of slope or geographic data.
- An increase of 1 km/h in vehicle speed is associated with a 5.8% increase in the probability of ABS activation. Higher speeds generate greater kinetic energy, leading to longer and stronger braking, which can trigger wheel lock prevention. The system may also respond during short braking events if the vehicle is moving at very high speeds.
- The influence of cruise control speed is minimal. A 1 km/h increase in cruise control speed corresponds to only a 0.2% increase in ABS activation probability, as cruise control maintains smooth, controlled driving and prevents abrupt braking.
- The heading_cos and heading_sin variables have negative coefficients, meaning that increases in these variables reduce the probability of ABS activation by 7.3% and 9.6%, respectively. This likely reflects road morphology or road types, such as uphill sections or long straight segments, where front-wheel braking demand is reduced, lowering the likelihood of ABS engagement. Nonetheless, this interpretation remains speculative in the absence of road infrastructure data.

4. Conclusion

The present study demonstrates the significant potential of connected vehicle data combined with machine learning techniques to enhance the understanding of driving behavior and its implications for road safety. Through the application of K-Means clustering, distinct driving profiles were identified, highlighting variations between cautious, typical, and aggressive driving patterns. Furthermore, the use of Random Forest and Binary Logistic Regression provided valuable insights into the factors influencing ABS activation, with engine-related variables and speed emerging as the most critical predictors.

Building on the findings and overall conclusions of the present study, several recommendations can be proposed to enhance road safety. Firstly, improving existing road infrastructure can strengthen traffic safety while also enhancing the quality of data collected in similar studies, as sensor readings would be less affected by road irregularities or vibrations. Secondly, raising awareness among government authorities and educational institutions is crucial. Individuals should develop an understanding of driver responsibility and road safety from an early age, with educational activities integrated into school programs to highlight the risks of violating traffic regulations.

Furthermore, studies that identify aggressive driving behaviors should be leveraged to provide guidance and interventions for improving driver behavior. Ensuring more frequent and stringent traffic enforcement can also help reduce violations, thereby creating a safer environment not only for drivers but also for other road users, including pedestrians and micro-mobility users. Finally, research focusing on driver behavior with the goal of enhancing road safety should be conducted more regularly to inform policy and practice.

For future work in this study, several avenues are suggested. Although the analysis was based on 262 trips from two vehicles, which limits the generalizability of the findings, future work should expand the dataset to include more vehicles and drivers across diverse road and traffic conditions. Additionally, similar studies could explore traffic conditions across different road types or investigate alternative variables using different machine learning algorithms. Incorporating geographic information into the analysis, such as whether a vehicle accelerates due to downhill slopes, could also provide deeper insights into driving behavior and system activation patterns.

Acknowledgements

The authors would like to thank OSeven Telematics for providing all necessary data exploited to accomplish this study.

References

Bagatelas, A., 2020. Logistic regression for rare events. Master's thesis, University of Piraeus.

Koliou, P., Peithis S., Yannis G. 2025. Data-driven urban road safety classification integrating telematics, machine learning, and spatial analysis. Proceedings of Road Safety on Five Continents (RS5C), 3–5 September 2025, Leeds, UK.

Kontaxi, A., Aivaliotis, A., Sideris, H., Oikonomopoulos, D., Yannis, G., 2025. Driver profiling through incentive-based cluster analysis in a naturalistic driving study. Proceedings of the 12th International Congress on Transportation Research, 16–18 October 2025, Thessaloniki, Greece.

Mohammadnazar, A., Arvin, R., Khattak, A.J., 2021. Classifying travelers' driving style using basic safety messages generated by connected vehicles: Application of unsupervised machine learning. Transportation Research Part C: Emerging Technologies 122, 102917.

Noble, W.S., 2006. What is a support vector machine? Nature Biotechnology 24.12, 1565–1567.

Oikonomou, M.G., Ziakopoulos, A., Chaudhry, A., Thomas, P., Yannis, G., 2023. From conflicts to crashes: Simulating macroscopic connected and automated driving vehicle safety. Accident Analysis & Prevention 187, 107087.

Petraki, V., Nikolaou, D., Yannis, G., 2025. Exploring safe and eco driving behavior through large-scale data using unsupervised learning. Proceedings of the 12th International Congress on Transportation Research, 16–18 October 2025, Thessaloniki, Greece.

Salman, H.A., Kalakech, A., Steiti, A., 2024. Random forest algorithm overview. Babylonian Journal of Machine Learning 2024, 69–79.

Sekadakis, M., Katrakazas, C., Santuccio, E., Mörtl, P., Yannis, G., 2021. Key performance indicators for safe fluid interactions within automated vehicles. Proceedings of the 10th International Congress on Transportation Research, September 2021, Rhodes, Greece.

Theodoraki, E.M., Garefalakis, T., Michelaraki, E., Yannis, G., 2025. Hybrid modelling for risky driving behavior classification: Insights from naturalistic driving study. Proceedings of the International Symposium Navigating the Future of Traffic Management, Athens, 29 June – 3 July 2025.

Theofilatos, A., Yannis, G., Antoniou, C., Chaziris, A., Sermpis, D., 2018. Time series and support vector machines to predict powered-two-wheeler accident risk and accident type propensity: A combined approach. Journal of Transportation Safety & Security 10.5, 471–490.

Uhlemann, E., 2015. Introducing connected vehicles [connected vehicles]. IEEE Vehicular Technology Magazine 10.1, 23–31.

Yang, K., Al Haddad, C., Yannis, G., Antoniou, C., 2022. Classification and evaluation of driving behavior safety levels: A driving simulation study. IEEE Open Journal of Intelligent Transportation Systems 3, 111–125.

Zhou, Y., Bridgelall, R., 2020. Review of usage of real-world connected vehicle data. Transportation Research Record 2674.10, 939–950.

Ziakopoulos, A., Karahlis, N., Yannis, G., 2025. Leveraging naturalistic connected LCV data for spatial surrogate safety measure applications. 16th ITS European Congress, Seville, Spain, 19-21 May 2025

Ziakopoulos, A., Yannis, G., 2020. A review of spatial approaches in road safety. Accident Analysis & Prevention 135, 105323.

Ziakopoulos, A., Vlahogianni, E., Antoniou, C., & Yannis, G. (2022). Spatial predictions of harsh driving events using statistical and machine learning methods. *Safety science*, *150*, 105722.