Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 00 (2018) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

World Conference on Transport Research - WCTR 2019 Mumbai 26-31 May 2019

Driving speed model development using driving data obtained from smartphone sensors

Dimitrios I. Tselentis[a]*, Christina Gonidi[a], George Yannis[a]

[a]*National Technical University of Athens, 5 Iroon Polytechniou St., GR-15773, Athens, Greece*

Abstract

The aim of this paper is the development of driver speed models based on detailed driving data collected from smartphone sensors. More specifically, this research investigates to which extent various driving behaviour parameters (harsh acceleration and deceleration events, driving distance, percentage of driving time per different road types, etc.) interact with each other and how these might potentially serve as driving speed predictors. Real time driving behaviour data collection was carried out using a smartphone application. Data were collected from 100 drivers and a total of 18,853 trips between July and December 2016. Six different linear regression models are developed to predict average driving speed under different driving conditions. One model for each road type (urban, rural and highways), one model for the risky hours period and one for the rest of the day, and a general model. The results indicated that the distance covered by the driver, the number of the harsh events occurred and the average deceleration are the strongest predictors of the average speed of a driver.

Peer-review under responsibility of WORLD CONFERENCE ON TRANSPORT RESEARCH SOCIETY.

Keywords: driving behaviour; driving speed; number of harsh manoeuvres; average acceleration; linear regression

1. Main text

Road safety is a complex issue, as it depends on many factors. The most important factors are the vehicle, the road environment, and the drivers (Frantzeskakis & Golias, 1994). In most cases, two or three of these factors are enough to cause an accident. The complexity and lack of detailed recorded data as well as the lack of an analysis of the driving conditions under which an accident takes place, do not always allow for the objective estimation of each factor's significance. However, several in-depth accident studies showed that the driver solely or in combination with the other

* Corresponding author. Tel.: +30-210-772-2210; fax: +30-210-772-1454.
E-mail address: dtsel@central.ntua.gr

Peer-review under responsibility of WORLD CONFERENCE ON TRANSPORT RESEARCH SOCIETY

two factors is the main cause of road accidents. More specifically, one of the most important problems in road safety, which greatly increases accident risk, is a driving speed. Driving over the speed limit is a determining factor in about 30% of road accidents. Approximately 40-50% of all drivers exceed the speed limit, while 10-20% of them exceed this limit by more than 10 km/hour (SafetyNet, 2009:3). Excessive driving speed increases the risk of not only collision but also of serious injury or/and fatal road accident.

Many studies have attempted to estimate the correlation between driving speed and its related factors. In this paper, two such groups of studies are examined: studies that have used in-vehicle recoding systems to investigate driving behaviour and studies aiming to define the factors that predict driving speed in relation to driver's behaviour and characteristics. In this research, methodologies that have been used in studies concerning in-vehicle special diagnostic systems, are taken into account.

Regarding the first group of studies, Ohta, Tohru, and Shouji Nakajima were among the first to use an in-vehicle driving data recorder (DDR). DDR is small and light and can be used for 9.000 driving hours or 100.000 kilometres. The data are initially stored in a memory stick and can be later analysed using a personal computer. The data are initially recorded through a time-based system, in which they are temporarily stored. Afterwards, data are stored in a frequency-based system that is based either on the frequency of use or on the frequency of appearance. DDR is designed to record the driving conditions under ordinary road circumstances and analyse driver's behaviour on the road, thus providing quantitative driving behaviour data (Ohta & Nakajima, 1994). Toledo et al. used the DriveDiagnostics, which is an in-vehicle data recorder with dimensions of 11x6x3cm and is charged using vehicle's battery. The system collects data such as vehicle's acceleration (in x, y, z coordinates), speed, position estimated through GPS, fuel consumption, total trip time, etc. (Toledo et. al., 2008). In addition, Zaldivar et al. (2011) used a recording system, called On Board Diagnostics (OBD-II), the development of an application for Android smartphones, which collects vehicle information and detects car accidents. Using the Bluetooth sensor, the mobile phone is connected to the OBD-II device and retrieves information on the condition of the vehicle. In addition, it pinpoints the vehicle's speed and exact position via the GPS sensor (Zaldivar et al., 2011). Tselentis et al., (2017) have further highlighted the importance of charging insurance premiums based on drivers' behaviour (Usage-Based motor Insurance (UBI)) and have analysed the willingness to pay for these insurance schemes (Tselentis et al., 2018). The main concept of these insurance schemes is that instead of a fixed price, drivers have to pay a premium based on their travel and driving behaviour. This research (Tselentis et al., 2017) examined the correlation between PAYD, PHYD and crash risk.

Regarding the second group of studies, researchers have examined various factors that predict driver's speed. According to Aarts and Van Schagen (2006), there is a contradiction in literature about whether or not vehicle crashes rate increases as the average speed decreases. Aarts and Van Schagen (2006) conclude that for a specific road, crash risk increases as speed of either an individual vehicle or total flow speed increases. However, they note that the precise rate depends on a great number of external factors, which do not allow its exact estimation and generalization for different road types (Aarts & Van Schagen, 2006). In addition, Cameron and Elvik (2010) concluded that there is no evidence that connects speed increase on urban arterial roads and crash severity. The power, applicable to serious casualties on urban arterial roads, was considerably less than that on rural roads, which in turn was considerably less than that on freeways (Cameron & Elvik 2010). Another study by Elvik (2009) analysed the relation between speed and road safety in order to evaluate the Power Model, which prevailed thanks to its parsimony and simplicity. Two models were developed: one model for urban roads or roads in residential areas, and one for rural roads and freeways. The conclusion was that in recent years, the effect of speed on road safety seems to be less important, although speed remains an important risk factor for accidents. In many countries, excessive speeding is one of the most important problems of road safety. Another study, conducted by Farmer et al., (2010) showed that most drivers fasten their seat belt as soon as they enter their vehicle. In this study, drivers who were alerted through sound messages on exceeding the speed limit, adjusted their driving behaviour right away since, because they were annoyed by the sound. For those who did not receive any sound alert, no compliance to safety rules was observed. Regarding harsh acceleration or deceleration events, the sound alert did not result in any significant improvement of the driving behaviour (Farmer et al., 2010). Wang et al. (2012) presented road safety data and theories, in order to explain how and why the risk factors affect road accidents and use findings in order to reduce wrong driving behaviours. Various factors were found to affect road safety and need further examination, including speed, traffic density, flow, congestion, demographics, driving behaviour (alcohol consumption, helmet usage and seat belt usage), and land use (Wang et al., 2012). By

unitizing the increase in available real-time traffic data and aiming at monitoring road safety, Theofilatos & Yannis (2014) examined the traffic characteristics and weather characteristics on road safety. It is found that the increase in speed limits is related with the increase in road accidents, and precipitation leads to increased accident frequency, but it does not seem to have a consistent effect on the severity of an accident.

Among methodologies which have been used in other studies concerning in-vehicle special diagnostic systems, those by Prato et al. (2010) are of particular interest for this research. They examined novice young drivers over a period of 12 months after licensure; during the first three months, a parent was present in the vehicle, while afterwards, the young drivers could drive alone. The results confirmed that male novice young drivers are more risk-prone and that when driving without parental supervision, they become more 'aggressive'. When driving under parental supervision, their accident risk was related to the driving behaviour of the adults. Collecting data through an IVDR system also played an important role in reducing young drivers' accident risk. The young drivers who were monitored by their parents throughout the study turned to be more careful in their solo driving period, while those who were not always supervised by their parents through the electronic application, did not appear willing to improve their risk-prone behaviour. Vaiana et al. (2010) examined driving behaviour (aggressive or not) by placing on an X, Y axes of a 2D plane the accelerations of the vehicle (the longitudinal and the lateral one, respectively). In addition, the friction circle of the vehicle was used, which depends on the tire and the road surface characteristics. Taking also into account the experience of the driver and the car type, the Driving Style Diagram (DSD) was created, which constitutes a combination of all the above-mentioned parameters. The same vehicle and the same mobile phone were always used in the experiments, and five drivers with different driving characteristics participated. The behaviour of each driver was estimated on the basis of the average acceleration above the limiting edges of the DSD. The study resulted in the following: the threshold value for distinguishing between aggressive and safe drivers can be set 9 % average acceleration above the limiting edges of the DSD. As literature review showed (Tselentis et al. (2017), Tselentis et al., (2018)), driving data collection through smartphone applications has become popular nowadays. The above studies and methodologies have greatly contributed to examining a series of factors related to driving speed. At the same time, they exploit a rather limited number of drivers, driving situations or variables. Thus, further research is necessary, using a larger number of drivers driving for longer periods in real time conditions, to further investigate a) which factors are significant predictors of average speed, b) how different road type affects the average speed, c) whether or not the same variables play a similar role on any road type (highway, urban, rural road) and d) whether a driver's speed is altered depending on the time of the day (risky hours timeframe or not).

The objective of this paper is to develop driving speed models based on a greater number of detailed driving data collected from smartphone sensors. More specifically, the paper investigates the extend at which the various parameters examined, which compose driving performance (the number of harsh acceleration/ deceleration events, distance travelled, percentage of driving time per different road types, etc.) interact with each other and affect driver's speed prediction under different driving conditions. To this end, employment and development of proper statistical methods of data analysis take place.

2. Data collection

The real-time recording and collection of the data required for examining driving behaviour took place using the OSeven smartphone application installed in driver's mobile phone. This smartphone application is part of the OSeven platform, which is used for data recording, collection, storing and visualization of the driving behaviour by using advanced Machine Learning algorithms and statistical methods for data processing to derive the necessary driving metrics. The data are collected using the mobile phone's sensors, transmitted through WiFi or cellular network and stored in large databases of the OSeven backend platform. Data were collected from 18.853 trips performed by 100 drivers, who drove between July and December 2016.

This procedure results to the creation of the risk exposure and driving behaviour 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), weather conditions, trip purpose combined with other data sources (speed limits and detailed accident maps). The driving behaviour indicators are speeding (duration of speeding, speed limit exceedance etc.), number and severity of harsh events, harsh braking (longitudinal

acceleration), harsh acceleration (longitudinal acceleration), harsh cornering (angular speed, lateral acceleration, course), driving aggressiveness (e.g. braking, acceleration), distraction from mobile phone use.

The initial large database obtained from OSeven includes the information collected for each trip (one trip per row). Since investigating the average speeding behaviour of each driver is the objective of this research, the initial database was modified so that the new database created, includes one row per each individual driver and not per trip. To avoid neglecting important information, the standard deviation, average, minimum and maximum value from all the trips were calculated for each driving variable. Furthermore, the initial database was split into five, so that individual models (apart from the general one - Model 1) can be developed to predict the driver's average speed in different driving states such as:

- Driving during less risky hours (on all road types) - Model 2
- Driving during risky hours (on any road type) - Model 3
- Driving on urban road (anytime of the day) - Model 4
- Driving on rural road (anytime of the day) - Model 5
- Driving on highway (anytime of the day) - Model 6

To shed light on the speed behaviour of each driver, the following charts are provided. These charts constitute 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.

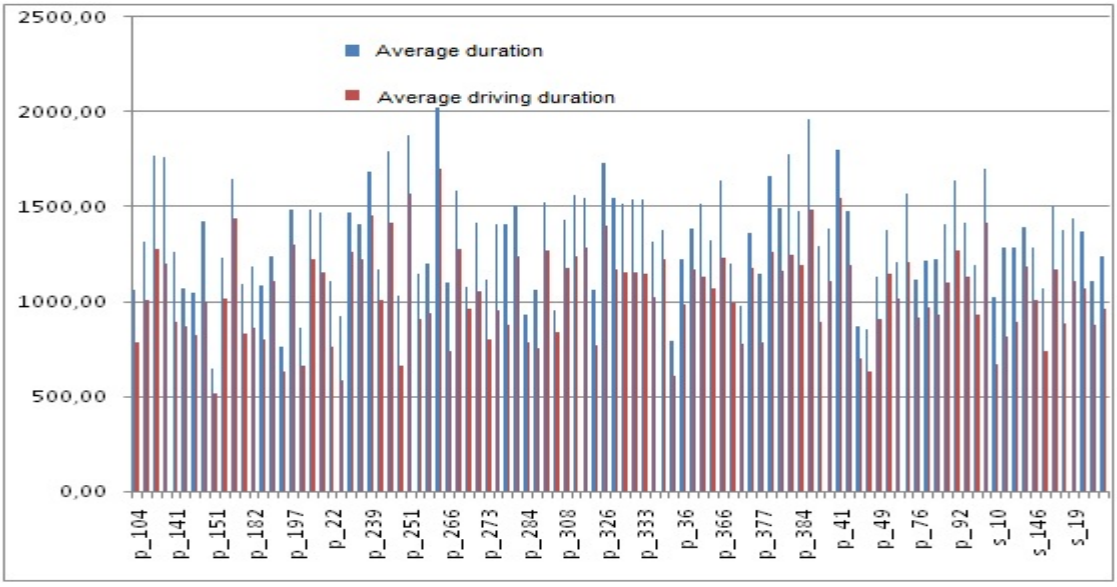

Fig. 1. Average duration and driving duration per driver

It is obvious that the higher the difference between the average driving duration and the average duration, the more a driver drives during traffic congestion. It is important to examine in which road type congestion this is more frequent. It is observed from figure 1 that the highest differences between average duration and average driving duration per driver exist on urban roads and rural roads while on highways, the differences are much lower for all drivers. Consequently, when driving on a highway a driver is rarely confronted with traffic congestion, in contrast to urban and rural roads. On urban roads, a significant difference exists between the average driving duration and the average duration, which is almost similar for all drivers. On the other hand, high differences between drivers are observed in rural roads probably because some drivers do not encounter traffic congestion (the time of average driving duration and the time of average duration are almost equal).

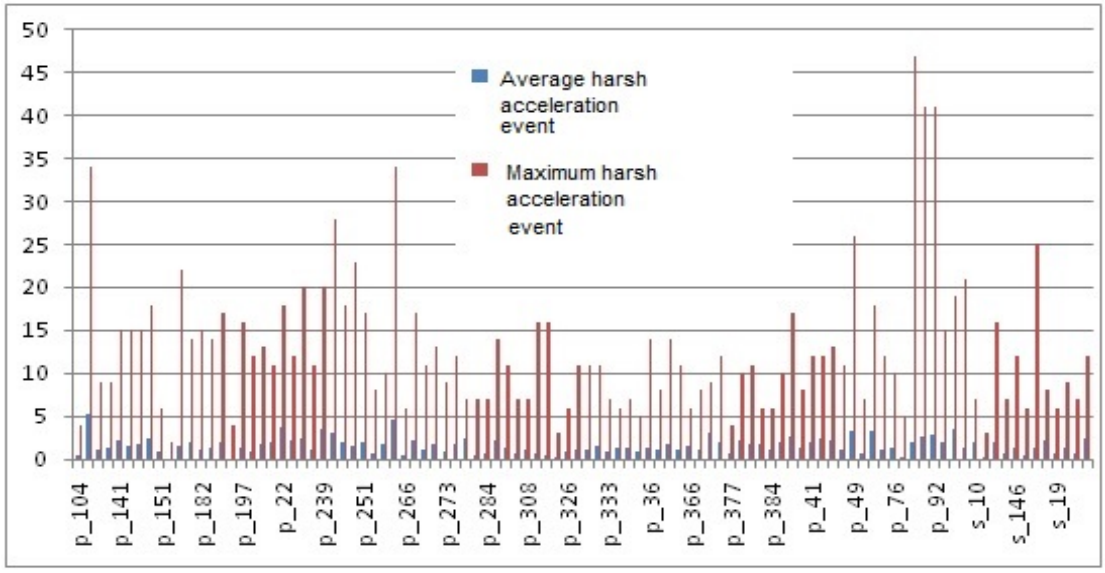

Fig. 2. Average and maximum harsh acceleration events per trip per driver

Various driving behaviours are observed in figure 2. The most noticeable observation is that drivers with high average number of harsh acceleration events also have a high number of maximum harsh acceleration events per trip. It should be noted though that the driver having the highest number of maximum harsh acceleration events, does not have the highest number of average harsh accelerations events. This can be attributed to a random aggressive behaviour that occurs occasionally and is not constant. On the contrary, drivers with the lowest number of maximum harsh acceleration events are those with the lowest number of average harsh acceleration events.

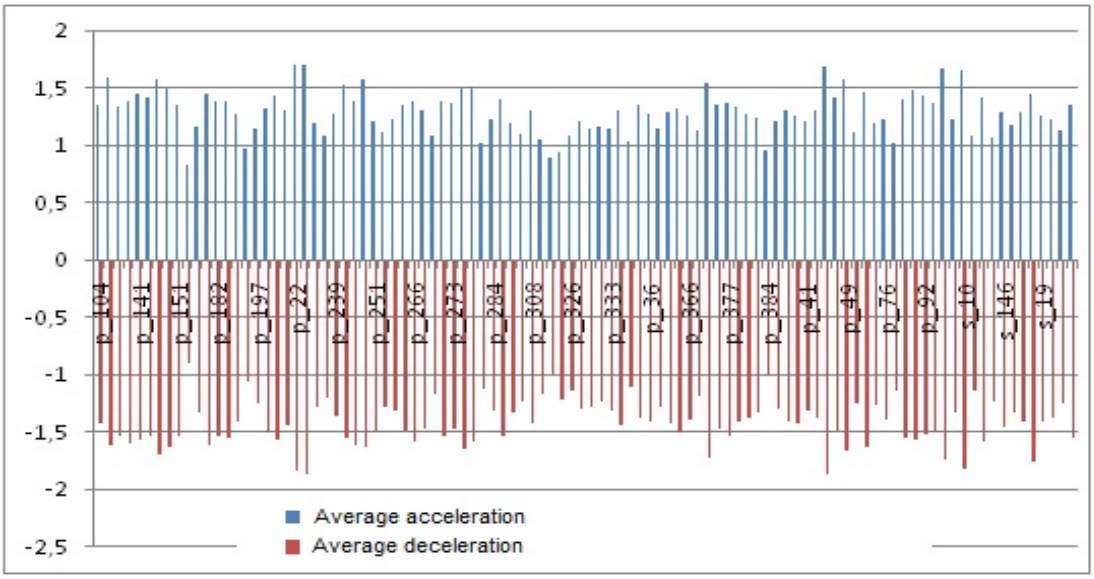

Fig. 3. Average Acceleration and deceleration per driver

Based on figure 3, the average acceleration appears to be approximately equal to the average deceleration for all drivers, which leads to the conclusion that drivers who accelerate much, also decelerate much. Additionally, the number of harsh acceleration events is higher than the number of harsh deceleration events despite the fact that the average acceleration appears to have lower values than the average deceleration.

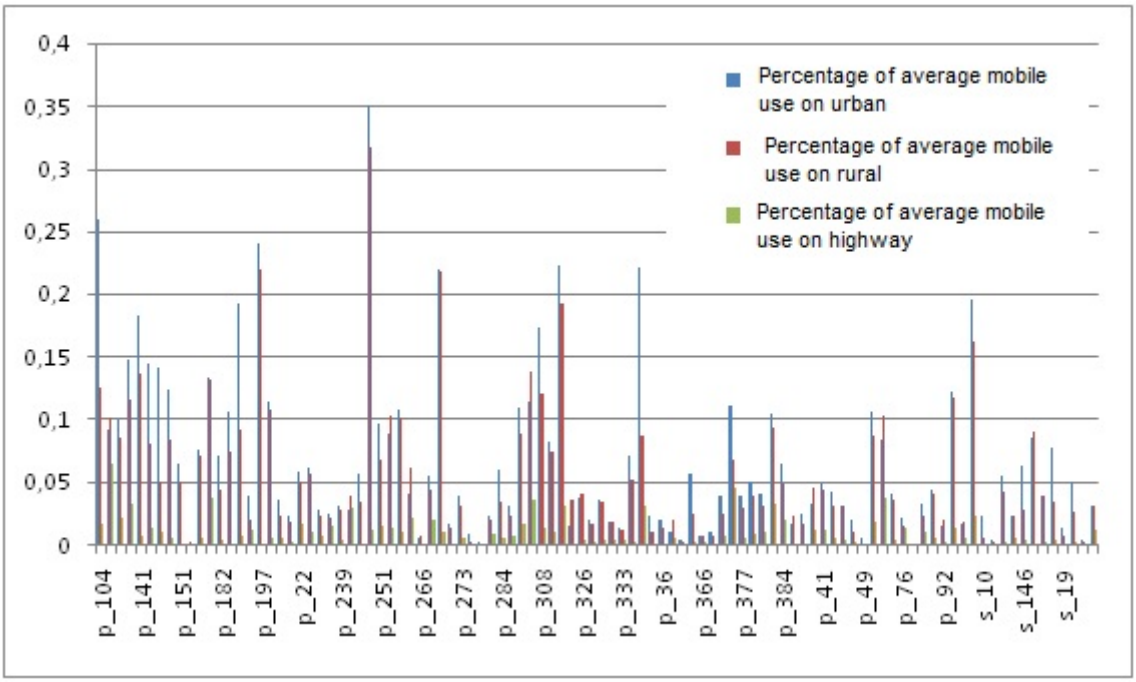

Fig. 4. Percentage of average mobile use per trip on any road type per driver

It is evident from figure 4 that drivers use their mobile phone more on urban roads and significantly less when travelling on highways.

3. Methodology

As mentioned above, models of the linear regression family are developed herein the theoretical background of which is given below. Simple linear regression aims to model the relationship between two quantitative variables, x and y, by fitting a linear equation to observed data. One variable is considered an explanatory variable, and the other is considered a dependent variable. The linear regression line has an equation of the form $Y = b_0 + b_i * X$, where X is the explanatory variable and Y is the dependent variable. Parameter b is the slope of the line, and b_i is the intercept (the value of y when x is equal to 0). Before attempting to fit a linear model to observed data, one should first determine whether there is a relationship between the variables of interest. This does not necessarily imply a causation between the two variables, but that there is some significant correlation between the two variables. A scatter plot can be a helpful tool in determining the strength of the relationship between two variables. If there appears to be no correlation between the proposed explanatory and dependent variables (i.e., the scatter plot does not indicate any increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model. 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 (Larry D. Schroeder, David L. Sjoquist, Paula E. Stephan, 2016, Claus Thorn Ekstrøm, Helle Sørensen, 2015).

In all linear regression models, the following steps are necessary. First of all, the values and the sign of the regression coefficients b must be explainable. Coefficient b indicates how much the dependent variable changes as the independent variable changes. Second, the constant coefficient a of the equation must be the lowest possible. The

constant considers the parameters that have not been taken into account. Third, the correlation coefficient R^2 needs to be as high as possible. The coefficient R^2 indicates the percentage of the dependent variable that is explained by the independent variables. Fourth, the t-statistic must be higher than 1.7 for significance level of 5%. The t-statistic measures shows whether or not the initial hypothesis is rejected. Fifth, the relative influence e_i*, used for quantifying the influence of each individual variable, should be checked, which allows for the comparison between the influence of different variables in a single model. In other words, e_i is the elasticity value and e_i* is the elasticity value normalized. All these steps are checked for the six mathematical models developed in this research. After an appropriate number of data processing tests performed, this study concluded to the linear regression models presented below.

4. Results

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 Variables	Description
av_distance	average distance travelled by the driver
dr_dur_perc	percentage of driving duration
hc	number of harsh cornering events occurred
ha	number of harsh acceleration events occurred
av_acc_st_dev	standard deviation of average acceleration
av_dec_st_dev	standard deviation of average deceleration
av_dec	average number of deceleration events
av_mu	average time of mobile phone use

The six mathematical models developed are summarized below in table 2. Results indicate that in all models, total distance in kilometres and a parameter for acceleration exists as significant variables and that in the general model and the models inside and outside risky hours, the very same variables have been set. All results are reasonably explained and confirmed by the findings of the existing literature.

Concerning the general model, it is observed that for each additional kilometre that a driver covers, the logarithm of the average speed increases by 0.004. In other words, the longer the total driving distance, the higher the vehicle's speed. Perhaps this happens because more driving time is needed to travel longer distances on rural roads, where speed is higher. It is also observed that for every additional harsh acceleration event that takes place, the logarithm of the average speed increases by 0.012, which is not an irrational conclusion to draw. Furthermore, when the standard deviation of the average deceleration recorded is increased by one, the logarithm of the average speed increases by 0.255, perhaps because this increase in the standard deviation of the average deceleration is the result of the changes (increase and decrease) of the speed. Thus, the higher the standard deviation of the average deceleration, the higher the driving speed. When the driving duration on urban roads increases by 1%, the logarithm of the average speed decreases by 0.183. Perhaps this is because congestions are more frequent on urban roads and road capacity decreases due to illegal parking.

Table 2. Summary table of the six model developed.

Independent Variables	Model 1 (General model)				Model 2 (Less risky hours)				Model 3 (risky hours)				Model 4 (Urban road)				Model 5 (Rural road)				Model 6 (Highway road)			
	βi	t	ei	ei*	βi	t	ei	ei*	βi	t	ei	ei*	βi	t	ei	ei*	βi	t	ei	ei*	βi	t	ei	ei*
Constant	1.554	38.873			1.726	49.829			1.548	38.632			1.214	32.176			1.593	99.944			1.915	71.427		
av_distance	0.004	4.772	0.182	5.222	0	2.228	0.167	5.044	0.005	5.035	0.197	6.527	0.019	5.109	0.201	3.415	0.005	7.119	0.225	3.069	0	2.430	0.071	-9.555
ha	0.012	1.965	0.035	1	0.014	2.505	0.062	1.860	0.012	1.969	0.030	1									0.005	1.764	0.054	-7.253
av_dec_st_dev	0.255	3.803	0.194	5.579	0.070	1.769	0.033	1	-0.186	-4.586	0.194	6.412												
dr_dur_perc	-0.183	-4.502	-0.174	-4.994	-0.239	-5.533	-0.090	-2.707	0.253	3.671	-0.182	-6.006												
av_dec													-0.098	-4.816	0.293	4.991								
av_acc_st_dev													0.131	2.589	0.059	1	0.097	3.146	0.073	1	0.238	2.694	0.025	-3.324
hc																	-0.010	-3.648	-0.078	-1.063				
av_mu																					-0.336	-3.430	-0.007	1
correlation coefficient	0.569				0.564				0.575				0.369				0.372				0.354			
Dependent variable: The logarithm of the average speed																								

Table 2 also demonstrates the elasticities and relative influences of the independent variables of the models developed. Regarding the elasticity values of the general model, it is observed that the influence of the av_dec_st_dev variable is the highest among all four independent variables. This shows the significance of the elasticity value of the av_dec_st_dev. For a 1% increase in av_dec_st_dev, the av_distance, the dr_dur_perc and ha, the dependent variable increases by 0.194%, 0.182%, -0.174% and 0.035% respectively. Regarding the relative influence values, it is shown that the ha variable has the lowest influence on the general model. The ha influences the general model by 5.5 times less than av_dec_st_dev, 5.2 times less than the av_distance, and about 5 times less than dr_dur_perc.

Concerning the model predicting the average speed when the driver travels during less risky hours, is observed that for each additional kilometre travelled, the logarithm of the average speed increases by 0.001 m/s. Additionally, for each additional harsh deceleration manoeuvre, the logarithm of the average speed increases by 0.014 and that for an increase of 1% in the standard deviation of the average deceleration, it increases by 0,07. Finally, when the driving duration on urban road increases by 1%, the logarithm of the average speed decreases by 0.239. Regarding the elasticity values of the less risky hours, it is evident that the influence of the variable of the av_distance is the highest among all four independent variables. This presents the significance of the av_distance's influence on the dependent value. For a 1% increase of the the av_distance, dr_dur_perc, ha and av_dec_st_dev the dependent variable increases by 0.167%, -0.09%, 0.062% and 0.033% respectively. Regarding the relative influence, it is observed that the ha variable has the lowest influence on the dependent variable. The av_dec_st_dev influences the dependent variable by approximately 5 times less than the av_distance, 2.7 times less than dr_dur_perc, and 1.86 times less than ha.

Regarding the risky hours model, it is observed that for each additional kilometre travelled, the logarithm of the average speed increases by 0.005 and that for any additional harsh acceleration events, the logarithm of average speed increases per 0.012. As standard deviation of the average deceleration increases by one unit, the logarithm of average speed increases by 0.253 while for a 1% increase in driving duration on urban road, it drops by 0.186. Regarding the elasticity values of the risky hours model, it is shown that the influence of the variable of the av_distance is the highest among all four independent variables. This demonstrates the significance of the av_distance's influence on the dependent variable. For a 1% increase of the av_distance, av_dec_st_dev, dr_dur_perc and ha, the dependent variable increases by 0.197%, 0.194%, -0.182% and 0.03% respectively. Regarding the relative influence values, it is demonstrated that ha has the lowest influence on the dependent variable. The ha variable affects the dependent variable by about 6.5 times less than av_distance, 6.4 times less than av_dec_st_dev and 6 times less than dr_dur_perc.

As for the urban road model, results showed that for each additional kilometre in the driving distance the logarithm of the average speed is increased by 0.019 and therefore, the longer the total driving distance, the higher the driver's speeds. This is probably because more driving time is needed to travel longer distances on rural roads, where speed is higher. As the standard deviation of the average acceleration is increased by one, the logarithm of the average speed is increased by 0.131 perhaps because the increase in the standard deviation of the average acceleration is caused by the increase and decrease of speed and therefore, the higher the speed used, the higher the standard deviation of the average acceleration. For a 1 m/s2 increase in average deceleration, the logarithm of average speed decreases by 0.98 perhaps due to the different road environment such as the existence of traffic lights or he higher traffic density that cause a speed reduction. Regarding the elasticity values of the urban road model, it is observed that the influence of the variable of the av_dec is the highest among all three independent variables. For a 1% increase in the av_dec, av_distance and av_acc_st_dev the dependent variable increases by 0.293%, 0.201% and 0.059%. Regarding the relative efficiency values, it is observed that av_acc_st_dev has the lowest influence on the dependent variable. The av_acc_st_dev variable influences the dependent variable by about 5 times less than av_dec and 3.4 times less than av_distance.

Regarding the rural road model, the logarithm of average speed appears to increase by 0.005 for each additional kilometre travelled and therefore, the longer the total driving distance covered, the higher the driving speed is. This is probably because longer distances signify more driving on rural roads, where the driving speed is higher. Furthermore, for each additional harsh cornering event that occurs, the logarithm of the average speed is reduced by 0.01, which is probably due to the fact that when a vehicle turns, the driver automatically slows down in order to make the turn. Finally, when the standard deviation of the average acceleration increases by one unit, the logarithm of the average speed increases by 0.097. Regarding the elasticity values of the rural road model, it is observed that the influence of av_distance is the highest among all three independent variables. For a 1% increase of the av_distance, hc and av_acc_st_dev the dependent variable increases by 0.225%, -0.078% and 0.073%. Regarding the relative influence, it

is observed that av_acc_st_dev has the lowest influence on the dependent variable. The av_acc_st_dev influences the dependent variable by 3.069 times less than the av_distance and 1.063 times less than hc.

Concerning the highway model, it is observed that for each additional kilometre travelled, the logarithm of the average speed increases by 2.99*10-4 and as a result, driving speed increases as total driving distance increases. This is explained by the fact that longer distances require longer driving time on highways where driving speed is higher. Additionally, it is observed that the logarithm of average speed increases by 0.005 for each additional harsh acceleration event that occurs which leads to the conclusion that more aggressive drivers tend to drive over the speed limits. For a unit increase in standard deviation of the average acceleration, the increase in the logarithm of the average speed is 0.238. Finally, for a second increase in the mobile phone usage time, the logarithm of the average speed is reduced by 0.336. This leads to the conclusion that mobile usage distracts driving attention and results in speed reduction. Regarding the elasticity values of the highway model, it is observed that the influence of av_distance is the highest among all three independent variables. For a 1% increase of the av_distance, ha, av_mu and av_acc_st_dev, the dependent variable is increased by 0.071%, 0.054%, 0.025% and -0.007%. Regarding the relative influence values, it is observed that the av_mu variable has the lowest influence on the highway model. The av_mu variable influences the dependent variable by 9.5 times less than the av_distance, 7.25 times less than hc and 3.3 times less than av_acc_st_dev.

Table 3. Relevant influence of the models' variables.

Independent Variables	Relevant influence		
	Model 1 (general model)	Model 2 (less risky hours)	Model 3 (risky hours)
av_distance	1.090	1.000	1.182
ha	1.152	2.036	1.000
av_dec_st_dev	5.872	1.000	5.858
dr_dur_perc	1.942	1.000	2.027

Table 3 presents the relevant influence of the models' variables that predict the average driving speed, the average driving speed during risky hours, and the average driving speed during less risky hours. It appears that the variable "av_dec_st_dev" affects the average speed 5.87 and 5.85 times more in model 1 and 3 respectively than in model 2. This is probably a result of higher traffic flow during less risky hours and therefore drivers decelerate more often but more smoothly as speed is lower and the need for braking is expected. Furthermore, the variable “dr_dur_perc” affects the average speed 2.02 times more for models 1 and 3 than for model 2 and as a result, it can be concluded that during risky hours, traffic volume is lower and therefore driving speed is higher. Additionally, "ha" variable influences average speed 2.04 times more for model 2 than for the other two models. This is probably because of the fact that traffic flow is higher during these hours and drivers tend to be more nervous and aggressive. On the other hand, drivers tend to be more relaxed during risky hours when commuting purposes are not often related with strictly scheduled obligations. Finally, the effect of the total driving distance percentage on the logarithm of the average speed is the same for all models. This is not surprising since driving distance is related to the road type rather than to the time-period. Distance certainly affects speed a lot but this happens regardless of the time-period.

Sensitivity analysis conducted are presented in figure 5. It shows to what extent the increase in the average driving distance influences average driving speed for the three models that have the same variables (model 1, 2 and 3). For this purpose, the values of all independent variables except from the average distance remained fixed and the average speed is plotted for these three models. The value remained fixed at 1.71, 0.34 and 0.52 for the variables “ha”, “av_dec_st_dev” and “dr_dur_perc”.

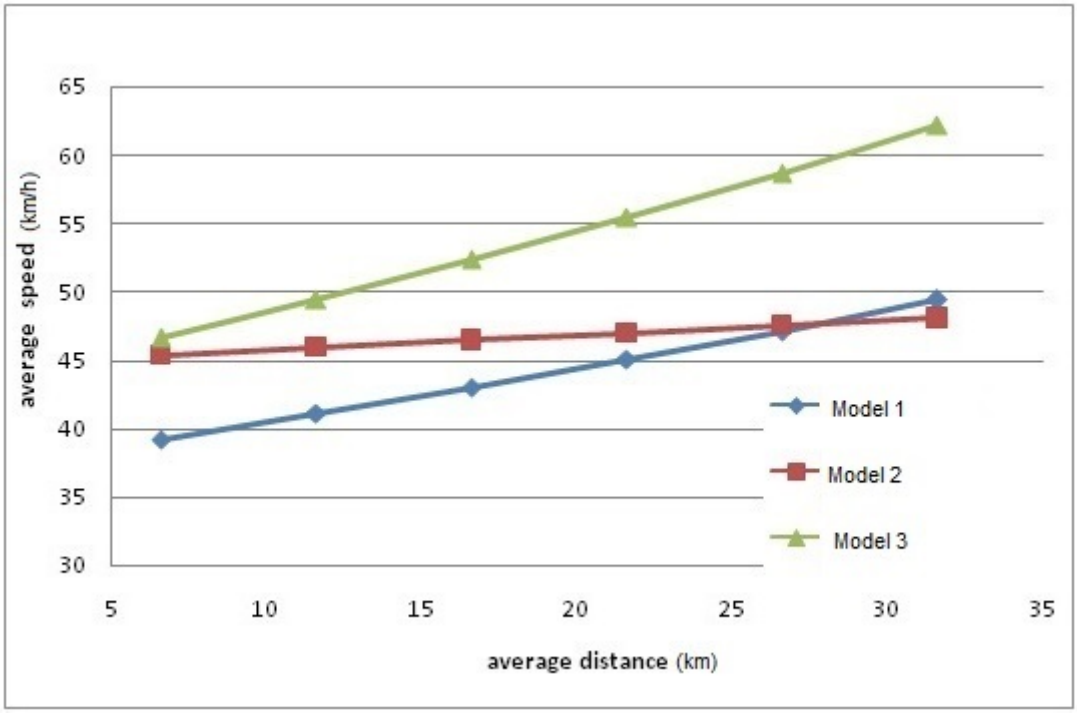

Fig. 5. Average speed to the average distance per driver.

All three models in figure 5 show an increasing trend in relation to the total driving distance. Therefore, it can be inferred that the longer the average driving distance, the higher the average speed. Another conclusion drawn is that model 3 has the highest increasing trend, and model 1 has the second highest whereas model 2 has a significantly lower increasing trend of average speed. The diagram also demonstrates that the model 3 always gives higher values of average speed compared to the other two models. At the same time, model 1 has the lowest speeds for short distances. Because of the higher increasing trend, model 1 has higher average speed for long driving distances than model 2. The model 1 and 2 have the same average speed for an average distance of about 27 km. The above results provide us with a setsquare for explaining the conclusions presented in the next section of the paper.

5. Conclusions

The aim of this paper is to develop driver speed models utilizing detailed driving data collected from smartphone sensors. To this end, six linear regression models (a general model, a model for driving during risky hours and one for during less risky hours and one model for each road type) predicting driver average speed were developed the results of which are discussed in the previous section.

One of the conclusions drawn here is that drivers who tend to accelerate harshly and frequently, also tend to drive at a higher speed. The same seems to apply for deceleration as well, but not to the same extent. In addition, drivers who accelerate more in average also have a higher average deceleration. This is probably because drivers that accelerate more often are usually aggressive and therefore a higher number of harsh braking events is also observed. It is also found that the average acceleration or harsh acceleration events influences driving speed more than the average deceleration or harsh deceleration events. This can be inferred from the fact that all models include an acceleration parameter, while only some of the models include a deceleration parameter. For all models, the coefficient of the acceleration variables is higher than that of the deceleration variables. Furthermore, the standard deviation of the acceleration and the deceleration is found to be a significant predictor of average driving speed. It appears that when large differences in the average acceleration and average deceleration exist between trips, it is likely that driver travels under different conditions and therefore speed changes.

It is also found that drivers traveling more in terms of distance and time, usually drive at a higher speed meaning that they are at higher exposure and behavioural risk. Moreover, the average trip distance appears to play a significant role in predicting the average driving speed, since it is included in all models in most of which, it is found to be the variable with the highest influence.

Regarding harsh driving manoeuvres, they are directly related to harsh speed changes and speed limit exceedance since it is found that the higher the number of harsh manoeuvres performed, the higher the number of speed changes and time of speed limit exceedance. Highways are featuring less harsh cornering events than any other road type and especially urban roads where the relative number is significantly higher. For the majority of the drivers, driving behaviour on rural roads is found to be similar or better to that on urban roads. This is because when driving in urban road, users drive in general more abruptly and their average acceleration is higher. On the other hand, average acceleration rate is much lower in highways where speed is not fluctuating much for longer periods. It is noticed that driving behaviour in rural roads is slightly better (lower average acceleration) than in urban roads.

The time-period of driving appears to affect only vehicle speed and not total driving behaviour. An increase in vehicle speed is observed during risky hours, which is attributed to the lower level of traffic flow. Nonetheless, it cannot be concluded that a driver is more careless or dangerous during risky hours, since the p-value of the risky hours variable is considerably low in the rest of the mathematical models.

Higher differences between duration and driving duration are observed in urban and rural roads but not in highways where such differences are not large for all drivers. This is attributed to the absence of traffic lights in highways and therefore reduced delays. Additionally, drivers use their mobile phone more in urban roads, and much less when driving in highways. It was also observed that mobile phone usage is a significant predictor of the average driving speed but less significant than other variables.

Variables related to the derivatives of acceleration or deceleration rates do not appear to have a significant impact on the predictability of driver’s speed. In all models developed, these variables have particularly low p-value, thus demonstrating an insignificant impact on the average driving speed examined in this research. The maximum and minimum values of each variable also did not appear to predict the average speed.

The investigation of additional dependent and independent variables is suggested for further research. First of all, it would be useful to develop models using the average speed on each road type as a dependent variable, thus taking advantage of more information. In addition, it would be useful to exploit information collected from questionnaires on drivers' traits, years of driving experience, driving habits etc., in order to combine and compare it with the actual driving information collected from smartphones. Moreover, it would be of great importance to conduct an analysis based on other type of data as well, including vehicle characteristics (horsepower, age etc.), road characteristics and road surface condition. Other important factors influencing average speed that should be taken into account in future modelling are the weather and traffic conditions such as traffic volume, speed and traffic density during driving. Moreover, it would be of particular interest to investigate the influence of distraction factors on driver's speed, such as the presence or absence of other persons (particularly children) inside or even outside the vehicle. Because of the fact that driving behaviour changes depending on emotional issues as well, a similar research should be conducted after the participants are examined by specialized psychologists. Overall, the scope of this research needs to be extended by taking into account an ever greater number of different drivers, so that the analyses and the results will concern a broader range of drivers and be more precise.

In order to statistically process collected data and develop the final mathematical models, this study used the multiple linear regression method. In terms of the methodological approach, the examination of additional methods of analysis are proposed, such as factor analysis, logistic regression, etc. Future studies would also benefit from exploiting more advanced technological equipment for recording the in-vehicle driving behaviour such as precise GPS equipment, radars measuring the reaction speed and headways as well as cameras inside and outside the vehicle. Nonetheless, cost is always a significant factor to consider and this is why this study uses a smartphone application for data collection, which is a user-friendly and affordable solution.

Acknowledgements

This research exploited data provided by OSeven Telematics.

References

Aarts, L., & Van Schagen, I. (2006). Driving speed and the risk of road crashes: A review. Accident Analysis & Prevention, 38(2), 215-224.

Cameron, M. H., & Elvik, R. (2010). Nilsson's Power Model connecting speed and road trauma: Applicability by road type and alternative models for urban roads. Accident Analysis & Prevention, 42(6), 1908-1915.

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

Elvik, R. (2009). The Power Model of the relationship between speed and road safety: update and new analyses (No. 1034/2009).

Farmer, C. M., Kirley, B. B., & McCartt, A. T. (2010). Effects of in-vehicle monitoring on the driving behavior of teenagers. Journal of safety research, 41(1), 39-45.

Frantzeskakis, G., & Golias, G. (1994). Road Safety. National Technical University of Athens Editions.

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 behavior of novice young drivers during the first year after licensure. Accident Analysis & Prevention, 42(2), 480-486.

SafetyNet, S. Web Text 2009, retrieved 3/3/2018.

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

Theofilatos, A., & Yannis, G. (2014). A review of the effect of traffic and weather characteristics on road safety. Accident Analysis & Prevention, 72, 244-256.

Toledo, T., Musicant, O., & Lotan, T. (2008). In-vehicle data recorders for monitoring and feedback on drivers' behavior. Transportation Research Part C: Emerging Technologies, 16(3), 320-331.

Tselentis, D. I., Yannis, G., & Vlahogianni, E. I. (2017). Innovative motor insurance schemes: A review of current practices and emerging challenges. Accident Analysis & Prevention, 98, 139-148.

Tselentis, D. I., Theofilatos, A., Yannis, G., & Konstantinopoulos, M. (2018). Public opinion on usage-based motor insurance schemes: A stated preference approach. Travel Behaviour and Society, 11, 111-118.

Tselentis, D. I., Vlahogianni, E. I., & Yannis, G. (2018). Comparative Evaluation of Driving Efficiency Using Smartphone Data. (No. 18-04182). In: Transportation Research Board 97th Annual Meeting, Washington, DC.

Vaiana, R., Iuele, T., Astarita, V., Caruso, M. V., Tassitani, A., Zaffino, C., & Giofrè, V. P. (2014). Driving behavior and traffic safety: an acceleration-based safety evaluation procedure for smartphones. Modern Applied Science, 8(1), 88.

Wang, C., Quddus, M. A., & Ison, S. G. (2012). Factors Affecting Road Safety: A Review and Future Research Direction (No. 12-1583).

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 Local Computer Networks (LCN), 2011 IEEE 36th Conference on (pp. 813-819). IEEE.