Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 14 (2016) 3399 – 3405

Transportation Research Procedia

www.elsevier.com/locate/procedia

6th Transport Research Arena April 18-21, 2016

Predicting road accidents: a rare-events modeling approach

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*

Abstract

Modeling road accident occurrence has gained increasing attention over the years. So far, considerable efforts have been made from researchers and policy makers in order to explain road accidents and improve road safety performance of highways. In reality, road accidents are rare events. In such cases, the binary dependent variable is characterized by dozens to thousands of times fewer events (accidents) than non-events (non-accidents). Instead of using traditional logistic regression methods, this paper considers accidents as rare events and proposes a series of rare-events logit models which are applied in order to model road accident occurrence by utilizing real-time traffic data. This statistical procedure was initially proposed by King and Zeng (2001) when scholars study rare events such as wars, massive economic crises and so on. Rare-events logit models basically estimate the same models as traditional logistic regression, but the estimates as well as the probabilities are corrected for the bias that occurs when the sample is small or the observed events are very rare. Consequently, the basic problem of underestimating the event probabilities is avoided as stated by King and Zeng (2001). To the best of our knowledge, this is the first time that this approach is followed when road accident data are analyzed. Instead of applying a traditional case-control study, the complete dataset of hourly aggregated traffic data such as flow, occupancy, mean time speed and percentage of trucks, 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 modeling results showed an adequate statistical fit and reveal a negative relationship between accident occurrence and the natural logarithm of speed in the accident location. This study attempts to contribute to the understanding of accident occurrence in motorways by developing novel models such as the rare-events logit for the first time in safety evaluation of motorways.

* Corresponding author. Tel.: +30-210-7721575; fax: +30-210-7721454.
E-mail address: atheofil@central.ntua.gr

Peer-review under responsibility of Road and Bridge Research Institute (IBDiM)
doi:10.1016/j.trpro.2016.05.293

Peer-review under responsibility of Road and Bridge Research Institute (IBDiM)

Keywords: accident occurrence; rare events; logistic regression; traffic parameters

1. Introduction

Accidents impose serious problems to society in terms of human costs, economic costs, property damage costs and medical costs. Understanding the various factors that affect accident occurrence is of particular concern to decision makers and researchers.

Recently, the incorporation of real-time (high resolution) traffic and weather data in freeways has proven to be a promising approach. A common methodology adopted to predict accident occurrence is to consider both accident and non-accident cases. This methodology was followed by many studies in the past (Abdel-Aty and Pande, 2005; Abdel-Aty et al., 2007; Ahmed and Abdel-Aty, 2012; Yu and Abdel-Aty, 2013a). Ahmed et al. (2012a) found that increased speed variation at any given accident segment combined with a decrease in average speed in the respective downstream segment, can lead in increased likelihood of rear-end accident occurrence. Ahmed and Abdel-Aty (2012), found that the probability of an accident increases when the variation in speed increases and the average speed decreases at the segment of the accident at 5–10 minutes prior to the accident occurrence. However, variation in speed is not always a risk factor. Kockelman and Ma (2007) have indicated that 30-sec speed changes are not correlated with risk of accidents.

Studies examining accident occurrence usually include traffic and weather characteristics as well as the combined effect of traffic and weather. Ahmed et al. (2012b) investigated the impact of geometrical, traffic and weather variables on accident occurrence on freeways. In winter, it was found that low visibility, high precipitation and speed variation increase the likelihood of accidents. Surprisingly, for dry season, low average speeds and low visibility increase the odds of an accident.

Yu et al. (2013), found that the 5-min average speed of the crash segment during 5–10 min prior to the crash time was found to significantly affect accidents. It is interesting that the authors suggested a negative correlation, which means that the crash occurrence likelihood increases as the average speed decreases 5–10 min before the crash occurrence.

Xu et al. (2012) developed accident 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 accident 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.

The literature review indicated that the main approach for analyzing accident occurrence (often referred as crash likelihood) is the case-control method, where a number of non-accident cases are collected under appropriate assumptions. However, in reality, accident might be considered as rare events. Consequently, a different approach might be more appropriate. For that reason, the aim of the study is to link accident occurrence with traffic characteristics, by considering accidents as rare events.

2. Methodology

Most of significant events in reality are rare events. They occur very rarely-that is we have dozens to thousands of times fewer events (e.g. wars, volcano explosions) than non-events. King and Zeng (2001a) identified two major causes for problems when analyzing such kind of data. Firstly, traditional statistical procedures underestimate the probability of rare events, and second, the inefficient data-collection strategies. In general, 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 rely on logit coefficients. To solve these problems,

King and Zeng (2001a, b) proposed an adapted version of the logistic regression, the so-called rare-events logistic regression.

This approach, applies a number of corrections. The first correction concerns data collection. The authors suggest to perform a case-control sampling design, based on stratified sampling. Thus, 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. Then the next equation applies:

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

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 by for correction is the "weighting" method, which was not used in this study and thus not described here. Moreover, the underestimation of the 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 pi. If we assume the corrected logit form based on the corrected constant term:

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

then,

$$p_i^{'} = p_i + C_i \quad (3)$$

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

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

where p_i is the probability of an event estimated using the corrected estimated coefficient a_0, x_0 is the 1*(m+1) vector of values for each independent variable, V(β) is the variance-covariance matrix, and lastly x_0' is the x_0 transposed. In this study, the rare-events model is applied under the assumption that accidents in the Attica Tollway are rare events, and thus the term "event" corresponds to an occurrence of an accident.

3. Data collection and preparation

In this study, the required accident and traffic data were extracted from Attica Tollway ("Attiki Odos") located in Greater Athens Area in Greece. Attica Tollway is a modern motorway extending along 65.2 km. Entry to the freeway is through 39 toll plazas with 195 toll lanes. 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. Being a closed motorway, it has controlled access points and consists of two sections, which are perpendicular to one another: The Eleusina – Stavros – Spata A/P motorway (ESSM), extending along approximately 52 km, and the Imittos Western Peripheral Motorway (IWPM), extending along approximately 13 km. Attica Tollway also connects Athens with the International Airport "Eleutherios Venizelos".

Inductive loops (sensors) are placed every 500 meters 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. These sensors enable the prompt detection of any problems causing disruption to the smooth flow of traffic and automatically activates intervention procedures to deal with the problem. Aside from the extensive loop detector

system, traffic monitoring and management are conducted through a series of Closed Circuit TV cameras (CCTV), Variable Message Signs, Variable Speed Limit Signs and Over Height Vehicle Detectors. Figure 1 that follows, shows the blue line that indicates the Attika Tollway in the Greater Athens Area.

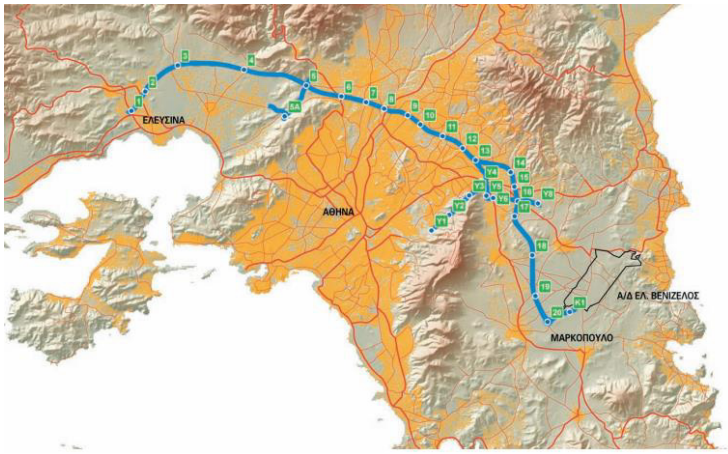

Fig. 1. Map of Attika Tollway.

Two different datasets were prepared for the need of the analysis; one dataset with accident data and one with traffic data. The required accident data for Attica Tollway were extracted from the Greek accident 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. It is located in Paiania and operates on a 24-hour basis. The Traffic Management and Motorway Maintenance Department contains one (1) 80 inches monitor, forty (40) 21 inches monitors and 7 workstations with access to motorway systems.The main purposes of the Traffic Management and Motorway Maintenance Department are among others: Traffic control and monitoring; Management of emergency incidents and planned activities; Inspection, maintenance and repairs, as may be required in order to preserve the good condition of the motorway; Maintenance of operation and maintenance vehicles and facilities; Routine works, such as cleaning the motorway. It is noted that only basic motorway segments (BFS) were considered for the analysis and not ramp areas.

More specifically, in order to proceed in the accident occurrence modeling, 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-per direction) were considered. Traffic and accident data from these three loop detectors were extracted and unified. The final unified dataset, consists of 17 accident cases (occurred nearby these three loop detectors) as well as 91118 non-accident cases. Traffic variables were measured in 1-hour intervals (flow, speed, occupancy and truck proportion). The flow is defined as the total number of vehicles on 1 hour 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 the 5-minutes intervals, which were automatically aggregated in 1-hour intervals.

Accident occurrence was defined as binary variable taking the values of 0 (non-accident) and 1 (accident). Therefore, in each 1-h time interval there is the information if an accident has occurred or not. In order to avoid the post-accident traffic conditions where low mean speeds may prevail due to the accident itself, the traffic time series before accident cases had to be checked. The aim was to identify potential sudden fall in mean speeds which would lead to erroneous estimates of the effect of traffic variables, because of the decrease in average mean speed, due to the effect of the accident itself. In such cases, the accident is assigned to the previous 1-hour time interval.

The next table (Table 1) provides the summary descriptive statistics of the final selected variables included in the models. In addition, all candidate variables were checked for correlation so as to avoid multicollinearity problems.

Table 1. Summary of independent variables descriptive statistics for Attica Tollway data.

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.Prop.	Average Truck Proportion (in %)	4.22	2.68

4. Results

Accident occurrence analyses were performed through the package *zelig* (Imai et al., 2007 and 2008) in R software. It is noted that 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 accident probability in motorways was explored with the application of the rare-events logistic regression. Therefore, the model results presented in this study, are a first trial and is attempted to observe whether this methodological approach creates promising results and thus may be potentially considered fruitful.

The main drawback of the rare-events 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 (accidents) to non-events (non-accidents) was used in each sample. As suggested by King and Zeng (2001a and b) all accident cases were retained in each sample. Therefore, in each trial, there were 17 accident cases and 170 non-accident cases.

It was not possible to include all explanatory variables in the models, due to serious multicollinearity problems among traffic variables. Therefore, 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 to 4 present the results of the rare-events logit models for the tree trials, each of them having a different sample of non-accident cases. The results include the logistic coefficients β, the standard error of β, the z-test values as well as the p-values for the explanatory variables. The standard error 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 (17/91118 = 0.00019) and γ is the proportion of events in the sample (17/170 = 0.1).

The 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. Furthermore, the change in the log-likelihood is significant in all three models. It is interesting though, that all three models showed very similar fit, either in terms of McFadden-R^2 or AIC. The three values of AIC are generally low, indicating good fit and have similar values, ranging from 106.6 to 106.9.

Table 2. Summary of rare-events logistic regression for trial 1.

Trial 1	β	S.E.	z value	p value
Constant	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		
AIC		106.9		
McFadden R^2		0.1141		

Table 3. Summary of rare-events logistic regression for trial 2.

Trial 2	β	S.E.	z value	p value
Constant	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		
AIC		106.6		
McFadden R^2		0.1168		

Table 4. Summary of rare-events logistic regression for trial 3.

Trial 3	β	S.E.	z value	p value
Constant	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		
AIC		106.8		
McFadden R^2		0.1150		

One drawback of the rare-events logistic regression is the potential dependency of the results on the endogenous stratified sampling. However, in the Attica Tollway data, when we compare the regression coefficient estimates, their standard errors and significance levels for the explanatory variables, we observe a number of trends between the three proposed models. The three models showed a consistent negative effect of the logarithm of average speed, while average truck proportion was not found to affect accident occurrence. Moreover, the constant term was significant in all three models having a positive sign. Therefore, the models managed to detect the statistically significant and the insignificant parameters successfully.

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 very 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 is noted though, that the core trends in accident occurrence were successfully detected.

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 on accident occurrence on freeways under clear weather. Therefore, considering the prevalence of good weather conditions in the Greater Athens Area, this negative effect of low speeds on accident occurrence may be considered consistent with the aforementioned study. Moreover, this finding may indicate that accidents 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.

5. Conclusions

The aim of the present study was to investigate accident occurrence in motorways by utilizing high resolution traffic data. Accident occurrence was explored by developing rare-events logistic regression models. The method of stratified sampling was followed and three models were developed. The main risk factor for accident occurrence was found to be the logarithm of average speed, as lower speeds were found to be more likely to result in accident occurrence. This finding was consistent with other past studies which examined accident occurrence by using real--time traffic data. Consequently, proactive management of freeways could rely on such findings, for example by setting variable speed limits.

This study contributes on the current knowledge, by applying a series of the rare-events. From a methodological point of view, the application of the rare-events logit model is considered promising, as it enables real-time prediction models of accident occurrence in segments or locations with a very low number of accidents. To the best of our knowledge, this is the first time that such models are applied in transport safety.

The application of this model and the produced results are considered promising, since the risk factors as well as the statistically insignificant parameters were identified. The beta coefficients of the independent variables were not found to be totally consistent across the three models, maybe due to the dependence of rare-events logistic on stratified sampling. However, the basic trends were successfully detected. In order to overcome this limitation and to improve similar models more efforts are needed. For example, rare-events logit models with replications could be used (Guns and Vanacker, 2012). Moreover, it would be interesting to examine the impact of weather parameters by applying such models.

References

Abdel-Aty, M., Pande, A., 2005. Identifying crash propensity using specific traffic speed conditions. Journal of Safety Research 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. Journal of Intelligent Transportation Systems: Technology, Planning & Operations 11(3), 107–120.

Ahmed, M., Abdel-Aty, M., 2012. The viability of using Automatic Vehicle Identification data for real-time crash prediction. IEEE Transactions of intelligent transportation systems 13(2), 459–468.

Ahmed, M., Abdel-Aty, M., Yu, R., 2012a. A bayesian updating approach for real-time safety evaluation using AVI data. Journal of Transportation Research 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. Transportation Research 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. Accident Analysis & Prevention 43, 1581–1589.

Guns, M., Vanacker, V., 2012. Logistic regression applied to natural hazards: rare event logistic regression with replications. Natural Hazards & Earth System Sciences 12, 1937–1947.

Imai, K., King, G., Lau O., 2007. Zelig: Everyone's statistical software. Retrieved from: http://GKing.harvard.edu/zelig.

Imai, K., King, G., Lau O., 2008. Toward a common framework for statistical analysis and development. Journal of Computational & Graphical Statistics 17(4), 892–913.

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

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

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

Xu, C., Liu, P., Wang, W., Li, Z., 2012. Evaluation of the impacts of traffic states on crash risks on freeways. Accident Analysis & Prevention 47, 162– 171.

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. Accident Analysis & Prevention 50, 371–376.