Accident Analysis and Prevention 98 (2017) 295–302

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

A multivariate-based conflict prediction model for a Brazilian freeway

Felipe Caleffi [a], Michel José Anzanello [b], Helena Beatriz Bettella Cybis [a,*]

[a] Laboratory of Transport Systems, Federal University of Rio Grande do Sul, Porto Alegre, RS 90035-180, Brazil
[b] Department of Industrial Engineering, Federal University of Rio Grande do Sul, Porto Alegre, RS 90035-180, Brazil

ARTICLE INFO

Article history:
Received 22 July 2016
Received in revised form
30 September 2016
Accepted 20 October 2016
Available online 27 October 2016

Keywords:
Conflict prediction model
Bhattacharyya distance
Principal component analysis
Linear discriminant analysis
Brazilian freeway

ABSTRACT

Real-time collision risk prediction models relying on traffic data can be useful in dynamic management systems seeking at improving traffic safety. Models have been proposed to predict crash occurrence and collision risk in order to proactively improve safety. This paper presents a multivariate-based framework for selecting variables for a conflict prediction model on the Brazilian BR-290/RS freeway. The Bhattacharyya Distance (BD) and Principal Component Analysis (PCA) are applied to a dataset comprised of variables that potentially help to explain occurrence of traffic conflicts; the parameters yielded by such multivariate techniques give rise to a variable importance index that guides variables removal for later selection. Next, the selected variables are inserted into a Linear Discriminant Analysis (LDA) model to estimate conflict occurrence. A matched control-case technique is applied using traffic data processed from surveillance cameras at a segment of a Brazilian freeway. Results indicate that the variables that significantly impacted on the model are associated to total flow, difference between standard deviation of lanes' occupancy, and the speed's coefficient of variation. The model allowed to asses a characteristic behavior of major Brazilian's freeways, by identifying the Brazilian typical heterogeneity of traffic pattern among lanes, which leads to aggressive maneuvers. Results also indicate that the developed LDA-PCA model outperforms the LDA-BD model. The LDA-PCA model yields average 76% classification accuracy, and average 87% sensitivity (which measures the rate of conflicts correctly predicted).

1. Introduction

Freeway safety is one of the most studied topics in transport engineering. Since crashes are a direct measure of traffic safety, many studies have focused on finding a relationship between freeway operational characteristics and collision risk. Real-time crash prediction models are often used to evaluate traffic collision risk based on aggregated traffic data gathered from loop detectors or surveillance cameras on freeways (Li et al., 2014).

Collision risk can be estimated as the probability of conflict occurrences based on existing freeway traffic flow characteristics, such as average speeds, flow, occupancy or density. A traffic conflict happens when two or more road users approach each other in space and time to an extent that there is a risk of crash if their movements remain unchanged (Davis et al., 2011). Collision risk prediction helps to identify hazardous traffic conditions where proactive crash prevention strategies are needed to mitigate the high crash probability. Several studies have proposed models for predicting freeway crash risk based on traffic operational characteristics. A systematic review on the impacts of traffic characteristics on freeway crash occurrence presented by Roshandel et al. (2015) summarizes the main papers in the area.

The literature does not report crash prediction models for Brazilians freeways. Brazilian freeways present high congestion levels and significant heterogeneity of traffic pattern among lanes. Such lanes usually present different average speeds, flow rates and traffic composition. Overtaking is frequent and, under these conditions, very dangerous. Therefore, the development of multivariate models aimed at predicting crash occurrence can help to identify hazardous traffic conditions, and to proposed strategies to enhance transport safety.

This paper seeks to understand the relationship between traffic characteristics and probability of conflicts occurrence on the Brazilian BR-290/RS freeway. For that matter, we first list a set of variables collected from surveillance cameras that potentially offer explanation on conflict occurrence. Next, we propose two alternative indices based on the Bhattacharyya distance (BD) and Principal Component Analysis (PCA) aimed at identifying the most relevant variables to categorize occurrences into conflict or non-

* Corresponding author.
E-mail addresses: f.caleffi@yahoo.com (F. Caleffi), michel.anzanello@gmail.com (M.J. Anzanello), helenabc@producao.ufrgs.br (H.B.B. Cybis).

http://dx.doi.org/10.1016/j.aap.2016.10.025
0001-4575/

conflict classes. Those indices guide the removal of less important variables in the selection process. After each variable is deleted, we assess the classification ability of the remaining variables to categorize events into conflict or not. The selected subset of variables is then inserted into a Linear Discriminant Analysis (LDA) model developed to estimate the conflict probability through a matched control-case using traffic data collected from surveillance cameras.

The contributions emerging from this manuscript are two-fold. First, it proposes a multivariate-based framework to select the most relevant variables describing factors related to traffic conflicts aimed at classifying such instances into proper categories (e.g.; conflicting our non-conflicting). In that selection process, a first step relies on data mining and multivariate tools (i.e., KNN, LDA, PCA and Bhattacharyya Distance) aimed at refining the initial set of variables; the retained variables, which allegedly present a higher discriminant ability, are then inserted in to a final LDA function. By doing so, we ensure that the final LDA model is not comprised of irrelevant or noisy variables, thus increasing its predictive power. Most previous studies do not carry out that selection step, and obtained models tend to present lower prediction accuracy. In addition, we understand this as the first time that the Bhattacharyya distance, a technique originally used in forensic and chemometrics schemes, was applied to a scenario dealing with transportation data aimed at identify the most relevant factors for classification of conflicting/non-conflicting instances. The second contribution comes from the particularities of the case study. This study focuses on a segment that represents traffic behavior on a typical Brazilian freeway access ramp. Traffic on these roads presents a highly heterogeneous behavior due to several factors: (i) the diversity of vehicle fleet, (ii) variable speed limits for different classes of vehicles, and (iii) drivers aggressiveness that leads to high number of lane changes and overtaking. These factors generate different traffic performances among freeway lanes.

2. Background

Logistic regression is one of the most used statistical techniques to link crash likelihood with traffic characteristics (Abdel-Aty et al., 2006, 2004; Islam et al., 2013; Li et al., 2014; Park and Oh, 2009; Xu et al., 2014, 2013, 2012; Zheng et al., 2010). Most studies applied matched case-control, where data were obtained from crash or conflict intervals and then matched with crash-free intervals. In general, results indicate that the models can help to predict crash probability satisfactorily, but inconsistent performance and high prediction errors were reported. Moreover, most studies do not report their models' predictive performance.

Several other studies have proposed the use of alternative techniques to evaluate collision risk, e.g. log-linear (Lee et al., 2003), Bayesian (Oh et al., 2005; Ahmed et al., 2012), probabilistic neural network (PNN) (Abdel-Aty et al., 2005); Abdel-Aty and Pande 2005), neural network (Pande and Abdel-Aty 2006), and support vector machine (SVM) (Yu and Abdel-Aty 2013). Lee et al. (2003) results suggested that the coefficient of variation (i.e., standard deviation of a variable divided by its average value) in speed, traffic density, and speed difference between upstream and downstream loop detector stations were significantly correlated with the crash risk. Ahmed et al. (2012) results indicated that average speed and speed standard deviation affected the likelihood of crash occurrence on freeways.

Abdel-Aty and Pande (2005) applied a probabilistic neural network (PNN) model to predict crash occurrences on freeways using multiple speed derivatives. In their propositions, logarithms of the coefficient of variation in speed were inserted into classification models. Pande and Abdel-Aty (2006) results indicated that the average speed, the difference in occupancy between adjacent lanes, and standard deviation of speed and volume contributed to increase crash risk.

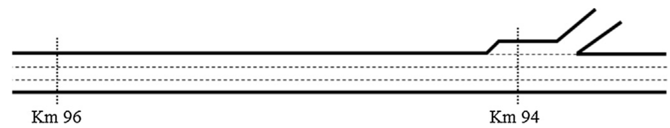

Fig. 1. Freeway section under study.

Most studies use aggregate traffic data obtained from loop detectors. Zheng et al. (2010) indicated that the speed standard deviation verified in intervals of 10 min could be considered a surrogate safety measure for collisions under congested traffic conditions. However, data is usually aggregated in 5 min intervals, ranging from 5 to 30 min prior to conflict or crash event. Abdel-Aty et al. (2004) reported that occupancy average upstream detector (0–5 min prior to collisions), and the coefficient of variation of speed downstream detector (5–10 min prior to collisions), have significant influence on collision occurrence. In addition, Islam et al. (2013) reported that occupancy average upstream detector (0–5 min prior to collisions), and the coefficient of variation of speed downstream detector (5–10 min prior to collisions), have significant influence on collision occurrence. Contradictory conclusions have been reported regarding the variables impact on the models; e.g. Lee et al. (2003) suggest that increasing the value of coefficient of variation of speed would reduce crash risk, while Abdel-Aty et al. (2006) reported opposite conclusions. These inconsistent findings may be due to different freeway geometry, traffic flow characteristics and human factors that influence drivability.

As claimed by Roshandel et al. (2015) the majority of previous studies did not seem to rigorously assess and/or report their models' predictive performance, such as false positive and negative prediction rates. Only a few studies have provided metrics to support the performance of proposed models: Abdel-Aty et al. (2004) stated that their model predicted 69% of crashes correctly with a false negative rate of 38.8%, and a false positive rate of 5.39%, while Hossain and Muromachi (2012) achieved 66% accurate predictions and 20% of false positive rates using a Bayesian-based approach. Hourdos et al. (2006) stated that their model prediction accuracy was 80%, with a false positive rate of 15%, while Islam et al. (2013) proposition yielded 68.5% accurate crash event prediction, and 26.1% false alarms. Since most studies did not present any measure of their model's performance, it becomes hard to evaluate how well these models perform in practice. A comprehensive validation of the model's accuracy is critical before any prediction model is implemented.

This paper presents a conflict prediction model based on a robust methodology for variables selection. The model development included most variables reported by previous studies, and the selection methodology allowed the identification of the most relevant variables that lead to the best prediction performance.

3. Study site and data processing

The study site was a segment of the Brazilian freeway BR-290/RS, kilometer 94, northbound direction, which is the main access to the city of Porto Alegre. The location was selected among other freeway sections based on several criteria: good quality of traffic surveillance cameras, congestion extent, and the existence of a high flow access ramp that disturbs traffic stream. The studied segment consists of a 3 lanes freeway with a single lane access ramp. Fig. 1 depicts the freeway section under study; Fig. 2 presents a speed-flow relationship for data collected in km 96 in May 2013, which illustrates the traffic differences between lanes. The left lane (Lane 1) features the highest speeds and flow rates, while the right lane