Quantifying the Need for Driving Data Collection in Driving Behaviour Assessment Using Smartphone Data

Dimitrios I. Tselentis[a]*, Eleni I. Vlahogianni[a], George Yannis[a], Nectarios Koziris[b]

[a] National Technical University of Athens
School of Civil Engineering
Department of Transportation Planning and Engineering
5, Iroon Polytechniou str., Zografou Campus, GR-15773
Athens, Greece

[b] National Technical University of Athens
School of Electrical and Computer Engineering
Department of Computer Science
5, Iroon Polytechniou str., Zografou Campus, GR-15773
Athens, Greece

Abstract

Despite the fact that nowadays there exist unprecedented opportunities to accurately monitor and analyze driving behaviour, the exact amount of the necessary driving data that should be collected for each driver in driving behaviour assessment is not determined yet. The objective of this paper is to quantify the need for driving data collection in driving performance assessment based on data collected through Smartphone devices from a sample of one hundred and seventy one (171) drivers that participated in the 7-months designed experiment. A statistical analysis was conducted to determine the period at which driving behaviour metrics (speed, mobile usage, harsh acceleration/braking events etc.) rate converges to a stable value. The impact of this methodology lies on the fact that it is essential for researchers nowadays. The impact of this methodology is significant since both small and big data samples lurk the risk of leading to doubtable results, by acquiring a sample either biased or computationally expensive to analyze.

Keywords: Driving Behaviour; Driving Sample; Big Data; Naturalistic Driving Experiment; Smartphone Data

Introduction

Rapid technological progress, especially in Telematics, and Big Data analytics, along with the increase in the information technologies' penetration and use by drivers (e.g. Smartphones), provide unprecedented opportunities to accurately monitor and analyse driving behaviour. First results from related applications (Tselentis et al. (2017), Theofilatos et al. (2017), Araújo et al. (2012), Enev et al. (2016), Vlahogianni and Barmpounakis (2017)) have confirmed the efficiency and usefulness of such big data collection schemes. Nevertheless, the exact amount of the necessary driving data that should be collected for each driver in driving behaviour assessment is not determined yet. Since both small and big data samples lurk the risk of leading to doubtable results, by acquiring a sample either biased or computationally expensive to analyse, it is a matter of great importance to specify exactly how much driving data should be recorded from each participant in the experiment.

Assessing driving behaviour has been has been studied in the past in many research papers (Matthews et al. (1998), Young et al. (2011)). It is extremely significant for road safety experts to identify the parameters that influence driving behaviour and therefore traffic risk. Several studies (Tselentis et al. (2017)) have been carried out regarding mobile phone usage distraction and methodologies for collecting and analysing driving behaviour data but only a few has focused on the required amount of data to be collected (Shichrur et al. (2014)). Among the most common methodologies applied are driving simulators (Desmond et al. (1998)), questionnaires (Gerald et al. (1998)) combined with simulators and

* Corresponding author. Tel.: +30-210-772-2210; fax +30.210.7721454.
E-mail address: dtsel@central.ntua.gr

naturalistic driving experiments (Toledo et al. (2008)), while the most common methods for driving measures recording are the in-vehicle-data-recorders (Shichrur et al. (2014)) and smartphones (Vlahogianni and Barmpounakis (2017)).

A significant number of risk factors that affect the probability of participating in a road traffic accident have been identified in literature. Among others, the most important risk factors recognized in literature (WHO 2015) are human factors such as speeding, distracted driving, driving under the influence of alcohol and other psychoactive substances etc. Human factors are considered one of the main causes of road traffic fatalities and injuries every year and therefore it is highly important to study how these factors can affect traffic risk.

Literature has shown that mobile phone usage has a significant influence on driving behaviour because it causes drivers to drive more slowly with more variation in speed, have higher variation in accelerator pedal position and report a higher level of workload (Md Mazharul and Washington 2015) regardless of conversation difficulty level. Drivers tend to select larger vehicle spacing (Nilsson 1982), longer time headways (Mohammad et al. 2015) suggesting possible risk compensatory behaviour (Md Mazharul and Washington 2015) and their reaction times (Patten et al., 2004) increase significantly when conversing. Speeding is another very important factor that affects accident probability (e.g. reaction distance reduction, loss of control) and crash impact. According to (OECD, 2006) speeding has been a contributory factor in 10% of the total accidents and more than 30% in fatal accidents. According to (Nilsson 1982) the probability of a crash involving an injury is proportional to the square of the speed, the probability of a serious crash is proportional to the cube of the speed and the probability of a fatal crash is related to the fourth power of the speed. Finally, harsh events such as acceleration, breaking and cornering are three significant indicators for driving risk assessment (Bonsall et al. 2005) especially for evaluating driving aggressiveness. These attributes are strongly correlated with unsafe distance from adjacent vehicles, possible near miss accidents, lack of concentration, increased reaction time, poor driving judgement or low level of experience and involvement in situations of high risk. The correlation between HA and HB events with driving risk has been highlighted in the scientific papers published by (Tselentis et al., 2017, Bonsall et al. 2005) and it has been widely recognized by the insurance and telematics industry.

As a result, it would be valuable to determine the amount of data that is necessary to be recorded, in order to obtain a complete picture for each driver, where the rate of those metrics described above per km travelled converges to a stable value. The objective of this paper is to quantify the need for driving data collection in driving performance assessments based on data collected through Smartphone devices. The basis of this framework is an innovative data collection scheme that is continuously recording the driving behaviour analytics of each participant in real time, using smartphone device sensors.

Experimental Data Collection

Data recording and transmission

For the purposes of this study, a mobile App developed by OSeven Telematics is employed to record driving behaviour of the participating users. The application exploits the hardware sensors of the smartphone device and a variety of APIs to read sensor data and transmit it to a central database. Recorded data come from various smartphone sensors and data fusion algorithms provided by Android (Google) and iOS (Apple). The frequency of the data recording varies depending on the type of the sensor with a minimum value of 1Hz. It should be noted that all procedures described below are implemented by OSeven and they do not constitute part of the work done within this research.

All recorded data are transmitted to the OSeven central database the end of each trip where data is stored in the cloud server for central processing and data reduction. As a result of big data handling and processing data is converted into meaningful behavioral and safety related indicators. This is achieved by using the two Big data processing methods which include two families of techniques, Big Data mining techniques and Machine Learning (ML) algorithms.

Machine learning methods are applied for data cleaning from existing noise and errors, and for identification of repeating patterns in the data. The methods applied allow for data filtering and outlier detection, data smoothening, speeding regions, harsh acceleration events, harsh braking events, harsh cornering events, mobile usage, risky hours driving and driver or passenger recognition. Subsequently, these data patterns are processed by means of big data mining techniques, to calculate the necessary parameters and derive behaviour indicators to be used in the analysis.

The indicators derived from the ML process are divided into two categories, risk exposure and driving behaviour indicators. The main risk exposure indicators are total distance travelled, driving duration, type(s) of the road network used, time of the day driving, while the main driving behaviour indicators are speeding over the speed limit, mobile phone usage, number and severity of harsh events such as harsh braking, harsh acceleration and harsh cornering.

Aggregated Data are analyzed and filtered to retain only those indicators that will be used for the analysis conducted herein. Data filtering and analysis is performed in Python programming language and several scripts are written for this reason. More details on the algorithmic implementation are given below.

Experiment design

One hundred and seventy one (171) drivers participated in the designed experiment that endured 7-months and a large database of 49,722 trips is created. All drivers chosen to be included in the analysis should had driven at least for 10 hours and 40 trips that approximately equals the typical monthly number of trips for a driver assuming that each driver drives 2 trips of 15 minutes a day for 5 working days a week.

The indicators per trip travelled that were used in this study are the number of harsh acceleration events, the number of harsh braking events, mobile usage and speeding. The definition of these indicators is given below:

- Driving Duration: number of driving hours including stops.
- Harsh Acceleration: positive longitudinal acceleration (m/s2).
- Harsh Braking: negative longitudinal acceleration (m/s2).
- Mobile Phone Usage: seconds of mobile phone usage while driving.
- Speeding: seconds of driving over the speed limit.

As illustrated in figure 1, driving duration and distance covered by drivers is normally distributed within the sample taken.

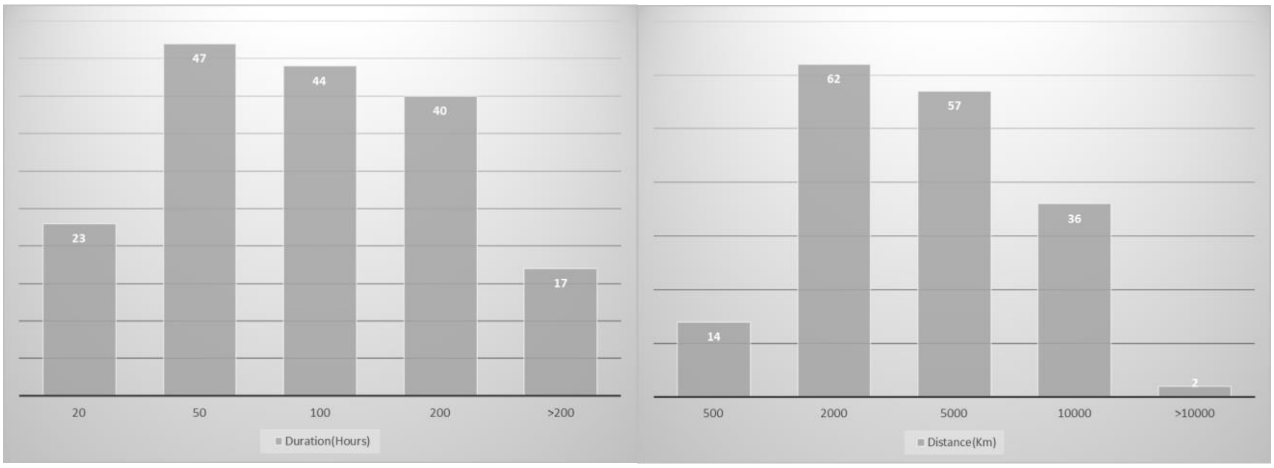

Fig. 1 (a) Histogram of driving duration for the sample of drivers used; (b) Histogram of driving distance for the sample of drivers used

It appears that the majority of the drivers (131 out of 171) have driven between 20 and 200 hours, while only a few spent more than 17 hours on road. Regarding distance covered, 155 drivers drove between 500 and 10,000 km. Only two drivers demonstrated a mileage of more than 10,000km.

It should be highlighted that the required amount of driving data will be quantified in driving time and not in driving distance in order for the results to be comparable with other past studies (Shichrur et al. (2014)).

Method

The statistical analysis using the data collected from the smartphone was conducted to determine the driving time at which the rate of the driving indicators converges to a stable index and therefore no more data are required to be collected. In this research the magnitude of change measurement in a time-series is employed which is decreasing over time as the specific magnitude converges on its final rate. At the same time, this means that the event rate also converges to its final event rate i.e. the average rate of events for the specific driver. For each driver, the above metrics were calculated by diving the total number of harsh events by the total driving hours until that time after every trip that took place thus constructing the time-series of average harsh events per hour.

Convergence Index

Assuming that we have calculated the time-series of the harsh events per hour (HE/hr) for trips, the following formula will be used for calculating the convergence index of harsh events:

$$|(HE/hr_i - HE/hr_{i-1})/HE/hr_{i-1}| \qquad \forall i \in N^* \ in\ [2, n] \tag{1}$$

A moving average of this magnitude is calculated taking into account the threshold set above for including a participant in the analysis. The time window considered is 40 trips and it is deemed to be within acceptable margins when the average change measurement for the time window of 40 trips and for at least 10 driving hours is less or equal to 5% (0.05). As stated above, the reason of choosing 40 trips and 10 driving hours is that all drivers included in the analysis should had driven at least 40 trips that approximately equals the typical monthly number of trips for a driver assuming that each driver drives 2 trips of 15 minutes a day for 5 working days a week.

The value 5% reflects the per mille change in measured change of the respective harsh event rate i.e. the moving average is attempting to capture the time after which the average per mille change is steadily less than 0.05. The reason why this value was selected can be better explained using figure 2.

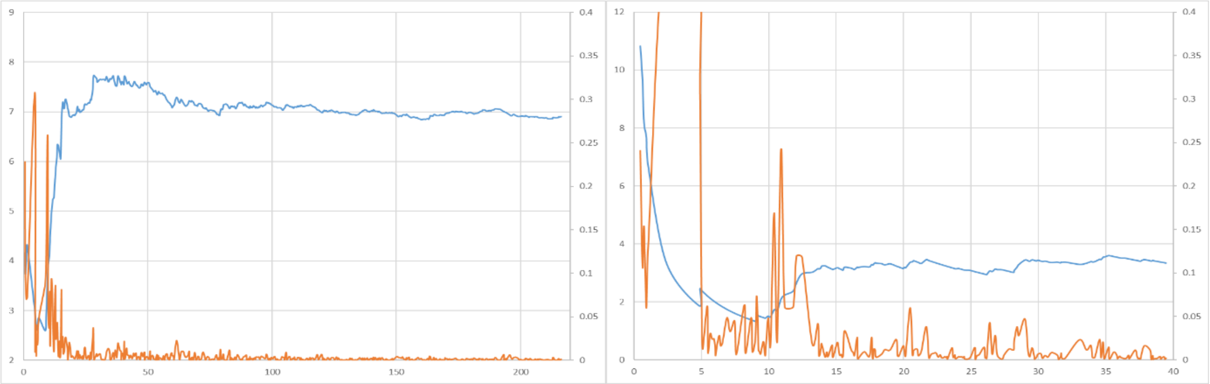

Fig. 2 Average Harsh Acceleration Rate and Convergence Index in a (a) converged rate; (b) non converged rate

Figure 2 demonstrates the average harsh event per hour rate and the convergence index of two individual drivers a and b of total driving durations, 216 and 39 hours respectively. The blue line refers to the harsh acceleration events rate per hour and the orange line to the convergence index. Respectively, the blue line is plotted based on the values of the primary Y axis while the orange to the secondary Y axis. It is evident that in figure 2a the driver's behaviour is gradually converging after approximately 100hrs of driving as the average rate of harsh events per hour and the average convergence index are not significantly altered from that time and after. On the other hand, it can be said from figure 2b that the driver's convergence index and rate of harsh acceleration events is not converged since the average harsh event rate is fluctuating and the convergence index is significantly increased over 0. The same analysis was conducted for all drivers and the optimum value for average measurement change was found to be 0.05. This value was secondarily chosen because it can be considered a secure low per mille change to draw statistically significant conclusions. The results of the analysis conducted are presented in the following section.

Results

In figure 3, the number of necessary driving hours for the examined metrics to converge are plotted with the respective metric value in order to observe any obvious trend or correlation between these two magnitudes. To begin with, the number of seconds of mobile usage and the number of harsh braking events per driving hour show an apparent decreasing trend which appears to be more steep for mobile usage rather than for harsh braking events. Regarding the seconds of speeding and the number of harsh acceleration events, no clear conclusion can be drawn from the graph. There is no obvious trend for these two metrics which are probably better represented by a constant.

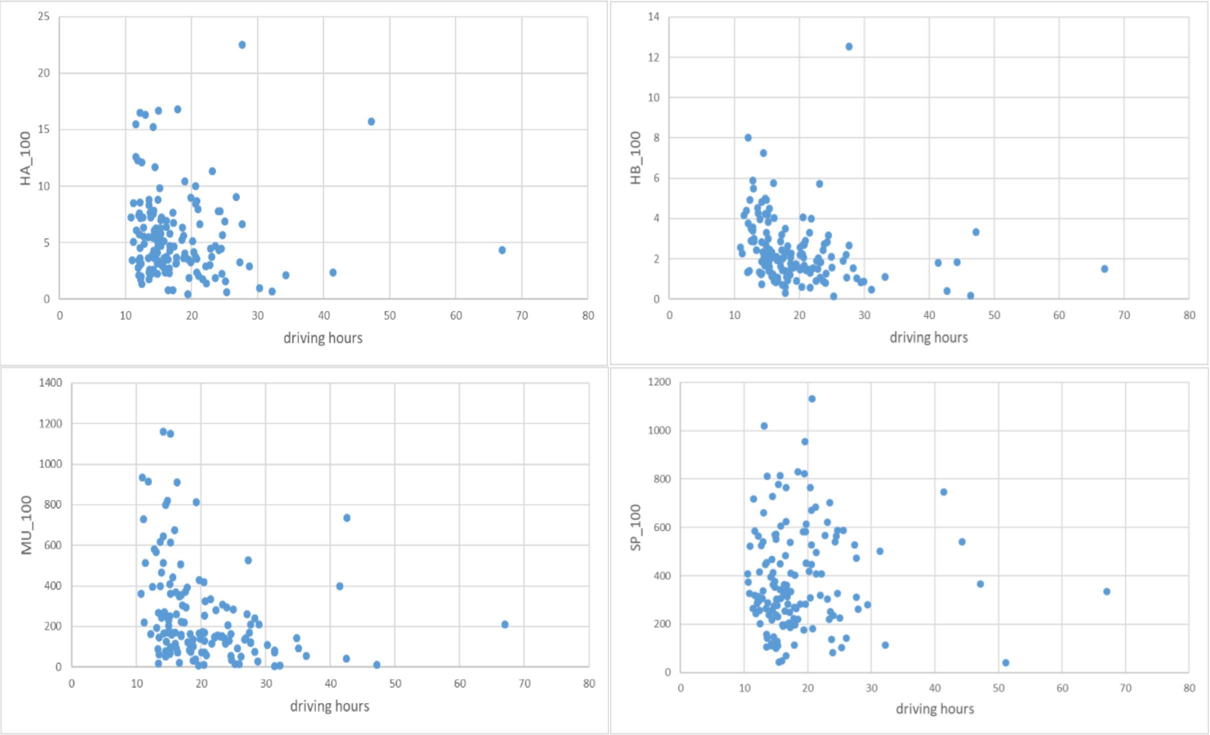

Fig. 3 Necessary driving hours for each of the following metrics to converge: (a) Number of harsh acceleration events, (b) Number of harsh braking events, (c) Seconds of mobile usage and (d) Seconds of driving over the speed limit

For each metric, the procedure described above was implemented to find whether a driver's behaviour is converged or not and after how many driving hours. For the total sample of converged drivers, the analysis per metric conduced per four different value categories, which were defined by the three percentiles of 25%, 50% and 75% of the converged sample of drivers. This categorization is implemented to enable the investigation of necessary recording time for drivers of different levels of aggressiveness 1, 2, 3 and 4. The results of each metric and category are presented in table 1. It is clear that different sampling periods are required in most cases for more and less aggressive drivers in order to avoid unexpected errors due to undersized data sample. This is probably because drivers of lower aggressiveness 1 presented a higher standard deviation as appears in figure 2 and therefore it takes more time for their driving characteristics to converge to their average level.

As for the characteristics per driving aggressiveness level, more aggressive drivers perform 10.7 and 4.5 harsh acceleration and braking events per driving hour on average whereas less aggressive drivers perform less than 2 and 0.9 harsh acceleration and braking events per driving hour respectively. On the other hand, less aggressive drivers use their mobile phone approximately for 0.8 min/ hour and overspeed for 2.6 min/ hour. Finally, aggressive drivers demonstrated an overspeed of 11.4 min/ hour on average and a 10 min/ hour mobile usage.

Table 1. Average and standard deviation of metric value and duration per aggressiveness category.

	Metric		Duration to convergence (hr)		Limits	
	Average	StDev	Average	StDev	Upper	Lower
HA	2.04	0.74	19.47	6.99	-	3.13 ha/ hour
	3.79	0.46	18.57	8.8	3.13 ha/ hour	4.63 ha/ hour
	5.84	0.72	16.29	3.85	4.63 ha/ hour	7.21 ha/ hour
	10.66	3.73	17.06	6.81	7.21 ha/ hour	-
HB	0.91	0.32	21.82	7.71	-	1.36 hb/ hour
	1.65	0.19	21.54	9.98	1.36 hb/ hour	1.99 hb/ hour
	2.44	0.28	18.24	4.28	1.99 hb/ hour	2.9 hb/ hour
	4.54	1.77	16.49	6.44	2.9 hb/ hour	-
MU	50.49	27.54	23.01	8.17	-	90.76 sec_mu/ hour
	133.05	21.58	21.4	5.44	90.76 sec_mu/ hour	165.56 sec_mu/ hour
	239.44	51.23	20.49	9.39	165.56 sec_mu/ hour	347.93 sec_mu/ hour
	597.08	224.84	16.84	7.04	347.93 sec_mu/ hour	-
SP	154.7	58.42	19.01	7.03	-	238.86 sec_sp/ hour
	287.71	26.55	16.89	4.9	238.86 sec_sp/ hour	335.4 sec_sp/ hour
	422.78	57.9	19.39	10.54	335.4 sec_sp/ hour	533.03 sec_sp/ hour
	681.65	142.02	19.38	6.96	533.03 sec_sp/ hour	-

According to table 1, mobile usage appears to be the most critical metric in determining the required sampling time for data analysis for less aggressive drivers since the adequate period for the relevant metric to converge seems to be the maximum value of the table. The relevant critical recording period for aggressive drivers is found to be determined by speeding. The least required time for convergence is demonstrated for the number of harsh acceleration events on average followed by the seconds of speeding and the number of harsh braking events, which is slightly less than that required for the number of speeding to converge.

The results of table 1 confirm the conclusions drawn from figure 3 regarding the decreasing trend of number of harsh braking events and seconds of mobile phone usage since the required driving time for convergence is increasing as the value of the final converged metric is decreasing. This does not apply for the driving metrics of the number of harsh acceleration events and seconds of speeding, which appear to remain stable for every driving aggressiveness level. Concluding, according to the aforesaid and table 1 the required sampling time for aggressive drivers is found to be approximately 23 hours for normal drivers and 19.4 for less aggressive.

Conclusions

As technology advances radically, especially in Telematics and Big Data analytics, and information technologies penetrate in the market by drivers (e.g. Smartphones), unprecedented opportunities to accurately monitor and analyze driving behaviour are emerging. The efficiency and usefulness of such big data collection schemes is confirmed by related applications. The exact amount of the required driving data in order to assess driving behaviour is not determined yet though, which is the outcome of this research i.e. to quantify the need for driving data collection through Smartphone devices. As stated above, this need emerges from the fact that collecting either more or less driving data can be risky because it might lead to non-significant conclusions or to excessive computational effort when it comes to large-scale data.

This framework is based on an innovative data collection scheme, which is continuously recording driving behaviour analytics in real time, using smartphone device sensors. One hundred and seventy one (171) drivers participated to the designed experiment during a 7-months timeframe and a large database of 49,722 trips is acquired. Driving behaviour variables collected include distance travelled, speed, accelerations, braking, cornering and smartphone usage. The driving period at which driving behaviour metrics rate converges to a stable value is determined by the statistical analysis that was conducted herein. Results indicated that the critical metrics for less and more aggressive drivers are mobile usage and speeding respectively and that sampling time should approximately 23 and 19.4 hours in order for the driving data sample collected to be deemed as adequate. It is highlighted that despite the fact that the number of harsh braking events and seconds of mobile usage is decreasing while required time for convergence is increased, the same does not seem to apply for the number of harsh acceleration and seconds of speeding as well. For future research, a larger sample is recommended to be exploited to acquire a clearer picture of the relationship between the number of harsh acceleration events and seconds of speeding and the total driving duration at which those metrics converge. The impact of this methodology lies on the fact that it is essential for researchers nowadays to allocate available resources efficiently due to the emerging necessity of expensive big data computations.

Acknowledgements

This research has exploited data provided by OSeven Telematics, London, UK.

References

In Intelligent Vehicles Symposium (IV), 2012 IEEE (pp. 1005-1010). IEEE.

Bonsall, Peter, Ronghui Liu, and William Young. Modelling safety-related driving behaviour—impact of parameter values. Transportation Research Part A: Policy and Practice Vol. 39, No. 5, 2005, pp. 425-444.

Desmond, Paula, Peter Hancock, and Janelle Monette. Fatigue and automation-induced impairments in simulated driving performance. Transportation Research Record: Journal of the Transportation Research Board No. 1628, 1998, pp. 8-14.

Enev M., Takakuwa A., Koscher K, and Kohno T. (2016) Automobile Driver Fingerprinting. Proceedings on Privacy Enhancing Technologies (1):34-51

Mazharul Haque Md, and Simon Washington. The impact of mobile phone distraction on the braking behaviour of young drivers: a hazard-based duration model. Transportation research part C: emerging technologies Vol. 50, 2015, pp. 13-27.

Matthews, Gerald, Lisa Dorn, Thomas W. Hoyes, D. Roy Davies, A. Ian Glendon, and Ray G. Taylor. Driver stress and performance on a driving simulator. Human Factors Vol. 40, No. 1, 1998, pp. 136-149.

Nilsson G. The effects of speed limits on traffic accidents in Sweden. Sartryck, Swedish National Road and Transport Research Institute, 1982.

Patten, Christopher JD, Albert Kircher, Joakim Östlund, and Lena Nilsson. Using mobile telephones: cognitive workload and attention resource allocation. Accident analysis & prevention Vol. 36, No. 3, 2004, pp. 341-350.

Saifuzzaman, Mohammad, Md Mazharul Haque, Zuduo Zheng, and Simon Washington. Impact of mobile phone use on car-following behaviour of young drivers. Accident Analysis & Prevention Vol. 82, 2015, pp. 10-19.

Shichrur, R., Sarid, A., & Ratzon, N. Z. (2014). Determining the sampling time frame for in-vehicle data recorder measurement in assessing drivers. Transportation research part C: emerging technologies, 42, 99-106.

Speed management. Paris, OECD/ECMT Transport Research Centre (JTRC), 2006

Theofilatos A., Tselentis. I.D., Yannis. G., Konstantinopoulos M. (2017) ”Willingness – to - Pay for Usage-Based Motor Insurance”, Proceedings of the 96th Annual meeting of the Transportation Research Board, Washington, D.C, January 8-12, 2017.

Toledo, Tomer, Oren Musicant, and Tsippy Lotan. In-vehicle data recorders for monitoring and feedback on drivers’ behavior. Transportation Research Part C: Emerging Technologies Vol. 16, No. 3, 2008, pp. 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.

Vlahogianni, Eleni I., and Emmanouil N. Barmpounakis. Driving analytics using smartphones: Algorithms, comparisons and challenges. Transportation Research Part C: Emerging Technologies Vol. 79, 2017, pp. 196-206.

Young, Mark S., Stewart A. Birrell, and Neville A. Stanton. Safe driving in a green world: A review of driver performance benchmarks and technologies to support 'smart'driving. Applied ergonomics Vol. 42, No. 4, 2011, pp. 533-539.