Factors Influencing Speed Limit Violations on Athens Road Network

Daphne Kyprouli[1], Dimitrios Nikolaou[1*[0000-0002-3106-2584]], George Yannis[1[0000-0002-7017-0756]]

[1]Department of Transportation Planning and Engineering, School of Civil Engineering - National Technical University of Athens, Athens, Greece
*dnikolaou@mail.ntua.gr

Abstract. The objective of this paper is to examine the factors influencing speed limit violations across the entire road network of Athens. The study utilizes data collected by OSeven Telematics and OpenStreetMaps, which includes information such as road geometry indicators, safety measurements and driving behavior metrics. Statistical models and machine learning algorithms were developed for two scenarios aiming to predict speeding violations and understand the factors influencing them. The first scenario focused on determining the presence or absence of speeding violations, while the second analyzed speeding relative to a specific threshold. Overall, ten models were developed. Through these different models, it was possible to highlight the importance of specific factors and their ability to predict the probability of speed limit violations. The results demonstrated that speeding has a statistically significant relationship with various variables, and improving driving behavior will consequently lead to a reduction of road crashes. The main factors affecting the likelihood of speeding in the examined road sections are the number of trips, the road section length and the percentage of mobile phone use, while slopes presented the least impact.

Keywords: Road Safety, Speed Limit Violations, Prediction Models, Mobile Phone Usage, Driver Behavior.

1 Introduction and Background

Speeding is defined as driving beyond the legally established speed limits [1] and has been recognized as one of the five fatal factors that lead to road crashes, together with driving under the influence of alcohol, fatigue, distraction, and the non-use of safety measures. International literature shows that speed affects both the frequency and severity of crashes. A U-shaped relationship has been observed, according to which vehicles traveling slightly faster than the average speed of a road tend to have the lowest crash rates, while those driving significantly faster or slower exhibit higher crash rates [2]. This theory was reinforced by studies [3],[4] which demonstrate that a reduction in average speed results in a proportional reduction in crash probability, thus establishing a linear relationship between

speed and risk. Statistical data indicate that human error is the main contributing factor in crashes (67%), while fewer crashes are attributed to road conditions (29%) and vehicle defects (4%). Moreover, is higher in areas with higher speed limits, though risk increases diminish at very high limits [5]. Earlier studies also found that a change in mean speed by 1 km/h can alter crash rates by 2% on motorways, 4% on roads with a 50 km/h limit, and between 2.5% and 5.5% on urban and rural roads, respectively [6], [7].

The violation of speed limits is attributed to factors related to the driver, the vehicle, and the road environment. With respect to driver characteristics, age plays a decisive role as drivers aged 25-34 exceed limits about 1.5 times more often than older drivers, while those over 60 display the most cautious behavior. Men and higher-income individuals are slightly more likely to speed, and penalty points increase future violations [8]. Regarding vehicle characteristics, type and capabilities have a strong influence. Sports cars and high-performance vehicles are more often associated with speeding. Automatic transmissions and lack of driver assistance systems can encourage higher speeds [9].

The road environment also plays a crucial role. On motorways, drivers tend to drive faster, whereas on winding or steep roads, many perceive limits as overly conservative and exceed them, increasing risk. Speeding often involves small deviations (~10 km/h) [9]. In addition, low posted limits and insufficient enforcement are among the main reasons for violations, while psychological factors such as risk-taking and inattention also affect speeding [10]. Speeding therefore remains a major factor in crash causation, influenced by driver, vehicle, and road characteristics.

This study analyzes the main factors influencing speeding behavior within the road network of Athens. Using data from OSeven Telematics combined with OpenStreetMap information, the analysis incorporates key variables to identify patterns and determinants of speeding. The objective is to develop a tool that identifies contributors to speed limit violations and generates a map of the network visualizing the probability of speeding, enabling targeted interventions to improve safety.

2 Data Collection

The present study utilizes naturalistic driving data obtained from a smartphone application developed by OSeven Telematics (https://www.oseven.io). Data recording begins automatically when driving is detected and stops after idle periods longer than five minutes. The application records user behavior through accelerometer, gyroscope, magnetometer, and GPS sensors, capturing variables such as speed, heading, latitude, and longitude. Additionally, iOS and Android provide data on deviation rate, intensity and rolling, linear acceleration, and gravitational acceleration. Data are processed centrally using Machine Learning algorithms to retain only information relevant to road safety and eco-driving.

Following a multi-level analysis process, raw data are transformed into high-precision insights about driving behavior, including risk exposure metrics (e.g., peak-hour

driving, total trip duration) and behavioral indicators such as harsh acceleration, harsh braking, turning intensity, and overall aggressive driving patterns.

For this study, the dataset covers the Athens area. An initial analysis of the road network was conducted using OpenStreetMap (OSM) to extract geometric road characteristics. The dataset included 35,282 road segments with key variables, including morphological road indicators, safety and protective measures, driver behavior, municipal area road crash statistics and environmental footprint indicators. From these variables, critical data were selected for analysis. While a wider set was available, some lacked sufficient coverage or relevance and were excluded. The selected variables were categorized as dependent and independent and included the following:

- Geographical information: osm_id, road segment centroid longitude (centroidlon), road segment centroid latitude (centroidlat)
- Road geometry characteristics: road type (highway), road segment length (length), road segment linearity (efficiency), road segment slope (slope), road segment slope category (slope_class)
- Driver behavior (field observations): seatbelt use by passenger car drivers (PC_D_Seatbelt_Yes%), helmet use by PTW drivers (PTW_D_Helmet_Yes%), hand-held mobile phone use by passenger car drivers (PC_D_Phone_No%), speeding by passenger car drivers (PC_Speeding_No%)
- Road crash statistics in the examined municipality 2016-2020: road crashes (Crashes2016-2020), road fatalities (Fat2016-2020), killed and seriously injured (KSI2016-2020)
- Driver behavior (telematics): Duration of exceeding the speed limits [sec] (speeding_count), speeding [yes/no] (speeding_di), Duration of mobile phone use [sec] (mobile_usage_count), Number of harsh acceleration events [count (harsh_acc_count), Number of harsh braking events [count] (harsh_braking_count)
- Environmental indicators: (CO2_count), (CO_count), (HC_count), (NOx_count)
- Trip information (telematics data): Number of trips per segment [count] (trip_count)

As shown in Fig. 1, correlations and potential interactions among variables were examined prior to statistical modeling to ensure strong associations while minimizing multicollinearity. The correlation analysis highlights key relationships among the variables. Total crashes are strongly correlated with fatalities, while seatbelt and helmet use also show a positive association, reflecting compliance with safety measures. A negative correlation exists between killed and serious injuries (KSI) and the proportion of non-speeding drivers, underlining the importance of speed compliance. Speeding incidents are positively associated with trip count and road length, while harsh events are correlated with each other

and with speeding. Mobile phone use shows weaker correlations overall but appears linked to more aggressive driving behavior.

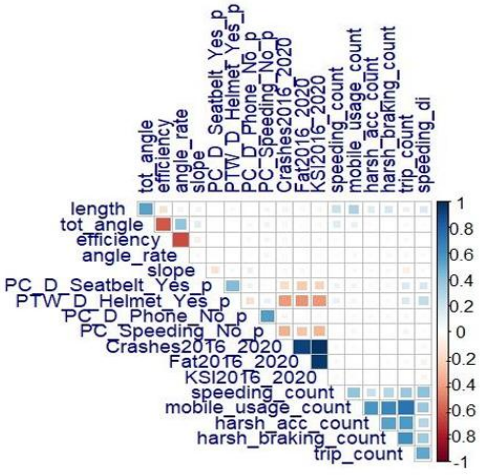

Fig. 1. Correlation matrix of the examined variables.

3 Methodology

After concluding the described data collection process, for statistical analysis, Binary Logistic Regression and Naïve Bayes Regression were employed. Additionally, Machine Learning algorithms were applied, including Random Forest, Decision Tree Regression, and Support Vector Machine (SVM) Regression. Each of these methods was selected for its suitability in binary classification, albeit with different theoretical foundations.

Binary Logistic Regression is a parametric model that estimates the probability of an outcome belonging to one of two classes by modeling the log-odds of the dependent variable as a linear combination of independent variables. Naïve Bayes is a probabilistic classifier based on Bayes' theorem, assuming conditional independence among predictors. Decision Trees classify observations through recursive partitioning, whereas Random Forest improves stability and predictive performance by aggregating multiple trees. SVM seeks to identify the hyperplane that maximizes the margin between classes, with kernel functions allowing for nonlinear separation when necessary.

To assess model performance, the dataset was divided into 80% for training and 20% for testing. Three main evaluation metrics were employed. Accuracy reflects the overall proportion of correct classifications but may be misleading in the presence of class imbalance. Sensitivity (recall) measures the proportion of true positives correctly identified and is especially important in contexts where missing a speeding event carries significant

consequences. Specificity captures the proportion of true negatives correctly identified, thus limiting false alarms.

The choice of the best-performing model depends on balancing these metrics with respect to the objectives of the study. While accuracy provides a global measure of predictive success, sensitivity and specificity highlight the trade-offs between detecting true speeding instances and avoiding misclassification of non-speeding cases. Consequently, model selection is guided not only by overall accuracy but also by the alignment of sensitivity and specificity with the practical requirements of road safety analysis.

4 Results

The classification models were evaluated under two scenarios:

- **Scenario 1:** Speeding indicator, where 0 = no speeding was recorded on the road segment, and 1 = speeding was recorded for at least one second.
- **Scenario 2:** Speeding indicator, where 0 = speeding duration on the road segment was less than 15.55 seconds, and 1 = speeding lasted at least 15.55 seconds. The value of 15.55 seconds corresponds to the mean speeding duration across all examined road segments.

In both scenarios, Binary Logistic Regression revealed consistent patterns regarding factors influencing speeding (Table 1). Road length and trip count showed positive and statistically significant effects, indicating that longer road segments and a higher number of trips are associated with greater speeding probability. Slope categories showed no meaningful effect. In Scenario 1, road efficiency also had a positive and significant impact, whereas mobile phone use showed a negative association: harsh events were positively related to speeding. In Scenario 2, efficiency lost statistical significance, while road type demonstrated strong significance in most categories. remained negative and significant, harsh braking retained limited strength, and harsh accelerations were not significant.

Table 1. Evaluation metrics of the developed models for the test dataset under two different scenarios.

	Scenario 1				
	Binary Logistic	**Random Forest**	**Decision Tree**	**SVM**	**Naive Bayes**
Accuracy	0.778	**0.784**	0.771	0.781	0.767
Sensitivity	0.355	**0.409**	0.330	0.319	0.323
Specificity	0.942	0.928	0.941	**0.958**	0.938
Scenario 2					

Accuracy	0.941	**0.945**	0.938	0.939	0.916
Sensitivity	0.355	0.392	0.265	0.275	**0.486**
Specificity	0.989	0.989	0.993	**0.994**	0.951

The evaluation of the developed models under both scenarios shows that all methods achieve high overall accuracy and very strong specificity, indicating their ability to correctly identify non-speeding cases. However, sensitivity remains comparatively low across models, reflecting the difficulty of detecting speeding events. In Scenario 1, Random Forest achieved the highest sensitivity, while in Scenario 2, Naïve Bayes achieved the highest sensitivity at the expense of lower overall accuracy and specificity. Overall, Random Forest demonstrates consistently strong performance across both scenarios, combining high accuracy with relatively better sensitivity.

In the Random Forest Analysis, variable importance was evaluated to identify the factors contributing most to speeding prediction (see Fig. 2). In both scenarios, trip count emerged as the most influential variable. In contrast, harsh events and slope class consistently recorded lower importance, suggesting a limited role in predicting speeding. The ranking of variables remained similar between the two scenarios, though in Scenario 2 the importance of mobile phone use increased slightly, while road length and efficiency contributed less compared to Scenario 1. Nevertheless, the differences were minor.

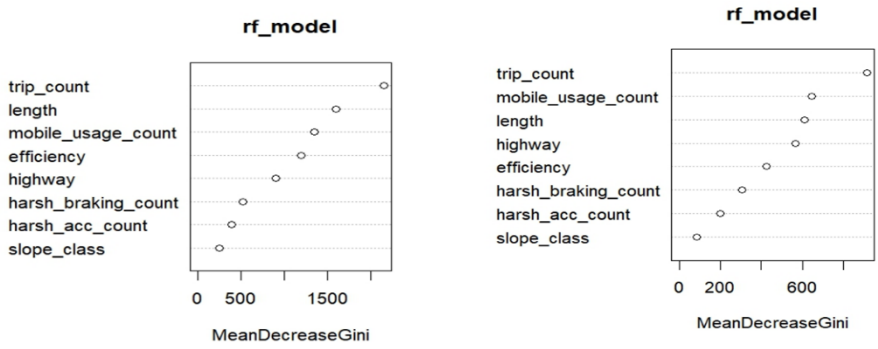

Fig. 2. Variable Importance Ranking – Random Forest model (Scenario 1: left, Scenario 2: right)

5 Map Visualization

As a final step, an interactive cartographic visualization of the results of the statistical model selected in the previous subsections was implemented. This visualization aims to geographically represent the probability of speeding violations across the entire road network of Athens. Initially, the spatial data (GeoJSON) were imported and processed, and subsequently converted into sf objects to enable spatial manipulation and visualization

using the leaflet library. For the mapping, Binary Logistic Regression combined with Scenario 1 was selected, as it was deemed to offer the optimal balance between interpretability, practical value, and comprehensive coverage of speeding events. This method estimates probabilities based on linear combinations of variables, and although other models performed better, the ease of integrating results into the map was prioritized. Comparing the two scenarios shows that Scenario 1 allows for the identification of all locations where any speeding event, regardless of duration, was recorded, whereas Scenario 2 significantly reduces the number of identified cases and may conceal critical risk areas. Consequently, Scenario 1 provides a more comprehensive picture of the spatial distribution of speeding across the road network, thereby supporting the design of targeted road safety measures.

The final map was then generated using a color scale ranging from green, indicating low probability, to red, indicating high probability, thus making it possible to identify the most hazardous road segments. Analysis of Fig. 3 shows that red areas are mainly observed on major central roads with multiple lanes and long segment lengths, suggesting that increased trip count and longer travel distances are the primary contributors to higher speeding probabilities in these segments. This approach renders the results accessible in a visual and comprehensible manner and can be utilized both for research purposes and for the formulation of targeted road safety policies.

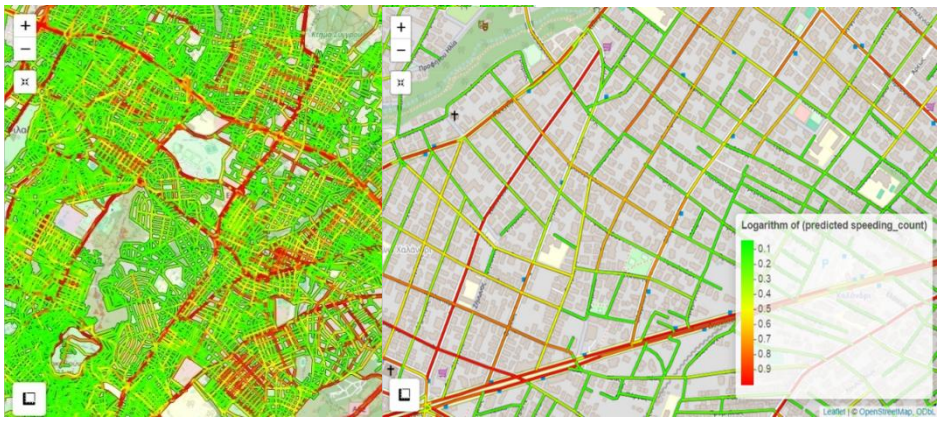

Fig. 3. Speeding Probabilities on the Road Network of Athens (Zoomed-out: left, Zoomed-in: right)

6 Conclusion

This study confirms that speeding is a critical determinant of road safety. Speeding is positively related to exposure indicators such as trip count and road length, as well as harsh driving behaviors. Interestingly, mobile phone use while driving exhibited a negative association with speeding, possibly reflecting compensatory cautiousness, although distraction still poses considerable risks. Among the developed models, Random Forest

consistently outperformed the others, identifying the most important predictors. From a policy perspective, the results highlight the need for targeted interventions to reduce speeding and promote safer driving. Measures such as adaptive speed management systems, telematics-based driver feedback, and awareness campaigns on the dangers of speeding and distracted driving could be particularly effective. The interactive map of speeding probabilities provides a practical tool for identifying high-risk segments and supporting evidence-based, location-specific safety measures.

Future research should therefore aim to expand the dataset to other regions and populations in order to capture geographical and social differences. Additionally, the relatively low sensitivity observed in the models may be related to potential class imbalance in the dataset or to the limited number of explanatory variables capable of adequately distinguishing speeding events. Model sensitivity could be further improved by incorporating additional contextual variables such as traffic density, weather conditions, time of day, and more detailed driver behavior indicators. Including such factors may enhance the model's ability to better capture complex driving environments and improve predictive performance.

References

1. Liu, C., Chen, C. L.: An analysis of speeding-related crashes: definitions and the effects of road environments. NHTSA Technical Report DOT HS 811 090 (2009).
2. Solomon, D. H.: Accidents on main rural highways: related to speed, driver, and vehicle. U.S. Department of Commerce, Bureau of Public Roads (1964). [Reprinted by FHWA, 1974]
3. Nilsson, G.: Traffic Safety Dimensions and the Power Model to Describe the Effect of Speed on Safety. Lund Institute of Technology, Bulletin 221 (2004).
4. Elvik, R.: Speed and Road Safety: Synthesis of Evidence from Evaluation Studies. Transportation Research Record 1908 (2013). https://doi.org/10.3141/1908-08
5. Davis, G. A., Davuluri, S. U., Pei, J.: Speed as a risk factor in serious run-off-road crashes: Bayesian case-control analysis with case speed uncertainty. Journal of Transportation and Statistics 9, 17-28 (2006).
6. Nilsson, G.: The Effects of Speed Limits on Traffic Accidents in Sweden. OECD Symposium on the Effects of Speed Limits (1982).
7. Taylor, M. C., Lynam, D. A., Baruya, A.: The Effects of Drivers' Speed on the Frequency of Road Accidents. TRL Report 421 (2000).
8. Perez, M. A., Sears, E., Valente, J. T., Huang, W., Sudweeks, J.: Factors modifying the likelihood of speeding behaviors based on naturalistic driving data. Accident Analysis & Prevention 159, 106267 (2021). https://doi.org/10.1016/j.aap.2021.106267
9. Liang, Z., Xiao, Y.: Analysis of factors influencing expressway speeding behavior in China. Plos One15(9), e0238359. (2020). https://doi.org/10.1371/journal.pone.0238359
10. Gabany, S. G., Plummer, P., Grigg, P.: Why drivers speed: The speeding perception inventory. Journal of Safety Research 28, 29-35 (1997). https://doi.org/10.1016/S0022-4375(96)00031-X