Accident Analysis and Prevention 94 (2016) 207–215

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Evaluation of the predictability of real-time crash risk models

Chengcheng Xu[a,b], Pan Liu[a,b,*], Wei Wang[a,b]

[a] Jiangsu Key Laboratory of Urban ITS, Southeast University, Si Pai Lou #2, Nanjing, 210096, China
[b] Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, Si Pai Lou #2, Nanjing, 210096, China

ARTICLE INFO

Article history:
Received 24 September 2015
Received in revised form 7 May 2016
Accepted 7 June 2016
Available online 19 June 2016

Keywords:
Crash risk
Predictability
Real-time crash risk modeling
Predictive performance
Freeways

ABSTRACT

The primary objective of the present study was to investigate the predictability of crash risk models that were developed using high-resolution real-time traffic data. More specifically the present study sought answers to the following questions: (*a*) how to evaluate the predictability of a real-time crash risk model; and (*b*) how to improve the predictability of a real-time crash risk model. The predictability is defined as the crash probability given the crash precursor identified by the crash risk model. An equation was derived based on the Bayes' theorem for estimating approximately the predictability of crash risk models. The estimated predictability was then used to quantitatively evaluate the effects of the threshold of crash precursors, the matched and unmatched case-control design, and the control-to-case ratio on the predictability of crash risk models. It was found that: (*a*) the predictability of a crash risk model can be measured as the product of prior crash probability and the ratio between sensitivity and false alarm rate; (*b*) there is a trade-off between the predictability and sensitivity of a real-time crash risk model; (*c*) for a given level of sensitivity, the predictability of the crash risk model that is developed using the unmatched case-controlled sample is always better than that of the model developed using the matched case-controlled sample; and (*d*) when the control-to-case ratio is beyond 4:1, the increase in control-to-case ratio does not lead to clear improvements in predictability.

1. Introduction

Crashes are rare and random events whose occurrences are influenced by a set of factors that are partly deterministic and partly stochastic. The circumstances that lead to a crash in one event will not necessarily lead to a crash in a similar event. The randomness of crashes introduces natural fluctuation in crash counts over time; and the average crash frequency in the short term may be significantly different from the expected crash frequency in the long term. The natural variability in crash counts makes it difficult to identify whether the changes in the observed crashes are due to the natural variability or to the changes in site conditions. Without being accounted for properly, the randomness of crashes will eventually introduce regression-to-the-mean bias, resulting in a biased estimate of safety.

To account for the natural variability in crash data, traditional crash modeling framework uses the Poisson family distributions, such as the Poisson distribution, Poisson-gamma distribution and Poisson-lognormal distribution for estimating the expected number of crashes in the long term. Crash frequency models are developed on the basis of the data that are aggregated over time and space. The process of aggregation may result in the loss of useful information with regard to the causal factors to crashes. In essence, the major task of traditional crash frequency models is not to predict crashes. Instead, they aim at estimating the long-term average crash frequency under a given set of geometric design, traffic control, and traffic conditions.

During the past two decades, with the widespread use of freeway traffic surveillance systems, increased attention has been given to identifying the traffic flow conditions prior to crash occurrences using high-resolution traffic data. Real-time crash risk models have been developed to link crash risks to dynamic traffic flow parameters (Oh et al., 2001; Lee et al., 2003; Abdel-Aty et al., 2004, 2005; Abdel-Aty and Pemmanaboina, 2005, 2006; Abdel-Aty and Pande, 2006; Xu et al., 2012). Unlike crash frequency models, the real-time crash risk models treat each individual crash as the unit of analysis. The central idea is to estimate the relative risks of crashes for a relatively small time interval (usually 5 min) on the basis of the hazardous traffic conditions that commonly occur before crashes (crash precursors) combined with other influence factors such as geometric design characteristics and weather information. With

* Corresponding author at: Jiangsu Key Laboratory of Urban ITS, Southeast University, Si Pai Lou #2, Nanjing, 210096, China.
E-mail addresses: xuchengcheng@seu.edu.cn (C. Xu), pan_liu@hotmail.com (P. Liu), wangwei@seu.edu.cn (W. Wang).

http://dx.doi.org/10.1016/j.aap.2016.06.004
0001-4575/© 2016 Elsevier Ltd. All rights reserved.

crash risk models, various proactive safety management techniques can be applied to make interventions before the occurrences of crashes (Lee et al., 2006a, 2006b; Khoury and Hobeika, 2007; Allaby et al., 2007).

The case-control study design forms the foundation of crash risk modeling. In a case-controlled dataset, the traffic data prior to crashes are taken as cases while those under crash free conditions are taken as controls. The logistic regression technique can then be used to distinguish between crash precursors and normal traffic conditions, and to establish a relationship between crash precursors and crash risks. In the case-control study design, a typical case-to-control ratio of 1:4 has been widely accepted. That is, for each crash case, researchers may select four non-crash cases. By doing so rare events can be investigated in a relatively quick and cheap manner by greatly reducing the total number of controls.

In the real world, however, the number of non-crash cases is actually enormous compared with that of the crash cases. For example, considering a freeway section with 70 loop detector stations, the total number of non-crash cases in three years is close to $70 \times 365 \times 3 \times 24 \times 60 = 1.1\mathrm{E}+08$, if the data are aggregated based on one-minute time intervals. Thus, the case-controlled samples are biased towards over-representing the crash cases. Assuming that a typical case-to-control ratio of 1:4 is used, a crash is expected to occur in one of every five records. However, the actual probability of a crash in a 5-min time interval is extremely small considering the large number of non-crash cases on the entire population. As a result, existing crash risk models provide only a measure of the relative risks of crash occurrences given the composition of the samples.

The probabilistic nature of both crash frequency and crash risk models raises fundamental questions: what is the role of the randomness in crash occurrences, and to what extend are crashes predictable? The primary objective of the present study was to investigate the predictability of real-time crash risk models. More specifically the present study sought answers to the following questions: (*a*) how to evaluate the predictability of a real-time crash risk model; and (*b*) how to improve the predictability of a real-time crash risk model.

2. Literature review

In some early studies, only the data prior to crashes were considered when modeling crash risks (Golob and Recker, 2004; Golob et al., 2004a, 2004b; Lee et al., 2006a, 2006b). Golob et al. (2004a) classified freeway traffic flow before crash occurrences into different states using clustering analysis, and then conducted nonlinear canonical correlation analysis to relate the characteristics of crashes to different traffic states. A procedure was also developed to predict the type of crashes that were most likely to occur for the traffic states being monitored (Golob et al., 2004b). Lee et al. (2006a, 2006b) compared the traffic flow conditions prior to sideswipe and rear-end crashes. A logistic regression model was developed to estimate the relative risks of sideswipe crashes compared to rear-end crashes given the fact that a crash has occurred. This type of models mainly focuses on the relative risk of a particularly type of crashes given the fact that a crash has occurred. Accordingly, they cannot be used to predict the probability of crash occurrences.

More recent studies took into consideration the traffic flow data both prior to crashes and in crash-free conditions when developing crash risk models. Most of these studies followed a case-control study design in which the traffic conditions that were associated with crash occurrences and normal traffic conditions were investigated to identify crash precursors and their effects on crash risks. The use of the case-controlled dataset allows for the measurement of the relationship between traffic conditions and crash risks. The matched and unmatched case-control designs are two dominant approaches for modeling crash risks. The major difference between these two approaches lies in the methods of selecting non-crash cases. In the matched case-control design, the non-crash cases are matched with crash cases according to some confounding factors such as the time and the locations of crashes, while in the unmatched case-control design the control samples are randomly selected.

Both matched and unmatched case-control designs control for the impacts of confounding variables (Bruce et al., 2008). The major difference is that the matched case-control design accounts for the impacts of confounding factors at the stage of selecting controls; while the unmatched case-control design takes into account the impacts of confounding factors at the stage of data analysis (Bruce et al., 2008). The case-control study design reduces the non-crash cases greatly and accounts for the selection bias. However, both matched and unmatched case-controlled samples are biased samples in which the crash cases are over-represented.

With the case-controlled dataset, the logistic regression models can be developed to link crash risks to crash precursors. The conditional logistic regression technique is usually used in the matched case-control design to account for the selection bias (Abdel-Aty et al., 2004, 2005; Abdel-Abdel-Aty and Pemmanaboina, 2006). In the conditional logistic regression model, the conditional likelihood function is used to make comparisons within each matched pair. Accordingly, it can deal with the confounding factors that lead to selection biases. In the unmatched case-control design, the conventional logit model is usually used (Xu et al., 2013; Ahmed et al., 2011).

The sensitivity and specificity have been widely accepted as the performance measures of the predictability of crash risk models. The sensitivity measures the proportion of the crash cases that are correctly identified. The specificity measures the proportion of the non-crash cases that are correctly identified, and (1-specificity) is usually called the false alarm rate, which represents the proportion of the non-crash cases that are mistakenly identified as crash cases. The use of specificity and sensitivity did not truly reveal the predictability of a crash risk model, and the reasons are twofold. First, most studies evaluate the sensitivity and false alarm rate on the basis of a case-controlled dataset. They are not indicative of the real prediction accuracy on the entire population. Second, because the crash risk models generate only the likelihood of crashes, a threshold needs to be selected to help identify crash precursors. The selected threshold heavily affects the sensitivity and false alarm rate that are associated with crash risk models. For example, one can deliberately increase the sensitivity, or in other words the prediction accuracy of crash cases, of crash risk models by setting up a very low threshold. In this condition, the false alarm rate will also get increased. The current practice is to select a threshold such that a balance can be reached between sensitivity and specificity. However, this cutoff is arbitrary in nature.

3. Predictability of crash risk models

The predictability of a crash risk model measures to what extend crashes are predictable given crash precursors. If the focus is on the prediction accuracy on the entire population, the predictability of crash risk models can be measured as the conditional probability of a crash given crash precursors. Based on the Bayes' theorem, the conditional probability of a crash given a crash precursor can be calculated as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)}{P(B)}P(A) \qquad (1)$$

where A represents the occurrence of a crash; B represents the crash precursor; $P(A|B)$ represents the risk of a crash when