Factors contributing to safety-critical events in urban areas: A driving simulator study

Fotini Bardi[1], Christos Katrakazas[1]*, George Yannis[1]

[1]National Technical University of Athens, Department of Transportation Planning and Engineering,5 Heroon Polytechniou str., GR-15773, Athens, Greece

Extended Abstract

1 Introduction

As the frequency of road traffic casualties and injuries is increased worldwide, road safety has become a critical social problem. In the recent past, indicators of dangerous driving behavior, inattention and unexpected events have been the focus of researchers in road safety (e.g. [1]–[3]).

Concerning unexpected events, studies have been looking into the real-time identification for safety assessment (e.g. [3], [4]), but the breakdown of events in phases (i.e. pre-event, during the event and post-event) and the identification of the most significant factors per phase is yet to be realized. This is the motivation behind the current paper, which aims at determining the most significant factors for classifying safety- critical events in urban roads. For that purpose, this research exploits driving simulator data and 41 drivers. Two statistical models are developed, namely a binomial logistic regression and a random forest one, both of which aimed at predicting the occurrence of an event. Finally, factor analysis was applied with regards to data concerning one minute before the event, the duration of the event as well as the combination of these cases, in order to examine the existence of common factors in the group of independent variables.

2 Methodology

The identification of unexpected events is a binary classification problem, and therefore two binary classification models were developed in this paper (i.e. binomial logistic regression and random forests).

In order to evaluate the two classification models (i.e. binary logistic regression and random forests), the confusion matrix was utilized in both models, along with the coefficient signs and the z-test for statistical significance for the logistic regression model. In the confusion matrix, the occurrence of an event was defined as the positive class and the non-occurrence as the negative class, respectively. The performance metrics that were used and based on the confusion matrix were accuracy, precision, recall, specificity, f-measure and false alarm rate.

In order to identify relationships between variables before and during the occurrence of unexpected events, factor analysis [5] was used to to identify groups of inter-related variables and examine how they relate to each other.

3 Analysis and Results

3.1 Data collection

Data were exported from a large driving simulator dataset created for an experiment of a doctoral thesis [6] and concern driving indicators. Useful indicators were decided to be total distance traveled, vehicle deviation from the middle of the road, speed, distance from the vehicle in front, distance from the right and left lane, steering wheel position, time to collision with the vehicle in front, time to collision with all the obstacles, lateral and longitudinal

[1] * Corresponding author. Tel.: +302107721265;
*E-mail address:*ckatrakazas@mail.ntua.gr

acceleration. All of these indicators were chosen because they are crucial predictors of the motion and positioning of a vehicle and therefore seriously affect the probability of involvement in an unexpected incident. Each driver's age, gender, education, driving experience were also obtained from questionnaires. Traffic conditions as well as the existence or not of an unexpected event emerged from the different scenarios performed by the drivers in the simulator. From the above data, the UrbanControl database was created and used in the present research, which contains data on 41 drivers, men and women, from all ages and levels of education with different driving experience and only applies to urban areas.

3.2 Data Processing

From the above database (i.e. UrbanControl), three individual data tables were created using the R programming language. These include a table containing the data during events (DurEventU), a table containing data one minute before the event (PreEventU) and a total table containing the sum of the aforementioned data (EventU). The tables were separated based on the Event variable.

Subsequently, descriptive statistics were performed in DurEventU and PreEventU tables, and the correlation between the independent variables was determined. The variables which were highly correlated were:

- Distance from the vehicle in front with time to collision with the vehicle in front
- Distance from the left lane with distance from the right lane
- Age with Driving experience.

According to the above, the final data table (Model) was created by extracting the columns: event, speed, lateral acceleration, longitudinal acceleration, time headway with the vehicle in front, distance from the right lane, total distance travelled, vehicle deviation from the middle of the road, steering wheel position and driving experience from the EventU table. Both models (i.e. logistic regression and Random Forests) were developed using the Model table.

The Factor Analysis method was performed in the EventsU2, PreEventU2 and DurEventU2 tables. The EventsU2 table was created by isolating speed, lateral acceleration, longitudinal acceleration, time to collision with the vehicle in front, distance from the right lane, total distance traveled, vehicle deviation from the middle of the road and steering wheel position columns from the Model table. The PreEventU2 table was created by isolating the same columns from the Model table only when Event = 0, i.e. the data related to the non-existence of an event. The DurEventU2 table was created by isolating the same columns from the Model table only when Event = 1, i.e. the data related to the existence of the event. The development of the aforementioned tables is depicted in Figure 1, while the variable description is given in Table 1.

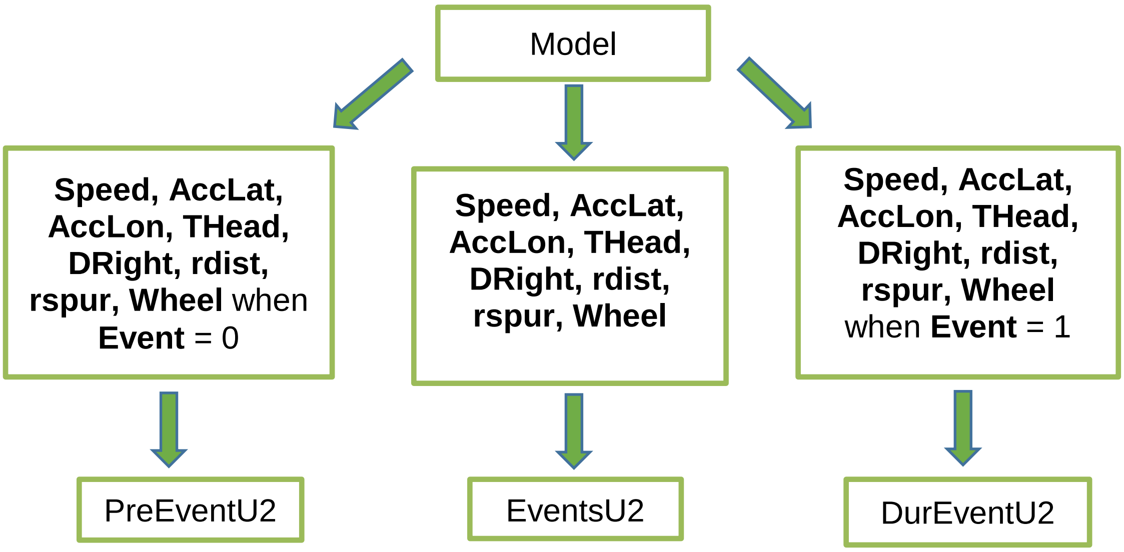
Model
Speed, AccLat, AccLon, THead, DRight, rdist, rspur, Wheel when Event = 0
Speed, AccLat, AccLon, THead, DRight, rdist, rspur, Wheel
Speed, AccLat, AccLon, THead, DRight, rdist, rspur, Wheel when Event = 1
PreEventU2
EventsU2
DurEventU2

Figure 1: Creation of PreEventU2, DurEventU2 and EventsU2 tables

Table 1: Variable description

Variables	Explanation
Speed	Speed (km/h)
AccLat	Lateral acceleration (m/s^2)
AccLon	Longitudinal acceleration (m/s^2)
HWay	Distance from the vehicle in front (m)
THead	Time to collision with the vehicle in front (s)
DLeft	Distance from the left lane (m)
DRight	Distance from the right lane (m)
rdist	Total distance traveled (m)
rspur	Vehicle deviation from the middle of the road (m)
Wheel	Steering wheel position (degrees)
Age	Age (years)
Driving experience	Driving experience (years)

3.3 Results

Two different independent variable settings were explored; Version A including all the independent variables in the EventsU2 dataset, and Version B, with the most important and statistically significant variables extracted from the Boruta algorithm [7]. The most important variables were speed, time headway, distance from the right side of the road, distance driven, steering wheel position and driving experience, as shown in Table 2.

Table 2: Variable importance

Variable	Boruta Importance
Wheel	228.78
rdist	177.71
speed	173.15
Dright	160.84
Driving Experience	143.86
Thead	129.7

Table 3: Models evaluation results

Performance metrics	Binomial logistic regression (Version A)(%)	Random Forest (Version A) (%)	Random Forest (Version B) (%)
Accuracy	79.26	87.17	81.19
Recall	13.44	65.56	53.51
Specificity	99.04	93.67	89.51
Precision	80.8	75.68	60.53
F-measure	23.05	70.26	56.81
False alarm rate	0.96	6.33	10.49

According to the results in Table 3, the **binomial logistic regression model** presents a very low recall index, so it fails to predict the occurence of an event, while it can predict satisfactorily the non-occurence of an event as observed by its high specificity. The overall accuracy, however is satisfactory. The false alarm rate is also quite satisfactory but the F-measure which expresses the harmonic means of accuracy and recall is very low.

The **random forest model** (Version A) can predict well the existence of events as well as the normal driving conditions as shown by the high specificity. The model also presents a high accuracy index and high precision. The possibility of incorrect classification of positive snapshots is also very low and the model can predict well both positive and negative instances as shown by the F-measure.

The random forest model (Version B) presents a marginally satisfactory recall index, along with high specificity. The overall accuracy is satisfactory but demonstrates a higher false alarm rate than version A.

From the above it can be concluded that the random forest model for Version A is the best in classifying the existence of an event, whereas the performance of Version B can be considered satisfactory.

Table 4: Factor analysis results

Table	Factor 1	Factor 2	Factor 3
PreEventU2 **(1 min before the event)**	DRight	Speed	rspur
DurEventU2 (during the event)	Speed	AccLat AccLon	
EventsU2 (before and during events)	AccLat	Speed	DRight

Finally, from the factor analysis results in Table 4, it was demonstrated that the situation one minute before the event can be described through the influence of the distance from the right borderline, the speed and the deviation of the vehicle from the middle of the road. The situation during the event can be expressed through the influence of speed and longitudinal and lateral acceleration while the whole event phenomenon is expressed through lateral acceleration, speed and distance from the right boundary line.

4 Conclusions

The present research determined critical factors for classifying safety-critical events in urban roads using driving simulator data. Two statistical analysis models, a binomial logistic regression and a random forest one, were used to predict the occurence of an event. Furthermore, factor analysis was applied to data one minute before the event, the duration of the event and the combination of the two aforementioned datasets.

Results of the developed models can be summarized as follows:

- The most important variables for identifying events are speed, total distance traveled, distance from the right lane, steering wheel position and time to collision with the vehicle in front.
- The random forest model provided the best classification results, compared to the binomial logistic regression model.
- Factor analysis validated that speed and acceleration along with lateral distances were found to be the most critical factors for identifying events in urban roads.

References

[1] P. Papantoniou, C. Antoniou, G. Yannis, and D. Pavlou, "Which factors affect accident probability at unexpected incidents ? A structural equation model approach," *J. Transp. Saf. Secur.*, vol. 0, no. 0, pp. 1–18, 2018, doi: 10.1080/19439962.2018.1447523.

[2] C. Katrakazas, M. Quddus, and W. H. Chen, "A simulation study of predicting real-time conflict-prone traffic conditions," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 10, pp. 3196–3207, 2018, doi: 10.1109/TITS.2017.2769158.

[3] C. Katrakazas, A. Theofilatos, M. A. Islam, E. Papadimitriou, L. Dimitriou, and C. Antoniou, "Prediction of rear-end conflict frequency using multiple-location traffic parameters," *Accid. Anal. Prev.*, vol. 152, no. December 2019, 2021, doi: 10.1016/j.aap.2021.106007.

[4] L. Zheng, T. Sayed, and F. Mannering, "Modeling traffic conflicts for use in road safety analysis: A review of analytic methods and future directions," *Anal. Methods Accid. Res.*, vol. 29, p. 100142, 2021, doi:

10.1016/j.amar.2020.100142.

[5] S. P. Washington, M. G. Karlaftis, and F. L. Mannering, *Statistical and Econometric Methods for Transportation Data Analysis*. CRC Press, 2010.

[6] D. Pavlou, "Traffic and safety behaviour of drivers with neurological diseases affecting cognitive functions," 2016.

[7] M. B. Kursa and W. R. Rudnicki, "Feature selection with the boruta package," *J. Stat. Softw.*, vol. 36, no. 11, pp. 1–13, 2010, doi: 10.18637/jss.v036.i11.