Identification of Critical Driving Parameters Affecting Speeding Using Data from Smartphones

Aristotelis Kokkinakis[1], Armira Kontaxi[2]*, Dimitrios Tselentis[3], George Yannis[4]

Department of Transportation Planning and Engineering – National Technical University of Athens, 5 Heroon Polytechniou Str., GR-15773 Athens, Greece

E-mail[1]: arkokkinakis@gmail.com
E-mail[2]: *corresponding author akontaxi@mail.ntua.gr
E-mail[3]: dtsel@central.ntua.gr
E-mail[4]: geyannis@central.ntua.gr

Abstract

The aim of this research is to identify the critical driving parameters that affect speeding using data from smartphones. To achieve this objective, data collected from sixty-eight (68) drivers who participated in a naturalistic driving experiment for fifteen (15) months, are analyzed. Through linear regression models, it was examined whether driving characteristics recorded by smartphone affect and can therefore predict the percentage of speeding time during driving. Four statistical linear regression models forecasting the percentage of speeding time during driving were developed: one overall model and three models for each road type (urban, rural, highway). The results revealed that six parameters namely, distance, high intensity harsh acceleration and braking events as well as cornering events, average deceleration and mobile usage, were found to be statistically significant in the regression models.

Keywords: *driving behavior, speeding, harsh braking, harsh acceleration, smartphone data, linear regression model, road safety*

Περίληψη

Στόχος της παρούσας έρευνας είναι ο προσδιορισμός των κρίσιμων παραμέτρων επιρροής της υπέρβασης των ορίων ταχύτητας με δεδομένα από έξυπνα κινητά τηλέφωνα. Για το σκοπό αυτό, αναλύονται στοιχεία που συλλέχθηκαν από εξήντα-οχτώ (68) οδηγούς οι οποίοι συμμετείχαν σε πείραμα οδήγησης σε πραγματικές συνθήκες για χρονικό διάστημα δεκαπέντε (15) μηνών. Μέσω μοντέλων γραμμικής παλινδρόμησης εξετάστηκε κατά πόσο χαρακτηριστικά της οδήγησης που καταγράφηκαν από αισθητήρες έξυπνων κινητών τηλεφώνων επηρεάζουν και επομένως μπορούν να προβλέψουν το ποσοστό υπέρβασης της ταχύτητας πάνω από τα όρια ταχύτητας. Αναπτύχθηκαν συνολικά τέσσερα μοντέλα πρόβλεψης υπέρβασης της ταχύτητας: ένα γενικό μοντέλο και τρία μοντέλα για κάθε τύπο οδικού δικτύου (αστικό, υπεραστικό, αυτοκινητόδρομο). Από την εφαρμογή των μοντέλων προέκυψε ότι οι έξι παράμετροι της απόστασης, των υψηλής εντάσεως απότομων επιταχύνσεων και επιβραδύνσεων, των απότομων στροφών, της μέσης επιβράδυνσης και της χρήσης του κινητού τηλεφώνου βρέθηκαν στατιστικά σημαντικές στα μοντέλα γραμμικής παλινδρόμησης.

Λέξεις κλειδιά: *οδηγική συμπεριφορά, υπέρβαση των ορίων ταχύτητας, απότομες επιβραδύνσεις, απότομες επιταχύνσεις, έξυπνα κινητά τηλέφωνα, μοντέλα γραμμικής παλινδρόμησης, οδική ασφάλεια*

1. Introduction

Road safety is a complicated scientific field of transport research due to the random nature of accident occurrence. Accidents impose serious problems to society in terms of human costs, economic costs, property damage costs and medical costs. According to World Health Organization (WHO), the total number of road fatalities worldwide continues to climb, reaching a high of 1.35 million in 2016. Regarding European Union, traffic accidents were the fifth cause of death in 2016, with roughly six people out of every 100,000 dying on the roads of the European Union because of road crashes. Consequently, understanding the various risk factors that cause road accidents is very crucial and has attracted great attention in the literature. Although there has been a very considerable research effort so far, there is still much to be investigated, especially in order to acquire a better knowledge of detailed pre-accident conditions in order to have a better proactive safety management in major roads of the transport network.

There is a significant number of risk factors identified in literature, which affect accident probability. The most important risk factors recognized in literature (WHO 2018, Elvik 2004) are human factors (speeding, distracted driving, driving under the influence of alcohol and other psychoactive substances etc.), unsafe road infrastructure, unsafe vehicles and inadequate law enforcement of traffic laws. Among them, human factors are likely to be the most crucial cause of road traffic fatalities and injuries every year and therefore the importance of studying how these factors can affect traffic risk is high.

Speeding is one of the most important human factors that influence road accident risk (WHO, 2018). Excess and inappropriate speed are responsible for a high proportion of the mortality and morbidity that result from road crashes. In high-income countries, speed contributes to about 30% of deaths on the road, while in some low-income and middle-income countries, speed is estimated to be the main contributory factor in about half of all road crashes. Controlling vehicle speed can prevent crashes happening and can reduce the impact when they do occur, lessening the severity of injuries sustained by the victims. However, it is of high importance to understand that the relationship between speed and road safety is a complex one; many physical and psychological factors play a role. In this paper, two such groups of studies are examined: studies aiming to define the factors that affect speeding and studies that have used in-vehicle recoding systems to investigate driving behavior.

The first studies that assessed the effects of speeding were case-control studies and date from the 1960s and 1970s in the United States (e.g. Solomon, 1964). At that time, the conclusion was that both drivers driving faster than the average speed on a road and drivers driving slower had a higher risk of getting involved in a crash. Other studies confirmed the higher crash risk of drivers driving above the average speed. In Australia, this conclusion was based on case-control studies (Kloeden et al. 1997, 2001, 2002). In Great Britain, a similar conclusion arose from a self-report study (Taylor et al., 2000). However, these recent studies did not find evidence for a higher crash risk for the driving below average speeds. This is most likely due to the fact that the older studies also included maneuvering vehicles. Maneuvering vehicles are more at risk and have, per definition, a low speed. These individual speed models in general estimate a higher individual risk, especially for excessive speeders, than do the aggregated Power and Exponential models based on mean speed.

More recently, during the last decades, with the evolution of technology, the automotive telematics market is growing steadily and a few innovative telematics and driver monitoring systems are introduced in our life. Nowadays, most drivers look for new services providing more options in order to identify their weak points in driving, adjust their driving style and techniques, reward their progress and promote maximum road safety for everyone. More specifically, in many studies that have taken place internationally, a device agnostic platform has developed with the ability to collect data from different sources such as smartphones, OBDs and 4G/5G connected cars. Additionally, recent works used a Driving Data Recorder (DDR) which can provide feedback on driver behavior for accident analysis and other insurance issues (Ohta et al., 1994, Gu et al., 2019). Also, a driving behavior and safety evaluation was conducted through a data recording system called Drive Diagnostics which is a dedicated In-Vehicle Data Recorder (IVDR) system with dimensions of 11x6x3cm, customized inside the vehicle (Toledo et al., 2006). It is worth highlighting that in a survey which took place in young drivers during the first year after their licensure was found that drivers who knew that their driving behavior was being monitored by this device, managed to be less aggressive and drove more ecologically (Prato et al., 2010). Moreover, real driving parameters of driver behavior have been assessed and analyzed through an On Board Diagnostics (OBD-II) device (Yannis et al., 2016). This recording system was developed in the United States of America and is aimed to detect road accidents and mechanical problems in vehicle that caused by high emissions above the acceptable limit values. An android smartphone connects via Bluetooth to the OBD-II and receives information about the vehicle status, such as speed, fuel, temperature, accelerometer values as well as accurate location with a specific latitude and longitude, via GPS updates (Zaldivar et al., 2011).

The objective of this study is to identify the critical driving parameters that affect speeding by utilizing detailed driving data from smartphone sensors. In particular, the authors are to investigate the extent to which the various driving performance parameters (travel distance, harsh acceleration, harsh deceleration, harsh turn, percentage of driving time the driver uses a mobile phone, etc.) interact and determine the percentage of net driving time at a speed above the speed limit.

2. Data Collection

For the purpose of this research, a naturalistic driving experiment was carried out involving 68 drivers during a 15-months timeframe from August 2016 to October 2017 and a large database of 21.610 trips was created. Driving behavior analytics were recorded in real time, using smartphone device sensors. More specifically, an innovative data collection system using a Smartphone Application that has been developed by OSeven Telematics was exploited. A set of sophisticated and personalized interactive tools is applied by OSeven Telematics, powered by breakthrough technology, smart algorithms and reliable metrics. The most advanced Machine Learning techniques are implemented to detect harsh events and speeding, identify the trip transport mode, recognize whether the user is a driver or a passenger and rate the driver's behavior. Consequently, the OSeven platform helps drivers understand their weak points and motivates them to improve their driving behavior and also reduce their fuel costs.

This procedure results to the creation of the risk exposure and driving behavior indicators. The risk exposure indicators are total distance (mileage), duration (total duration of the trip) and driving duration (total duration of the trip not including stops), type(s) of the road network used (given by GPS position and integration with map providers e.g. Google, OSM), time of the day driving (rush hours, risky hours), trip purpose combined with other data sources (speed limits and detailed accident maps). The driving behavior indicators are speeding (duration of speeding, speed limit exceedance etc.), number and severity of harsh events (braking and acceleration), harsh cornerings, driving aggressiveness (e.g. braking, acceleration), distraction from mobile phone use fatigue and driving in risky hours.

It should be highlighted that taking into consideration that privacy and security consist two of the main principles in the field of telematics, the OSeven platform has very clear privacy policy statements for the end users covering the type of data collected, the reason data is collected for, the time that data is stored and the procedures for data security based on encryption standards for data in transit and at rest. All this is done using state-of-the-art technologies and procedures in compliance with GDPR. In this framework all data has been provided by OSeven Telematics in an anonymized format.

The following chart constitutes a preliminary analysis of the variables, which allows for an initial better understanding of the data and the results and will be used to draw qualitative conclusions. Figure 1 illustrates the percentage of speeding time during driving for each different road type respectively. It is noted that in the urban environment the percentage of speeding is the greatest while on highways drivers exceed the speed limits the least, probably because the speed limits on highway are already enough high.

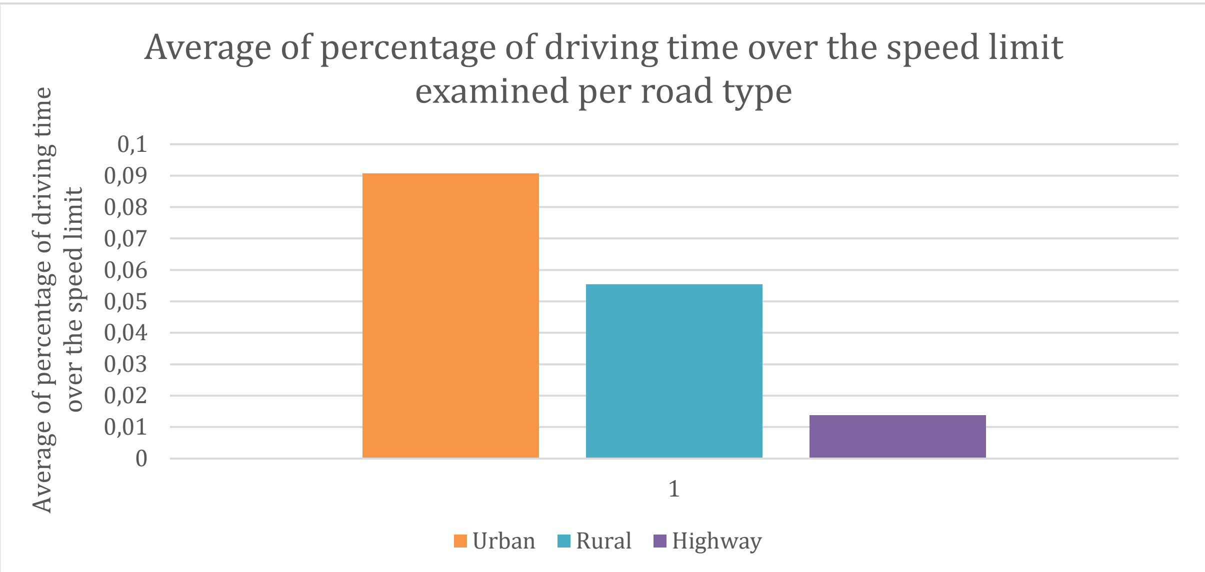

Figure 1: Average of percentage of driving time over the speed limit examined per road type

3. Methodology

As previously reported, data from the driving measures collected by OSeven backend office were analysed using Microsoft Excel and SPSS statistical program. Afterwards, statistical analyses were carried out using linear regression models. This type of analysis was developed to examine whether driving characteristics such as distance, harsh accelerations, harsh braking, smartphone usage (dialling, talking, texting etc.) and driving during risky hours affect and can therefore predict the percentage of driving duration with speed above the speed limit. The basic equation of the regression model is the following:

$$y_i = a + b*x_i + \varepsilon_i \qquad (1)$$

In this case, four statistical regression models forecasting the percentage of driving duration with speed above the speed limit were developed: one overall model and three models for each different road type (urban, rural, highway), as shown below:

- Model 1: Predicting the percentage of speeding – overall model
- Model 2: Predicting the percentage of speeding on urban road
- Model 3: Predicting the percentage of speeding on rural road
- Model 4: Predicting the percentage of speeding on highway

Specifically, all models were designed to have exactly the same variables in order to make it easier to compare models with each other. Then, it was examined for all models separately whether the numerical results quantifying relationships between variables, did satisfy the models' quality. Within the present research, the scope of this analysis was to determine which observed independent variables are highly correlated with the dependent variable and at the same time which of them are inconsistent with each other. A valuable numerical measure of two variables' association is the correlation coefficient, which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables (Schroeder et al., 2016, Ekstrom & Sørensen, 2014). For this reason, the correlation coefficient R^2, which indicates the percentage of the dependent variable that is explained by the independent variables, needs to be as high as possible. It is important to mention that the values and signs of the normal regression coefficients b_i must be reasonably explainable. Additionally, the value of the t-statistic control and the statistical significance level should be acceptable and compensatory for the confidence level that commonly used. The constant coefficient of the equation, which considers all the parameters that have not been taken into account, must be the lowest possible, as well. Lastly, the elasticity (ei) that shows how responsive one variable is to a change in another but also the relevant influence elasticity (ei*) used for quantifying the influence of each individual variable should be examined, process that allows for the comparison between the influence of different variables in a single model. In particular, point estimates of elasticities (ei) are provided by the following formula, for each value (i) in the sample:

$$ei = (\Delta yi \ / \Delta xi)\ (xi / yi) = \beta i\ (xi / yi) \quad (2)$$

4. Results

The results of the previous analyses led to several useful observations regarding the identification of critical driving parameters affecting speeding using data from smartphones. Table 1 provides a description of the parameters that were found to be significant in the linear regression models.

Table 1: *Description of the parameters used in the models*

Independent Variable	Description
totaldist	total distance in km (km)
ha_intensity_high	high intensity harsh acceleration
hb_intensity_high	high intensity harsh braking
hc	harsh cornering
avdecel	average deceleration
mobileusage	mobile usage while driving

In Table 2, the parameters of the driving experiment and the regression results for the overall model and the three models per road type are presented. Furthermore, it is highlighted that all the independent variables used in the models are inconsistent with each other, as it is required (figure 2). The developing models have given a more clear understanding of the relation of driver behavior and speeding, which can be directly translated into realistic coaching targets for drivers and road safety policy-makers.

		totaldist	ha_intensity_high	hb_intensity_high	hc	avdecel	mobileusage	perc_speeding
totaldist	Pearson Correlation	1	,137**	,176**	,237**	,271**	-,035**	,434**
ha_intensity_high	Pearson Correlation	,137**	1	,274**	,066**	-,223**	,019**	,230**
hb_intensity_high	Pearson Correlation	,176**	,274**	1	,055**	-,169**	,004	,242**
hc	Pearson Correlation	,237**	,066**	,055**	1	,037**	,036**	,073**
avdecel	Pearson Correlation	,271**	-,223**	-,169**	,037**	1	-,079**	,004
mobileusage	Pearson Correlation	-,035**	,019**	,004	,036**	-,079**	1	,014*
perc_speeding	Pearson Correlation	,434**	,230**	,242**	,073**	,004	,014*	1

Figure 2: Correlation coefficient of the examined independent variables

Table 2: *Summary table of the four mathematical models developed for speeding*

Independent Variables	**Model 1 (overall model)**				**Model 2 (urban road)**				**Model 3 (rural road)**				**Model 4 (highway)**			
	b_i	t	e_i	e_i*	b_i	t	e_i	e_i*	b_i	t	e_i	e_i*	b_i	t	e_i	e_i*
Constant	0,021	16,182			-10.466	15,990			0.014	10.234			0.002	3.382		
totaldist	0,002	63,679	0,626	9,468	0.014	60.955	1,629	14,808	0.006	63.515	1,385	9,819	0.002	57.137	1,516	7,664
ha_intensity_high	0,031	20,179	0,809	12,235	0.014	5.451	0,297	2,700	0.026	7.441	0,630	4,465	0.031	5.712	0,755	3,819
hb_intensity_high	0,054	19,794	1,006	15,219	0,070	14.091	0,963	8,755	0.032	6.145	0,539	3,818	0.096	12.028	0,831	4,203
hc	-0,002	-6,677	-0,167	-2,519	-0.009	-10.432	-0,406	-3,688	-0.017	-22.407	-1,016	-7,201	-0.033	-5.283	-0,812	-4,103
avdecel	-0,016	-8,674	0,676	10,218	-0.041	-22.659	1,426	12,963	-0.017	-15.455	0,500	3,541	-0.057	-28.715	0,664	3,358
mobileusage	0,014	3,709	0,066	1,000	0.026	6.159	0,110	1,000	0,036	6.666	0,141	1,000	0.104	13.894	0,198	1,000
R^2	0,239				0,193				0,200				0,296			

Regarding the overall model it is observed that that for each additional kilometer that a driver covers, the percentage of speeding increases by 0.002 units. The largest the total distance covered on the travel routes, the higher the ratio of percentage of speeding. This may be explained by the fact that longer distances mean more driving on the interurban road network as well as on highway where the traffic speeds are higher. Additionally, it is noted that for a further high intensity harsh acceleration and deceleration, the percentage of speeding is increased by 0.031 and 0,054 units respectively. This may be because by accelerating and decelerating harshly thus driving in a more nervous way, the driver probably exceeds the speed limits by producing maneuvering in order to maintain their main speed. In an opposite way, harsh cornering as well as mean deceleration result in decrease of the percentage of speeding, as the driver reduces his driving speed. Furthermore, for a 1% increase of the time using a mobile phone while driving, the percentage of speeding increases by more than 0.014 units. This is probably due to the driver's distraction. Due to the fact that the driver is concentrated on the phone and not on the driving process, he can develop high speeds that exceed the speed limits.

Table 2 also demonstrates the elasticities and relative influences of the independent variables of the models developed. Regarding the elasticity values of the overall model, it is observed that the influence of the hb_intensity_high variable is the highest among all independent variables. This shows the significance of the elasticity value of the hb_intensity_high. For a 1% increase in hb_intensity_high, the dependent variable increases 1,006%, as it is presented in figure 3. Furthermore, the variable "mobileusage" shows the least influence on the model. Specifically with respect to the most influential variable, it affects the model 15,2 times less. Regarding the variables "ha_intensity_high" and "avdecel", they have the highest impact on the model after the most influential variable. Impact values are 0,809 for the variable "ha_intensity_high" and 0.676 for the "avdecel" variable, respectively. Finally, the impact of the variables “totaldist” and “hc” are 9,4 and 2,5 times higher than the variable “mobileusage”.

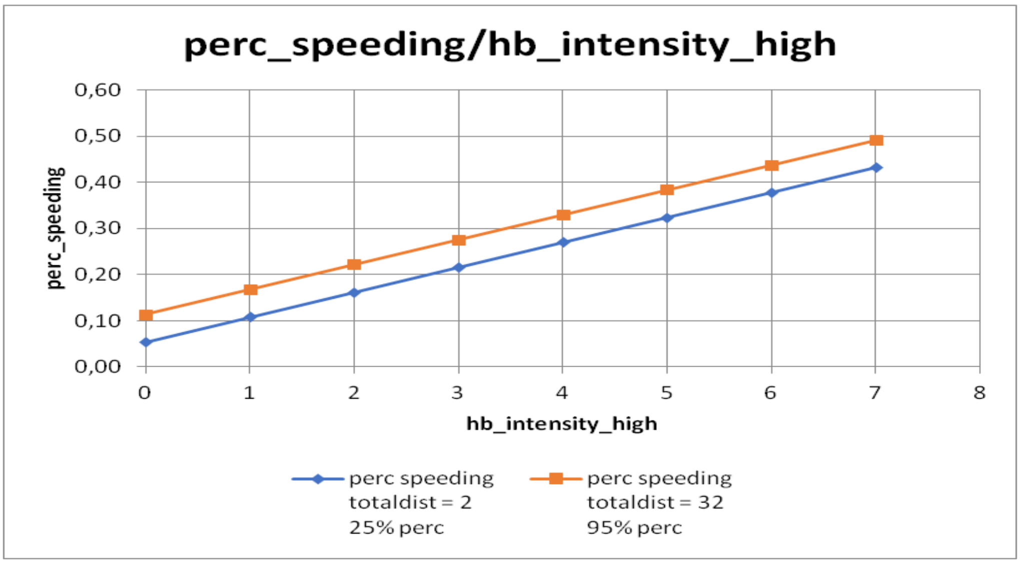

Figure 3: Percentage of speeding to high intensity harsh braking

Furthermore, with respect to urban road, the variable "distance_urban" is the variable with the highest impact on the model compared to the other five independent variables. This shows the significance of the variable in the overall performance indicator. The value of the influence is 1,629. The variable "mobileUsage_urban" has the lowest impact on the model. Specifically, with respect to the most influential variable, it affects the model 14,8 times less. As far as the variables "av_decel_urban" and "hb_intensity_urban_high" are concerned, they seem to have the highest impact after the most influential variable. Finally, the variables "hc_urban" and "ha_intensity_urban_high" have 3,6 and 2,7 times higher impact respectively, than the variable "mobileUsage_urban".

Regarding the rural road model, the variable with the highest impact is "distance_rural", namely 9,819 times higher than "mobileUsage_rural", which is the variable with the lowest impact on the speeding model. Additionally, it is found that "hc_rural" and "ha_intensity_rural_high" are the variables with the highest impact on the model after the most influential variable, namely 1,016 and 0,630 respectively. Regarding the two remaining variables, "hb_intensity_rural_high" and "av_decel_rural", they have a similar impact on the speeding model, namely 3,818 and 3,541 times higher than "mobileUsage_rural" respectively.

As for the highway model, once again, the variable with the highest impact is "distance_ highway", namely 7,664 times higher than "mobileUsage_ highway", which is the variable with the lowest impact on the speeding model. Regarding the variables "hb_intensity_highway_high" and "hc_highway", they seem to have the greatest impact after the most influential variable, namely 0,831 and -0,812 respectively. Finally, the variables "ha_intensity_highway_high" and "av_decel_highway" have a similar impact on the speeding model, namely 3,819 and 3,358 times higher than "mobileUsage_ highway" respectively.

Finally, with respect to the quality of the model, it is noted that the relatively low R^2 value indicate that the examined independent variables can partially predict the dependent. Still they are acceptable and might be improved with the examination of additional independent variables which were either not recordable or partially recorded, and thus excluded from this analysis.

5. Conclusions

The present research aims to identify the critical driving parameters that affect speeding using data from smartphones. To this end, four statistical linear regression models forecasting the percentage of speeding time during driving were developed: one overall model and three models for each road type (urban, rural, highway), the results of which are discussed in the previous section. Results indicate that the percentage of speeding time during driving has been more accurately predicted for the highway model, taking into account its relatively greater R^2 value compare to the other three prediction models. This may probably be explained by the fact that highways are characterized by a more normalized traffic flow, resulting in a driving behavior with less harsh events, namely acceleration, deceleration, cornering in comparison with the other types of road network.

Furthermore, one important finding is that the distance covered in a travel trip in an urban and rural environment, as well as on a highway is crucial for forecasting the percentage of speeding time during driving as it is included in all three models for a particular road type and is the variable with the highest impact on the models. The longer the distance is, the higher percentage of speeding. This conclusion can be interpreted by the fact that drivers are familiar with the conditions of the road environment and, in order to travel faster and reach their destination more quickly, they speed up, resulting in exceeding the speed limits. Additionally, it should be highlighted that the variable with the highest impact on the overall model is found to be the number of high-intensity harsh deceleration. This may probably be explained by the fact that harsh deceleration events are a determining factor because of their direct relation to aggressive driving, as drivers tend to harshly accelerate when the space is brought up, and therefore, when they encounter lower-speed vehicles, they are forced to slow down by performing harsh deceleration events.

Concluding, it is expected that this research can provide considerable gains to the society, since the stakeholders including policy makers and industry could rely on the results and recommendations regarding risk factors that appear to be critical for safe driving. As for further research, the examination of additional methods of analysis are proposed, such as factor analysis, logistic regression as well as microscopic data analysis of the database collected could be implemented through econometric techniques such as time-series analysis. Future studies would also benefit from exploiting more advanced technological equipment for recording the in-vehicle driving behavior such as precise GPS equipment, radars measuring the reaction time and headways as well as cameras inside and outside the vehicle. However, these solutions sometimes come at considerable costs, resulting in the investigation of affordable and ergonomic ways of monitoring and assessing driving behavior.

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

6. References

Global status report on road safety by World Health Organization, 2018 (https://www.who.int/violence_injury_prevention/road_safety_status/2018/en/, accessed 19 March 2019)

Elvik, R., Christensen, P., & Amundsen, A. (2004). Speed and road accidents. An evaluation of the Power Model. TØI report, 740, 2004.

Solomon, D. (1964). Accidents on main rural highways related to speed, driver, and vehicle.

Kloeden, C. N., McLean, A. J., Moore, V. M., & Ponte, G. (1997). Travelling speed and the risk of crash involvement volume 2-case and reconstruction details. Adelaide: NHMRC Road Accident Research Unit, The University of Adelaide.

Kloeden, C. N., McLean, J., & Glonek, G. F. V. (2002). Reanalysis of travelling speed and the risk of crash involvement in Adelaide South Australia. Australian Transport Safety Bureau.

Kloeden, C. N., Ponte, G., & McLean, J. (2001). Travelling speed and risk of crash involvement on rural roads. Australian Transport Safety Bureau.

Taylor, M. C., Lynam, D. A., & Baruya, A. (2000). The effects of drivers' speed on the frequency of road accidents. Crowthorne: Transport Research Laboratory.

Schroeder, L. D., Sjoquist, D. L., & Stephan, P. E. (2016). Understanding regression analysis: An introductory guide (Vol. 57). Sage Publications.

Ekstrom, C. T., & Sørensen, H. (2014). Introduction to statistical data analysis for the life sciences. CRC Press.

Gu, Y., Wang, Q., & Kamijo, S. (2019). Intelligent Driving Data Recorder in Smartphone Using Deep Neural Network-Based Speedometer and Scene Understanding. IEEE Sensors Journal, 19(1), 287-296.

Ohta, T., & Nakajima, S. (1994). Development of a driving data recorder. JSAE review, 15(3), 255-258.

Prato, C. G., Toledo, T., Lotan, T., & Taubman-Ben-Ari, O. (2010). Modeling the behaviour of novice young drivers during the first year after licensure. Accident Analysis & Prevention, 42(2), 480-486.

Yannis, G., Tselentis, D., Papadimitriou, E., & Mavromatis, S. (2016). Star rating driver traffic and safety behavior through OBD and smartphone data collection. Proceedings of the International Conference on Artificial Intelligence and Intelligent Transport Systems, Technical University of Brest, Brest, Belarus.

Zaldivar, J., Calafate, C. T., Cano, J. C., & Manzoni, P. (2011, October). Providing accident detection in vehicular networks through OBD-II devices and Android-based smartphones. In 2011 IEEE 36th Conference on Local Computer Networks, 813-819.