Accident Analysis and Prevention 130 (2019) 151–159

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Impact of real-time traffic characteristics on crash occurrence: Preliminary results of the case of rare events

Athanasios Theofilatos[a,*], George Yannis[a], Pantelis Kopelias[b], Fanis Papadimitriou[c]

[a] *National Technical University of Athens, Department of Transportation Planning and Engineering, 5 Heroon Polytechniou str., Athens, GR15773, Greece*
[b] *University of Thessaly, Department of Civil Engineering, Pedion Areos, Volos, GR38334, Greece*
[c] *Attica Tollway Operations Authority – Attikes Diadromes S.A., 41.9 km Attiki Odos Motorway, Paiania, GR19002, Greece*

ARTICLE INFO

Keywords:
Crash occurrence
Rare events
Real-time traffic
Logistic regression

ABSTRACT

Considerable efforts have been made from researchers and policy makers in order to explain road crash occurrence and improve road safety performance of highways. However, there are cases when crashes are so few that they could be considered as rare events. In such cases, the binary dependent variable is characterized by dozens to thousands of times fewer events (crashes) than non-events (non-crashes). This paper attempts to add to the current knowledge by investigating crash likelihood by utilizing real-time traffic data and by proposing a framework driven by appropriate statistical models (Bias Correction and Firth method) in order to overcome the problems that arise when the number of crashes is very low. Under this approach instead of using traditional logistic regression methods, crashes are considered as rare events In order to demonstrate this approach, traffic data were collected from three random loop detectors in the Attica Tollway ("Attiki Odos") located in Greater Athens Area in Greece for the 2008–2011 period. The traffic dataset consists of hourly aggregated traffic data such as flow, occupancy, mean time speed and percentage of trucks in traffic. This study demonstrates the application and findings of our approach and revealed a negative relationship between crash occurrence and speed in crash locations. The method and findings of the study attempt to provide insights on the mechanism of crash occurrence and also to overcome data considerations for the first time in safety evaluation of motorways.

1. Introduction

Crashes impose serious problems to society in terms of human costs, economic costs, property damage costs and medical costs. Understanding the various factors that influence crash occurrence is of particular concern to decision makers and researchers. Most of the existing studies that aimed to link traffic parameters and road safety used aggregate traffic data[1] and other traffic proxies for congestion (Ceder 1982; Caliendo et al., 2007; Martin 2002; Dickerson et al. 2000; Chang 2005; Anastasopoulos and Mannering, 2009; Noland and Quddus 2005; Wang et al. 2009; Lord et al. 2005; Kononov et al. 2012). However, aggregated traffic parameters such as AADT (Annual Average Daily Traffic) may be too aggregated to be directly linked with crashes.

Other studies used micro-simulation (specified or calibrated traffic models) or video analyses (experiments through observational vehicle tracking data) in order to obtain crash prone parameters. Safety performance is affected by various factors (behavioral, road characteristics, vehicle attributes, traffic parameters, environmental conditions) (Elvik et al. 2009; Evans, 1991; Guido et al. 2011). These two sources of data are very important in road safety research as they can capture pre-crash conditions, however they are not in the scope of the present paper. For more details the reader is encouraged to refer to Guido et al. (2011) who provide a detailed comparison between safety performance measures from micro-simulation and observational data.

Recently, the incorporation of real-time traffic and weather data in freeways has proven to be a promising approach, since an increasing number of researchers use disaggregate traffic and weather data (i.e. in 5-min intervals) to analyze crash occurrences in freeways. Ahmed et al. (2012a) found that increased speed variation at any given crash segment combined with a decrease in average speed in the respective downstream segment can lead to increased likelihood of rear-end crash occurrence. Ahmed and Abdel-Aty (2012) found that the probability of a crash increases when the variation in speed increases and the average speed decreases at the crash segment 5–10 min prior to crash occurrence. On the other hand, Kockelman and Ma (2007) have indicated that 30-sec speed changes are not correlated with risk of crashes.

Studies investigating crash occurrence often include real-time traffic as well as real-time weather characteristics. Ahmed et al. (2012b)

[*] Corresponding author.
E-mail address: atheofil@central.ntua.gr (A. Theofilatos).

[1] For a more detailed list of relevant studies the reader is encouraged to refer to Theofilatos and Yannis (2014).

https://doi.org/10.1016/j.aap.2017.12.018
Received 16 May 2016; Received in revised form 15 November 2017; Accepted 26 December 2017
Available online 05 January 2018

investigated the impact of geometrical, traffic and weather variables on crash occurrence in freeways. In winter, it was found that low visibility, high precipitation and speed variation seem to increase the likelihood of crashes. Surprisingly, dry season, low average speeds and low visibility increase the odds of a crash. Xu et al. (2013a) developed separate models for clear, rainy and low visibility weather and found that speed difference between upstream and downstream detectors increased crash risk for low visibility and/or rainy weather. Rainfall intensity and occupancy variations from the closest downstream detector were also found to be positively correlated with crash risk during adverse weather. During good weather conditions, the standard deviation of speed was the main risk factor. Similarly, Xu et al. (2013b) state that various traffic parameters, such as traffic density upstream, speed variance upstream and/or downstream, traffic volume difference between upstream and downstream station and occupancy difference between upstream and downstream stations are positively correlated with crash risk.

Zheng et al. (2010) investigated the impact of stop and go driving (traffic oscillations) on crash occurrence. It was found that the standard deviation of speed increased the likelihood of rear-end crashes. Yu et al. (2013) found that the 5-min average speed of the crash segment 5–10 min prior to the crash time to significantly influenced crash risk. It is interesting that the authors suggest a negative correlation, which means that crash occurrence likelihood increases as the average speed decreases. Xu et al. (2012) developed crash risk models for different traffic states. Traffic flow parameters were found to have different effects on safety for every traffic state. For instance, the average downstream occupancy seemed to reduce crash risk in two traffic states (in congested traffic as well as in transition from free flow to congested flow) but caused an increase in the overall model.

Hossain and Muromachi (2013) aimed to identify crash predictors on urban expressways. One essential contribution is that crash risk in freeway segments and ramp vicinities were analyzed separately. The findings suggest that crash mechanisms are not the same for basic freeway segments and ramps. Wang et al. (2015) developed real-time crash prediction models for expressway weaving segments. The authors concluded that the mainline speed at the beginning of the weaving segments, the speed difference between the beginning and the end of weaving segment and the logarithm of traffic volume all have a significant influence on crash risk.

In terms of methodology when dealing with real-time crash likelihood models, disaggregate data are used and crash occurrence is analyzed as a binary variable having two outcomes, namely crash and non-crash (Abdel-Aty and Pande, 2005; Abdel-Aty et al., 2007; Ahmed and Abdel-Aty, 2012; Yu and Abdel-Aty, 2013). As a consequence, binary logistic models are usually applied (Ahmed et al. 2012a and 2012b; Theofilatos 2017; Xu et al., 2012; Xu et al. 2013a and 2013b; Yu and Abdel-Aty, 2013). The model specification is frequently Bayesian or random effects. Other non-traditional approaches have also been followed. For instance, Yu and Abdel-Aty (2013) applied Support Vector Machine (SVM) models, whilst Pande and Abdel-Aty, 2006a, 2006b) applied Neural Networks (NN). These models usually performed very well but they are "black box" techniques (i.e. it is difficult to obtain relationships between the dependent and the independent variables).

Under this methodological framework, the case-control ratio usually varies from 1:1 to 1:5 (Roshandel et al., 2015). In other words, for each crash case, 1–5 non-crash cases were selected. For instance, Yu and Abdel-Aty (2013) and Zheng et al. (2010) used a 1:4 ratio, while Ahmed et al. (2012b) analyzed 301 crash cases and 880 non-crash cases (a roughly 1:3 ratio). Larger crash to non-crash ratios are also used but only rarely. Wang et al (2015) used a dataset including 125 crash and 1250 randomly selected non-crash cases (1:10 ratio). Xu et al. (2013a) also utilized a 1:10 ratio, while Xu et al. (2013b) used 794 crash cases and 15,880 non-crash cases (1:20 ratio). Roshandel et al. (2015) illustrate a discussion of the case-control design problems in real-time crash occurrence evaluation. In that context, the case-control design methods may result in loss of valuable information. This happens because in reality the number of non-crash cases greatly is much higher than crash cases and crashes might be considered as very rare events, as stated previously. In addition, when crash risk is investigated in specific segments or freeway exits or in rural areas and specific minor roads, a very low number of crashes is likely to exist. Therefore, when the number of crashes is particularly low, the traditional control-case-design and the traditional statistical methods may not be appropriate. Consequently, alternative data collection methods and statistical approaches should be explored[2].

For that reason, the aim of this study is add to the current knowledge by extending the investigation of crash occurrence with real-time traffic characteristics when the number of crashes is so low that they can be considered as rare events. Moreover, to the best of the authors' knowledge, this one of the very first times that the case of rare events is examined in real-time safety evaluation of freeways, as only one study has applied a rare events approach in workzones (Yang et al., 2015). The selected approach deals explicitly with the prevailing traffic conditions at the time and location of each recorded crash and non-crash case. Validity is improved by avoiding the use of a random sample of non-crash cases as controls and losing valuable information. Thus, alternative methods which are directly linked to data sampling are utilized. More specifically, our approach improves data collection and analysis methods by proposing: a) a collection of a 1:10 ratio of crashes to non-crash cases through stratified sampling and applying the bias correction method or b) the Firth method application when all the existing crash and non-crash cases are available. Therefore, this approach is considered novel and adds to current knowledge as it offers a holistic method on how to deal with the case of rare events in crash likelihood modeling.

The paper is structured as follows: In Section 2, the proposed framework with alternative statistical methods to deal with rare events is demonstrated, namely, the bias correction method and the Firth method (also known as penalized likelihood estimation method). Section 3 presents the data collection and data preparation for the case study of Attica Tollway, which is a modern urban freeway in the area of Athens. Section 4 presents the main findings of the study. The last section is dedicated to conclusions.

2. Methodology

2.1. Overview of the problem

The issues of data preparation and model development when dealing with real-time safety evaluation and crash likelihood modeling have received increasing attention by researchers (Imprialou et al., 2016a; Roshandel et al., 2015). Roshandel et al. (2015) mention the problems arising from the case–control design, in which the case–control ratio varies from 1:1 to 1:5. Authors state that "*using a 1:1 case–control design will bias false positive rates to be really low, because in practice there are far more 'real' controls then the one used in the study*". It is also mentioned that there is an absence of reporting this bias from the relevant literature. Therefore, there is a need to resort to different data collection strategies in order to set up a case-control study design. However, the selected data collection strategies will largely affect model development.

In reality, most significant events are rare events. They occur very rarely-meaning there are dozens to thousands of times fewer events (e.g. wars, volcano explosions) than non-events. Leitgöb (2013) suggested that in the case of rare events, maximum likelihood estimates of logistic regression may be biased. The authors compared three common methods dealing with rare events, namely Bias Correction method, Penalized Maximum Likelihood Estimation (also called Firth method) and Exact logistic regression. Exact logistic regression only works when the number of events is very low (< 200). Moreover, independent variables should be

[2] Similar methods exist when modeling count data and there is an excessive number of zeros (e.g. zero inflated Poisson or zero inflated Negative binomial models).

restricted to low numbers and also be dichotomous or discrete. Another drawback is that this method is computationally expensive.

As stated earlier, the choice of the most appropriate statistical method depends heavily on the sampling of the study. If the researcher is able to gather the entire sample of non-events, then the Firth or Exact logistic regression methods are considered to be the most appropriate. On the other hand, the Bias Correction method applies a number of mathematical corrections in the model and thus enables the researcher to gather the total sample of all events (i.e. crashes) but only a sample of non-events. Consequently, valuable time and effort are saved. For these reasons, only the Bias Correction and the Firth method are presented and applied in this paper under the assumption that crashes in the Attica Tollway are rare events, and that the term “event” corresponds to an occurrence of a crash. The next two subsections provide a theoretical background of these two methods.

2.2. Bias correction method

King and Zeng (2001a) identified two major causes for problems when analyzing rare events. Firstly, the fact that traditional statistical procedures underestimate the probability of rare events, and secondly the inefficient data-collection strategies, because too much time and effort are needed to record all non-events. Furthermore, serious problems arise due to the fact that maximum likelihood estimation of the logistic model suffers from small-sample bias, with the degree of bias being strongly dependent on the number of cases in the less frequent of the two categories of the dependent variable y. For example, even with a sample size of 100,000 cases, if there are only 20 events in the sample, substantial bias exists. Consequently, scholars cannot confidently rely on logit coefficients. To solve these problems, King and Zeng (2001a,b) proposed an adapted version of the logistic regression, the so-called Bias Correction method. This approach applies a number of corrections. The first suggested correction concerns data collection. King and Zeng (2001a,b) propose a case-control sampling design, based on stratified sampling, where it is recommended to include all events and a random selection of non-events. Then, in order to account for the biased estimation of constant term due to the case-control design, a prior correction has to be applied to the constant term. The next equation applies the correction:

$$\alpha_0 = \hat{\alpha} - \ln[(\frac{1-\tau}{\tau})*(\frac{1-\gamma}{\gamma})], \tag{1}$$

where α_0 is the new corrected constant, $\hat{\alpha}$ is the uncorrected constant, τ is the proportion of events in the population and γ is the proportion of events in the sample. Another method proposed for correction is the “weighting” method, which was not used in this study and thus not described here. Moreover, the underestimation of probabilities when using the corrected intercept α_0 needs a similar correction. For that reason, a correction factor C_i is added to the estimated probability p_i. If we assume the corrected logit form based on the corrected constant term:

$$\log it p_i = \ln(\frac{1-p_i}{p_i}) = a_0 + \sum \hat{\beta}_i x_i, \tag{2}$$

then,

$$p_i' = p_i + C_i, \tag{3}$$

where Ci is calculated according to King and Zeng (2001b):

$$C_i = (0.5 - p_i)*p_i*(1-p_i)*x_0*V(\beta)*x_0' \tag{4}$$

where p_i is the probability of an event estimated using the corrected estimated coefficient a_0, x_0 is the 1 × (m + 1)[3] vector of values for each independent variable, V(β) is the variance-covariance matrix, and lastly x_0' is the x_0 transposed.

2.3. Penalized maximum likelihood estimation (firth method)

In order to reduce bias in generalized linear models, Firth (1993) suggested a modification of the score equations. The principle is to extend the elements of the score vector by a penalization term. It is therefore suggested to maximize the penalized log likelihood:

$$LogL(\beta)^* = LogL(\beta) + \frac{1}{2} * \log\left|I(\beta)\right| \tag{5}$$

where, *I(β)* is the Fisher information matrix. Extending this to logistic regression, the score function *U(β)* is replaced by the new modified score function:

$$U(\beta)^* = U(\beta) + a \tag{6}$$

where, α has rth entry:

$$a_r = 0.5trI(\beta)^{-1}\left[\frac{dI(\beta)}{d\beta_r}\right] \tag{7}$$

and r = 1,…,k. The explicit formulas can be found in Heinze and Schemper (2002). For more details the reader is referred to Heinze and Ploner (2003) and Heinze et al. (2015) who provide detailed descriptions of the Firth method.

3. Data collection and preparation

In this study, the required crash and traffic data were extracted from Attica Tollway (“Attiki Odos”) located in Greater Athens Area in Greece in order to demonstrate the applications of the statistical methods for rare events.

Attica Tollway is a modern motorway extending along 65.2 km (Fig. 1). It constitutes the ring road of the greater metropolitan area of Athens and the backbone of the road network of the whole Attica Prefecture. It is essentially a closed toll motorway, within a metropolitan capital, where the problem of traffic congestion is acute. Entry to the freeway is through 39 toll plazas with 195 toll lanes. Inductive loop detectors are placed every 500 m inside the asphalt pavement of the open sections of the motorway and every 60 meters inside tunnels, providing information regarding the volume, speed and density of traffic.

Two datasets were prepared for the need of the analysis; one dataset with crash data and one with traffic data. The required crash data for Attica Tollway were extracted from the Greek crash database SANTRA provided by the Department of Transportation Planning and Engineering of the National Technical University of Athens.

Traffic data for the Attica Tollway were extracted after a close collaboration with the Traffic Management and Motorway Maintenance, which is located in Paiania and operates on a 24-hour basis. The complete traffic time series measured in 1-hour intervals from 2008–2011 in three random loop detectors in BFS areas with the same number of lanes (3 lanes per direction) were considered. The aim was to demonstrate the validity of the approach when there are very few crashes. Consequently, if a wider range of data was explored, traditional methods would apply and there would be no need for this alternative framework; for instance, if there were 200 crashes, a traditional logistic regression method would be applicable.

Traffic variables were measured in 1-hour intervals (flow, speed, occupancy and truck proportion). Traffic flow is defined as the total number of vehicles on a 1 h basis and is measured in vehicles per hour (veh/h). On the other hand, speed (km/h), occupancy (%) and truck proportion (%) are the averaged values of 5-min interval measurement, which were automatically aggregated in 1-hour intervals.

Crash occurrence was defined as a binary variable taking the values of 0 (non-crash) and 1 (crash). Therefore, in each 1-h time interval it represents the information of whether a crash has occurred or not. In order to avoid post-crash traffic conditions where low mean speeds may

[3] It is noted that m is the number of independent variables

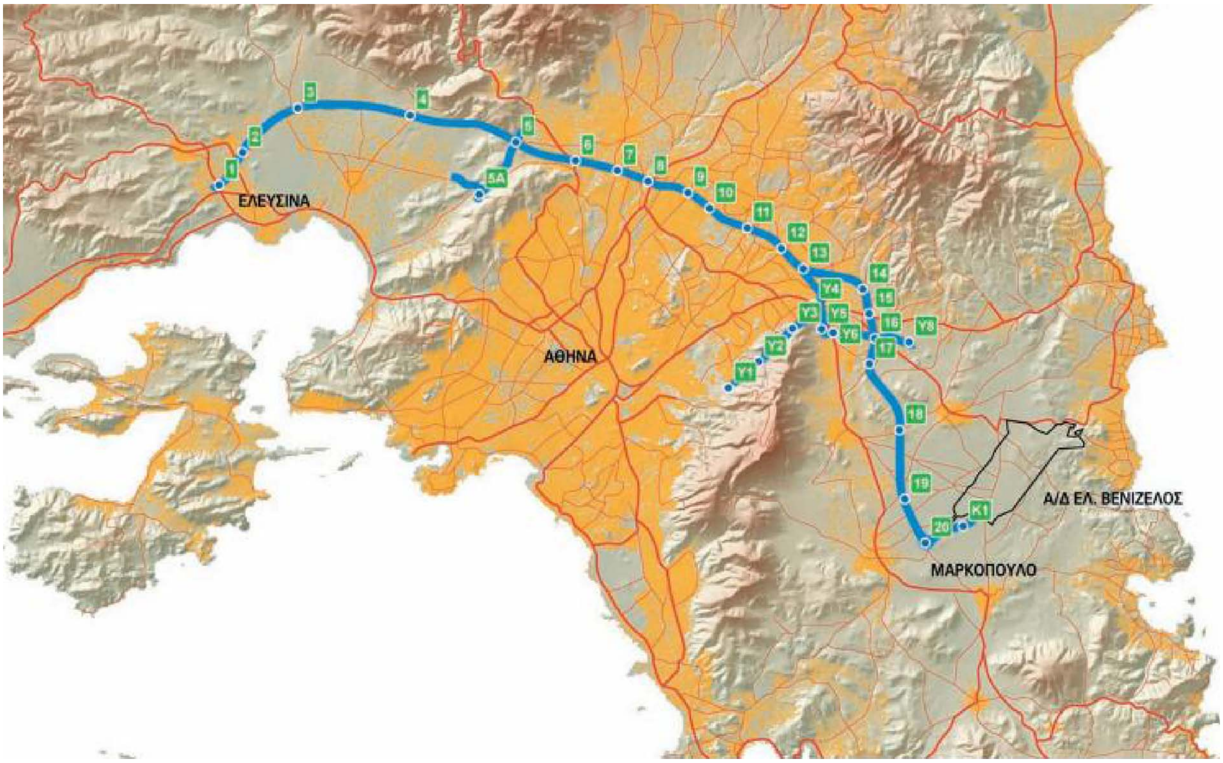

Fig. 1. Map of Attika Tollway.

Table 1
Summary statistics of independent variables descriptive statistics for Attica Tollway data (complete dataset).

Variables	Description	Mean	Standard Deviation
Flow	Average Flow (in veh/h)	1885.08	1244.60
Speed	Average Speed (in km/h)	107.61	7.58
Occupancy	Average Occupancy (in %)	3.12	2.32
Truck Proportion	Average Truck Proportion (in %)	4.22	2.68

prevail due to the crash itself, traffic time series before crash cases had to be checked. The idea was to identify potential sudden drops in mean speeds which would lead to erroneous estimates of the effect of traffic variables. In such cases, the crash is assigned to the previous 1-hour time interval. Sudden drops in speeds are usually used for detection of the time of a crash occurrence (Lee et al., 2003). Our approach is in line with the previous literature in the field and is similar to applying a time lag (Pande et al., 2011; Lee et al., 2011). More specifically, a time lag is often considered in order to avoid the impact of the crash itself on the traffic variables. Similar time lags have been applied in other real-time data analyses. For example, Christoforou et al. (2010), used a 12 min time lag. Quddus et al. (2009), used a more macroscopic approach for M25 motorway outside London and used a 30 min time lag. Moreover, Abdel-Aty and Pande (2005)and Zheng et al. (2010) utilized traffic flow data to detect abrupt and dramatic changes in traffic conditions at the upstream and downstream detectors. Since our dataset consisted of 1 h time intervals it was not possible to use such a microscopic time lag, as each time slice indicated if a crash had occurred or not. Therefore, if a sudden drop was detected on a time interval that a crash was assigned, it is considered high likely that this sudden drop in speed was caused by the crash itself.

The final dataset consists of 17 crash cases (occurred nearby these three loop detectors) as well as 91,118 non-crash cases. All types of crashes were considered regardless of the severity level. It is noted that in order to apply the Firth method, the entire dataset is utilized. On the other hand, the Bias Correction method supposes a stratified sampling, consequently all crash cases and a random sample of non-crashes cases are selected.

4. Results

4.1. Preliminary data analysis

Table 1 provides the summary descriptive statistics of the selected variables included in the final models.

Figs. 2–5 illustrate the kernel density estimation regarding non-crash and crash cases in respect to each independent traffic variable at a time: traffic flow, speed, occupancy and truck proportion. The kernel density estimation (KDE), which is a non-parametric way to estimate the probability density function of a variable, was used here to provide a first visual inspection of the data. It is noted that the y-axes are unitless as they represent the probability density.

4.2. Bias correction method

As stated previously, this procedure offers the option to correct the coefficients β in order to account for the rare events bias. This is the first time that crash probability in motorways was explored with the application of the Bias Correction method. The model results presented in this study are a first trial and an attempt to observe whether this methodological approach creates promising results and thus may be potentially considered fruitful.

One potential drawback of the Bias Correction logistic regression is the dependency of results on the stratified sampling. As a result, three trials were conducted and results are compared. For the stratified sampling, a proportion of 1:10 for the ratio of events (crashes) to non-events (non-crashes) was used in each sample. As suggested by King and Zeng (2001a,b) all crash cases were retained in each sample. Therefore, in each trial, there were 17 crash cases and 170 non-crash cases.

All candidate variables were checked for potential correlation

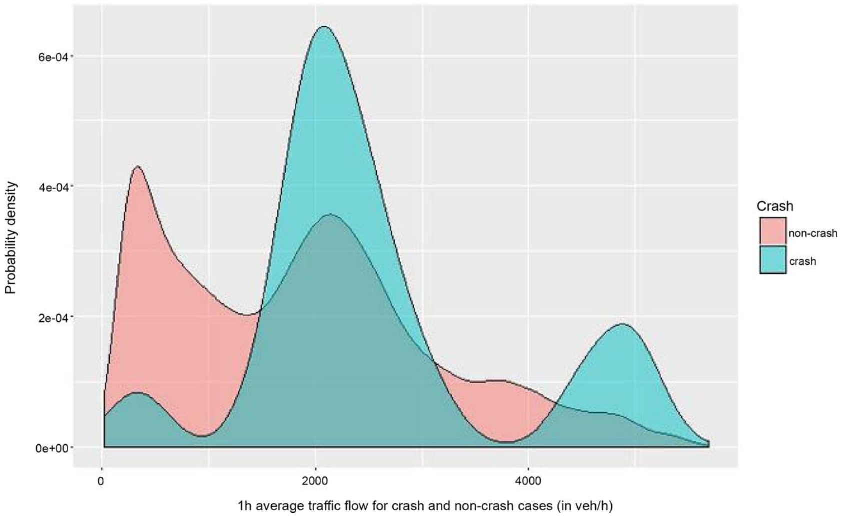

Fig. 2. Probability density of 1-hour average traffic flow for crash and non-crash cases.

before entered in the model in order to avoid multicollinearity problems. Consequently, it was not possible to include all explanatory variables in the models. For that reason, several tests had to be performed in order to find the best combination of independent variables. In order to illustrate the model but also to highlight which variables are consistently significant, non-significant variables are also included in the final models.

Tables 2, 3 and 4 present the results of the Bias Correction method for the tree trials, each of them having a different sample of non-crash cases. The results include the logistic coefficients β, the z-test values as well as the p-values for the explanatory variables. The standard error of β is presented as well, in order to further compare the models of the tree trials. All the models include the "prior correction", where τ is the proportion of events in the population (17/91,118 = 0.00019) and γ is the proportion of events in the sample (17/170 = 0.1).

The goodness-of-fit of the three models is reasonable. The values of McFadden-R^2 may be considered adequate, since it is suggested that values between 0.2 and 0.4 indicate a very good fit. It is interesting that all three models showed very similar fit in terms of the McFadden R^2. Furthermore, the change in the log-likelihood is significant in all three models.

For the Attica Tollway data, when the regression coefficient estimates, their standard errors and the significance levels for the explanatory variables are compared, a number of trends between the three proposed models can be observed. The three models showed a consistent negative effect of the logarithm of average speed, while average truck proportion was not found to affect crash occurrence. Moreover, the constant term was significant in all three models having a positive sign.

The constant term had the highest variation in the three models. The percentage of trucks in the traffic (Truck.Prop.) does not have the same sign across the models but the values of the beta coefficient (β) are similar ranging from −0.0444 to 0.0157. This can be attributed to the fact that all the beta coefficients are very close to zero. However, this parameter is not statistically significant in any of the models. It can be considered that the core trends in crash occurrence were successfully detected. Although it seems interesting that the number of trucks in traffic does not impact crash risk, Fig. 5 does not show any major differences between crash cases and non-crash cases in regard to truck percentage in traffic. Moreover, the fact that low speeds are associated with increased crash risk indicates that crashes seem to be caused by increased congestion rather than the number of trucks in traffic.

The only statistically significant explanatory variable was found to be speed, through its logarithmic transformation. The consistent negative sign of the beta coefficient of logarithm of average speed in all the models may seem counterintuitive, however is consistent with similar past studies (Ahmed et al., 2011, 2012b; Yu et al., 2013). For example, Ahmed et al. (2012b), found that low average speeds increase crash occurrence on freeways under clear weather. Therefore, considering the prevalence of good weather conditions in the Greater Athens Area, this

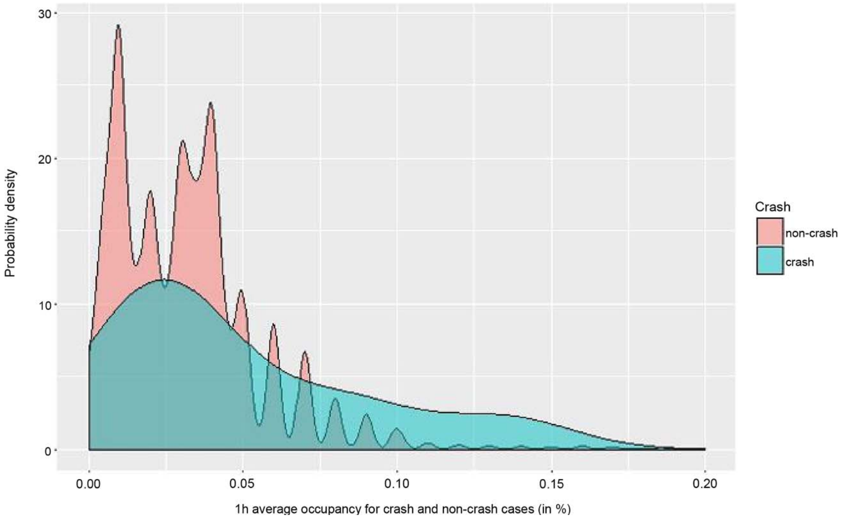

Fig. 3. Probability density of 1-hour average occupancy for crash and non-crash cases.

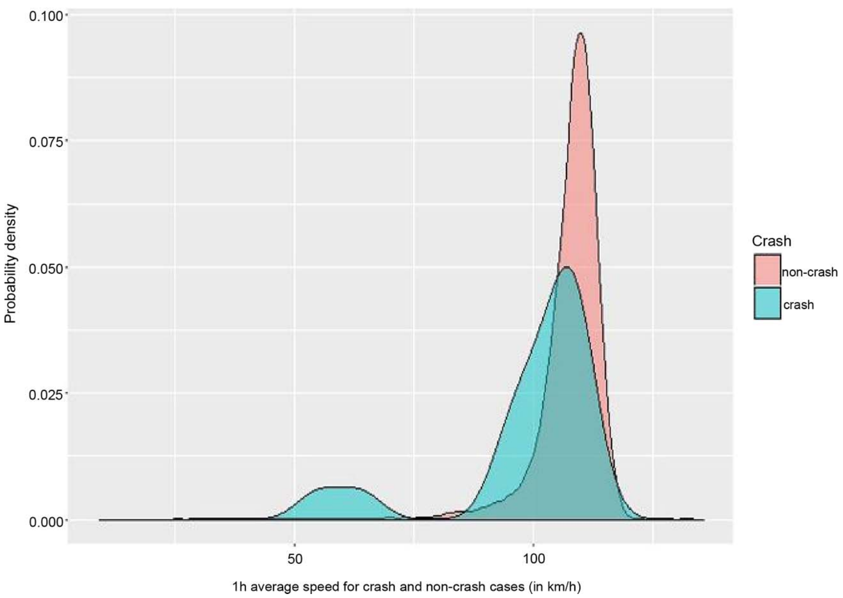

Fig. 4. Probability density of 1-hour average speed for crash and non-crash cases.

negative effect of low speeds on crash occurrence may be considered consistent with the aforementioned study. Moreover, this finding may indicate that crashes in Attica Tollway are more likely to occur in more dense traffic conditions with lower mean speeds or under adverse weather conditions which forced drivers to adapt their driving speeds.

4.3. Penalized maximum likelihood estimation (firth method)

In order to apply the Firth method the entire sample of crash and non-crash cases is considered which includes 17 crash and 91,118 non-crash cases. For considering all variables, each one is tested for its statistical significance to the target variable. Statistical tests for potential correlations among the independent variables had to be carried out before variables were entered in the models. Table 5 illustrates the modeling estimation results.

The likelihood ratio test is significant for this model as well, although the overall fit is lower than the fit of the Bias Correction models (McFadden $R^2 = 0.07$). A number of similarities and differences with Bias Correction models can be further observed from the results of Table 5. The proposed model includes the absolute value of average speed and not the logarithm of speed as in the previous analysis. However, the sign of the beta coefficient of average speed is negative here as well. The effect of the proportion of trucks in the traffic (Truck.Prop.) is non-significant here as well.

Table 6 illustrates a summary overview of the proposed statistical methods included in this framework. As per the aim of this research, both methods work well when the number of crashes is very low (rare events). It can also be observed that even if they utilize different sampling frames, the core findings of the two modeling approaches are

Table 2
Summary of the Bias Correction method for trial 1.

Trial 1	β	S.E.	z value	p value
Constant term	26.4158	11.3706	2.3232	0.0212
Truck. Prop.	−0.0394	0.1072	−0.3684	0.7129
log(Speed)	−7.4700	2.4369	−3.0653	0.0025
Log-likelihood at zero	−113.9			
Final log-likelihood	−100.9			
Likelihood ratio test	26.0			
McFadden R^2	0.1141			

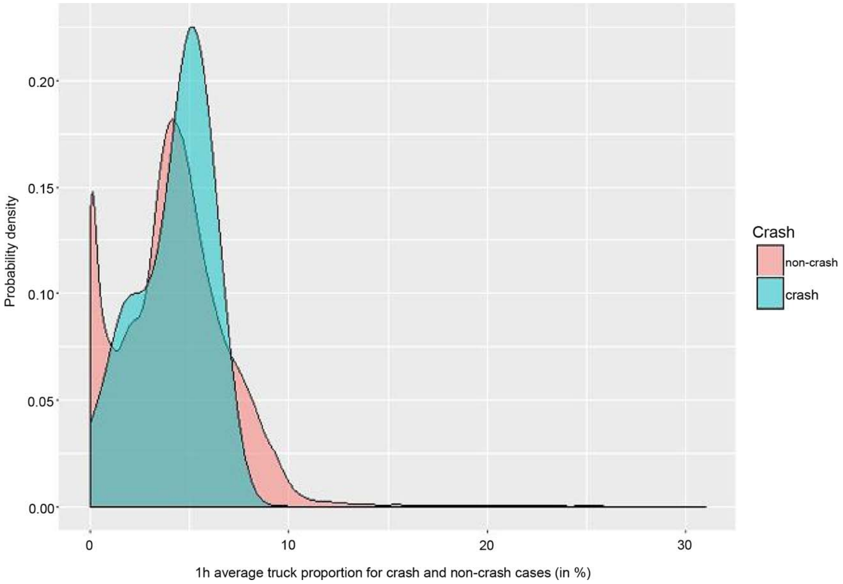

Fig. 5. Probability density of 1-hour average truck proportion for crash and non-crash cases.

Table 3
Summary of the Bias Correction method for trial 2.

Trial 2	β	S.E.	z value	p value
Constant term	33.2999	14.3741	2.3117	0.0216
Truck.Prop.	0.0157	0.0981	0.1597	0.8733
log(Speed)	−9.0004	3.0874	−2.9152	0.0039
Log-likelihood at zero	−113.9			
Final log-likelihood	−100.6			
Likelihood ratio test	26.6			
McFadden R^2	0.1168			

Table 4
Summary of the Bias Correction method for trial 3.

Trial 3	β	S.E.	z value	p value
Constant term	29.8363	12.6321	2.3619	0.0192
Truck.Prop.	−0.0444	0.0964	−0.4600	0.6460
log(Speed)	−8.2035	2.7063	−3.0311	0.0028
Log-likelihood at zero	−113.9			
Final log-likelihood	−100.8			
Likelihood ratio test	26.2			
McFadden R^2	0.1150			

Table 5
Summary of the Firth method.

Model	β	S.E.	Chi-square	p value
Constant term	−3.1819	1.0930	13.9363	0.0002
Speed	−0.0512	0.0106	11.4104	0.0007
Truck.Prop.	0.0071	0.0861	0.0056	0.9402
Log-likelihood at zero	−162.98			
Final log-likelihood	−151.566			
Likelihood ratio test	22.828			
McFadden R^2	0.07			

similar as both methods were able to capture the effect of the candidate variables on real-time crash risk. More specifically, decreased speeds were found to be positively associated with crash risk, while truck proportion had no impact. However, the magnitude of the effects is not exactly the same. The fact that the two models have not produced the exact same results could be expected, since they utilize: a) different data samples and b) different correction approaches for correcting bias. Overall, both models provide good insight when crashes are considered as rare-events, however one should take into account the time, effort and cost to collect and process all cases and controls before applying the Firth method. On the other hand, the Bias correction method will save time and effort to researchers, although sensitive to stratified sampling.

5. Conclusions

The aim of the present study is to investigate crash occurrence in motorways by utilizing real-time traffic data when the number of crashes is particularly low. In the proposed approach, crashes are considered as rare events. From a methodological point of view, this study adds to the current knowledge by utilizing appropriate logistic regression models specifically designed for rare events. In this way potential biases are overcome. To the best of our knowledge, this is one of the very first times that such models are applied in transport safety. It is noted that similar models have been applied for crash frequency purposes so far (e.g. zero-inflated models), but not for dichotomous dependent variables (e.g. crash vs no crash).

Furthermore, another research gap is addressed when exploring crash likelihood with real-time traffic data. More specifically, the proposed approach enables the development of crash likelihood models in specific segments or locations with a very low number of crashes. So far, the impact of real-time traffic and weather parameters on urban and rural areas is not deeply explored (Theofilatos and Yannis, 2014; Yannis et al., 2014; Wang et al., 2013). For example, the number of crashes on a rural road or on a specific road segment might be too low in comparison with urban segments. Therefore, in order to examine the effect of traffic parameters on crash likelihood, existing approaches are not appropriate and either the Bias correction or the Firth methods should be used to handle the problem. Additionally, Powered-Two-Wheeler (PTW) safety is underrepresented (Theofilatos and Yannis 2014, 2015). Therefore, future studies could potentially explore PTW crash likelihood by utilizing real-time data and by applying the proposed approach of this study.

In terms of model estimation results, it was found that the main risk factor for crash occurrence was average speed, indicating that lower speeds were positively associated with crash risk. This may be attributed to the fact that lower operating speeds in motorways may imply a) congested traffic conditions and/or b) adverse weather (since speed drops from crash occurrence were excluded). Particularly, adverse weather in the region of Athens does not frequently occur; when it does drivers lower their speeds to compensate. This finding can be considered consistent with relevant past studies which examined crash occurrence by using real-time traffic data. Moreover, another study by Imprialou et al. (2016b), showed that slight injury multiple-vehicle crashes are not related to high speeds with congested traffic. Consequently, proactive management of motorways could potentially rely on such findings.

The application of the Bias Correction and the Firth model and the produced results are considered promising, since the risk factors as well as insignificant parameters were identified. The beta coefficients of the independent variables were not found to be totally consistent across the models, probably due to different sampling of the Bias Correction

Table 6
Overview of the proposed methods and results.

Method	Case-control ratio	Advantages	Disadvantages	Candidate parameters that affect crash risk in freeway	Effect of parameters affecting crash risk in freeway
Bias correction	10:1	● Saves time and effort to researchers. ● No need to collect all cases and controls. ● Specific location with low number of crashes (< 200) can be efficiently studied.	● Sensitive to stratified sampling.	● Speed. ● Truck proportion.	● Increased mean speed (logarithm) reduces crash risk. ● Truck proportion has no effect.
Firth logistic	All cases and controls	● Can capture the "real" relationship between the dependent variables and the predictors, without losing information due to random sampling. ● Specific location with low number of crashes (< 200) can be efficiently studied.	● Time consuming and expensive to collect all cases and controls.	● Speed. ● Truck proportion.	● Increased mean speed reduces crash risk. ● Truck proportion has no effect.

model and Firth model. However, the core trends were successfully detected. The R-square values of the models are not considered high, however this study can be considered as a first trial. The rarity of crashes in our approach has also to be considered. Moreover, only a few variables were used in our study due to data limitations. If more traffic parameters were available and added in the models, the statistical fit of the models would be improved. Consequently future studies should utilize more information regarding traffic parameters.

Furthermore, due to weather data unavailability, only traffic parameters were considered. Weather effects should be included in these models in future research, especially for countries such as Greece where the weather effects are of particular interest but not so straightforward as past studies indicate (Antoniou et al., 2013; Bergel-Hayat et al., 2013).

The two methods have different advantages and disadvantages, therefore the choice of the most appropriate method depends on several criteria. For example, the Firth method may be more detailed, as the entire sample of non-crash cases are involved (i.e. all time slices for all loop detectors for all segments). In the era of big data the processing of huge datasets is becoming easier, so thousands to millions of non-crash cases can be included. In that context, more disaggregate traffic and weather data, such as in 5 min or even 1 min intervals could be used. Nevertheless, relevant research is still at an early stage (Vlahogianni, 2015; Shi and Abdel-Aty, 2015). The utilization of more microscopic traffic data could have likely led to different results. From a methodological point of view, the results of the Firth method could perhaps change, since many more non-crash cases would be included in the model, however the results of the Bias correction method would not change respectively, because that method relies on a 10:1 non-crashes to crashes random sampling and correcting. From a data point of view, since the traffic data aggregation would be 5 min intervals, results would be more microscopic. Therefore, the 5 min traffic flow could have a different effect than the 1 h traffic flow.

However, the Bias Correction method saves time and effort because it requires all crash cases but only a sample of non-crash cases (preferably a ratio of 1:10).A possible limitation is that the Bias Correction method is potentially sensitive to stratified sampling. More efforts are needed in order to overcome this limitation and improve similar models. For example, logit models with replications could be potentially used, as a few studies have shown (Guns and Vanacker, 2012). Summing up, the methods applied in the paper are promising and should be considered when the number of crashes is less than 200.

Overall, our paper approach: a) provides a complete methodological guidance and also methodological alternatives, b) is directly applicable in any case when crashes crashes are considered as rare-events c) improves data collection because the most appropriate control-case ratio is collected in order to apply the bias correction method (1:10 ratio), and d) provides insight on the important issue of real-time crash likelihood modeling and relevant critical risk factors.

References

Abdel-Aty, M., Pande, A., 2005. Identifying crash propensity using specific traffic speed conditions. J. Saf. Res. 36, 97–108.

Abdel-Aty, M., Pande, A., Lee, C., Gayah, V., Dos Santos, C., 2007. Crash risk assessment using intelligent transportation systems data and real-time intervention strategies to improve safety on freeways. J. Intell. Transp. Syst. Technol. Plann. Oper. 11 (3), 107–120.

Ahmed, M., Abdel-Aty, M., 2012. The viability of using automatic vehicle identification data for real-time crash prediction. IEEE Trans. Intell. Transp. Syst. 13 (2), 459–468.

Ahmed, M., Abdel-Aty, M., Yu, R., 2012a. A bayesian updating approach for real-time safety evaluation using AVI data. J. Transp. Res. Board 2280, 60–67.

Ahmed, M., Abdel-Aty, M., Yu, R., 2012b. Assessment of the interaction between crash occurrence, mountainous freeway geometry, real-time weather and AVI traffic data. Transp. Res. Record 2280, 51–59.

Ahmed, M., Huang, H., Abdel-Aty, M., Guevara, B., 2011. Exploring a Bayesian hierarchical approach for developing safety performance functions for a mountainous freeway. Accid. Anal. Prev. 43, 1581–1589.

Anastasopoulos, P.C., Mannering, F.L., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accid. Anal. Prev. 41, 153–159.

Antoniou, C., Yannis, G., Katsohis, D., 2013. Impact of meteorological factors on the number of injury accidents. In: Proceedings of the 13th World Conference on Transportation Research. COPPE - Federal University of Rio de Janeiro at Rio de Janeiro, Brazil. pp. 15–18 July.

Bergel-Hayat, R., Debbarh, M., Antoniou, C., Yannis, G., 2013. Explaining the road accident risk: weather effects. Accid. Anal. Prev. 60, 456–465.

Caliendo, C., Guida, M., Parisi, A., 2007. A crash-prediction model for multilane roads. Accid. Anal. Prev. 39, 657–670.

Ceder, A., 1982. Relationships between road accidents and hourly traffic flow-II. probabilistic approach. Accid. Anal. Prev. 14 (1), 35–44.

Chang, L.-Y., 2005. Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network. Saf. Sci. 43, 541–557.

Dickerson, A., Peirson, J., Vickerman, R., 2000. Road accidents and traffic flows: an econometric investigation. Economica 67 (265), 101–121 London School of Economics and Political Science, London.

Elvik, R., Høye, A., Vaa, T., Sørensen, M., 2009. The Handbook of Road Safety Measures. Emerald Group Publishing Limited, Howard House, Wagon Lane, Bingley, UK.

Evans, L., 1991. Traffic Safety and the Driver. Van Nostrand Reinhold, New York, N.Y.

Firth, D., 1993. Bias reduction of maximum likelihood estimates. Biometrika 80 (1), 27–38.

Guido, G., Astarita, V., Giofré, V., Vitale, A., 2011. Safety performance measures: a comparison between microsimulation and observational data. Procedia-Soc. Behav. Sci. 20, 217–225.

Guns, M., Vanacker, V., 2012. Logistic regression applied to natural hazards: rare event logistic regression with replications. Nat. Hazards Earth Syst. Sci. 12, 1937–1947.

Heinze, G., Schemper, M., 2002. A solution to the problem of separation in logistic regression. Stat. Med. 21, 2409–2419.

Heinze, G., Ploner, M., 2003. Fixing the nonconvergence bug in logistic regression with SPLUS and SAS. Comput. Methods Prog. Biomed. 71, 181–187.

Heinze, G., Ploner, M., Dunkler, D., Southworth, H., 2015. Firth's Bias Reduced Logistic Regression. Statistical package. http://cemsiis.meduniwien.ac.at/en/kb/science-research/software/statistical-software/fllogistf/.

Hossain, M., Muromachi, Y., 2013. Understanding crash mechanism on urban expressways using high-resolution traffic data. Accid. Anal. Prev. 57, 17–29.

Imprialou, M.-I., Maher, M., Quddus, M., 2016a. Exploring crash-risk factors using Bayes' theorem and an optimization routine. In: Proceedings of the 95th Annual Meeting of the Transportation Research Board. January 10-14, Washington, D.C.

Imprialou, M.-I., Quddus, M., Pitfield, D.E., 2016b. Predicting the safety impact of a speed limit increase using condition-based multivariate poisson lognormal regression. Transp. Plann. Technol. 39 (1), 3–23.

King, G., Zeng, L., 2001a. Explaining rare events in international relations. Int. Org. 55 (3), 693–715.

King, G., Zeng, L., 2001b. Logistic regression in rare events data. Polit. Anal. 9 (2), 137–163.

Kockelman, K.M., Ma, J., 2007. Freeway speeds and speed variations preceding crashes, within and across lanes. J. Transp. Res. Forum 46 (1), 43–61.

Kononov, J., Durso, C., Reeves, D., Allery, B.K., 2012. Relationship between traffic density, speed, and safety and its implications for setting variable speed limitson freeways. Transp. Res. Record 2280, 1–9.

Lee, C., Hellinga, B., Saccomanno, F., 2003. Real-time crash prediction model for the application to crash prevention in freeway traffic. In: Proceedings of the 82nd Annual Meeting of the Transportation Research Board. January 12-16, Washington, D.C.

Lee, C., Park, P., Abdel-Aty, M., 2011. Lane-by-lane analysis of crash occurrence based on driver's lane-changing and car-following behavior. J. Transp. Saf. Secur. 3 (2), 108–122.

Leitgöb, H., 2013. The problem of rare events in maximum likelihood logistic regression - assessing potential remedies. In: 15-19 July, 2013, Lisbon, Portugal. 5Th Conference of the European Survey Research Association.

Lord, D., Manar, A., Vizioli, A., 2005. Modeling crash-flow-density and crash-flow-V/C ratio relationships for rural and urban freeway segments. Accid. Anal. Prev. 37, 185–199.

Martin, J.-L., 2002. Relationship between crash rate and hourly traffic flow on interurban motorways. Accid. Anal. Prev. 34, 619–629.

Noland, R.B., Quddus, M.A., 2005. Congestion and safety: a spatial analysis of London. Transp. Res. A Policy Pract. 39, 737–754.

Pande, A., Abdel-Aty, M., 2006a. Assessment of freeway traffic parameters leading to lane-change related collisions. Accid. Anal. Prev. 38, 936–948.

Pande, A., Abdel-Aty, M., 2006b. Comprehensive analysis of the relationship between real-time traffic surveillance data and rear-end crashes on freeways. Transp. Res. Record J. Transp. Res. Board 1953, 31–40.

Pande, A., Dasand, A., Abdel-Aty, M., Hassan, H., 2011. Real-time crash risk estimation are all freeways created equal? Transp. Res. Record 2237, 60–66.

Quddus, M.A., Wang, C., Ison, S.G., 2009. The impact of road traffic congestion on crash severity using ordered response models. road traffic congestion and crash severity: econometric analysis using ordered response models. J. Transp. Eng. 136 (5), 424–435.

Roshandel, S., Zheng, Z., Washington, S., 2015. Impact of real-time traffic characteristics on freeway crash occurrence: systematic review and meta-analysis. Accid. Anal. Prev. 79, 198–2011.

Shi, Q., Abdel-Aty, M., 2015. Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transp. Res. Part C 58 (2015), 380–394.

Theofilatos, A., Yannis, G., 2014. A review of the effect of traffic and weather characteristics on road safety. Accid. Anal. Prev. 72, 244–256.

Theofilatos, A., Yannis, G., 2015. A review of powered-two-wheeler behaviour and safety. Inj. Control Saf. Promot. 22 (4), 284–307.

Theofilatos, A., 2017. Incorporating real-time traffic and weather data to explore road accident likelihood and severity in urban arterials. J. Saf. Res. 61, 9–21.

Vlahogianni, E.I., 2015. Computational intelligence and optimization for transportation big data: challenges and opportunities. Comput. Methods Appl. Sci. 38, 107–128.

Wang, C., Quddus, M.A., Ison, S.G., 2009. Impact of traffic congestion on road acci-dents: a spatial analysis of the M25 motorway in England. Accid. Anal. Prev. 41, 798–808.

Wang, C., Quddus, M.A., Ison, S.G., 2013. The effect of traffic and road characteristics on road safety: a review and future research direction. Saf. Sci. 57, 264–275.

Wang, L., Abdel-Aty, M., Shi, Q., Park, J., 2015. Real-time crash prediction for expressway weaving segments. Transp. Res. Part C 61, 1–10.

Xu, C., Liu, P., Wang, W., Li, Z., 2012. Evaluation of the impacts of traffic states on crash risks on freeways. Accid. Anal. Prev. 47, 162–171.

Xu, C., Tarko, A.P., Wang, W., Liu, P., 2013b. Predicting crash likelihood and severity on freeways with real-time loop detector data. Accid. Anal. Prev. 57, 30–39.

Xu, C., Wang, W., Liu, P., 2013a. Identifying crash-prone traffic conditions under different weather on freeways. J. Saf. Res. 46, 135–144.

Yang, H., Ozbay, K., Xie, K., Bartin, B., 2015. Modeling 1 crash risk of highway work zones with relatively short durations. In: Proceedings of the 85th Annual Meeting of the Transportation Research Board. January 11-15, Washington, D.C.

Yannis, G., Theofilatos, A., Ziakopoulos, A., Chaziris, A., 2014. Investigation of road accident severity and likelihood in urban areas with real-time traffic data. Traffic Eng. Control 55 (1), 31–35.

Yu, R., Abdel-Aty, M., 2013. Investigating the different characteristics of weekday and weekend crashes. J. Saf. Res. 46, 91–97.

Yu, R., Abdel-Aty, M., Ahmed, M., 2013. Bayesian random effect models incorporating real-time weather and traffic data to investigate mountainous freeway hazardous factors. Accid. Anal. Prev. 50, 371–376.

Zheng, Z., Ahn, S., Monsere, C.M., 2010. Impact of traffic oscillations on freeway crash occurrences. Accid. Anal. Prev. 42, 626–636.