Transportation Research Part C 86 (2018) 202–219

Contents lists available at ScienceDirect

Transportation Research Part C

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

Real-time crash prediction in an urban expressway using disaggregated data

Franco Basso[a,*], Leonardo J. Basso[b], Francisco Bravo[c], Raul Pezoa[b]

[a] Escuela de Ingeniería Industrial, Universidad Diego Portales, Santiago, Chile
[b] Civil Engineering Department, Universidad de Chile, Santiago, Chile
[c] OPTILOG Consultant, Santiago, Chile

ARTICLE INFO

Keywords:
Real-time crash prediction
Support vector machines
Logistic regression
Automatic vehicle identification

ABSTRACT

We develop accident prediction models for a stretch of the urban expressway Autopista Central in Santiago, Chile, using disaggregate data captured by free-flow toll gates with Automatic Vehicle Identification (AVI) which, besides their low failure rate, have the advantage of providing disaggregated data per type of vehicle. The process includes a random forest procedure to identify the strongest precursors of accidents, and the calibration/estimation of two classification models, namely, Support Vector Machine and Logistic regression. We find that, for this stretch of the highway, vehicle composition does not play a first-order role. Our best model accurately predicts 67.89% of the accidents with a low false positive rate of 20.94%. These results are among the best in the literature even though, and as opposed to previous efforts, (i) we do not use only one partition of the data set for calibration and validation but conduct 300 repetitions of randomly selected partitions; (ii) our models are validated on the original unbalanced data set (where accidents are quite rare events), rather than on artificially balanced data.

1. Introduction

Car accidents in cities are an important externality caused by traffic. Accidents imply congestion, delays and sometimes fatalities. For example, in Chile, 1675 persons died in road accidents in 2016, the largest number in the last 8 years, while Rizzi and de Dios Ortúzar (2003) calculate that up to USD 1,300,000 are required in safety measures to avoid one death in interurban highways. Thus, understanding under what conditions accidents occur or, in different words, which traffic and external conditions increase the probability of a car accident, may have a sizeable impact. Furthermore, if those conditions were observed on line, then authorities or managers may have the chance to intervene in order to avoid accidents from happening. Nowadays, having traffic data on line is possible because of the new IT technologies which provides quality and bulk data to support monitoring traffic systems (Shi and Abdel-Aty, 2015). The purpose of this research is to study the precursors of car accidents in an urban expressway, using data that is available on-line to the expressway managers, in order to create a real-time accident prediction model which, in the future, may be transformed into a software tool. The on-line data is very rich: every car using this expressway has to have a transponder, so that the expressway can detect and charge them when they cross an Automatic Vehicle Identification (AVI) gate. One specific section of the expressway is studied, looking at data from AVI gates over a period of 18 months. We consider the afternoon rush-time, hence the focus is only on weekdays, using 80% (randomly selected) of the data for calibration purposes, while using the remaining 20% to test the predictive power of our model. Our results are promising: using the best classification model (logistic regression), we are able to

* Corresponding author.
E-mail address: franco.basso@udp.cl (F. Basso).

https://doi.org/10.1016/j.trc.2017.11.014
Received 28 January 2017; Received in revised form 5 October 2017; Accepted 13 November 2017
Available online 23 November 2017
0968-090X/

predict 67.89% of the accidents (sensitivity), while making only 20.94% of false predictions (false alarm rate). In the binary crash-prediction context, the false alarm rate is defined as the number of misclassified non-accident divided by the total number of observations. The sensitivity is defined as the total number of correct predicted accidents divided by the total number of accidents.

Our approach can be summarized in four steps: (i) The traffic data from AVI gates is aggregated to five minutes averages, and then used to calculate variables that are of interest, such as flows per type of vehicle, speeds, speed change, variance of speeds, density and density change. The data set then will have 0 and 1 s, corresponding to no accident or accident respectively. The data set is complemented from other sources that capture external conditions that may affect driving behavior such as, temperature, atmospheric pressure and rain. (ii) We then analyze this data both graphically and statistically, using a random forest procedure, in order to identify what are the variables that appear to be strong precursors of car accidents. (iii) The previous analysis are then used to calibrate two classification models, namely support vector machines (SVM) and logistic regression; for this, the first 80% of the data is used for calibration/training purposes and the remaining 20% for validation. (iv) In order to check for robustness of our models, the following is repeated 300 hundred times: randomly select 80% of the data base, calibrate both SVM and logistic models and then validate using the remaining 20%. This allow us to see dispersion in prediction power as the data changes, thus mimicking what would happen if an online prediction tool was at work, receiving new data continuously. With these results we compare the performances of our models.

There has been previous work on this area –some relevant references are reviewed below– however, there are two main general differences with previous efforts: data and the prediction/performance analysis. Regarding data, in this paper we work with data provided by a major tolled urban highway in Santiago, Chile, Autopista Central.[1] This highway spans for 60.5 km, crossing the metropolitan region from north to south, and connecting with the main interurban highway, Ruta 5. The highway is privately operated, and charge drivers according to the type of vehicle and distance by using AVIs and transponders installed in the vehicles. Since revenues come from AVIs, these devices have a very small failure rate, which enabled the acquisition of a detailed, disaggregated and rich traffic data set, that is, we know exactly at what time and at which speed each vehicle (separated by type) crossed an AVI. This contrast with previous efforts: as far as we know, the majority of the papers in the literature have worked with aggregated data, usually in periods of 30 s, without identification of type of vehicle, and using loop detectors which have a sizeable failure rate: according to Ahmed and Abdel-Aty (2012), loop detectors have a failure that ranges between 24% and 29%.[2] Even tough, last years some efforts have been made in order to include AVI data to analyze accident rates (Abdel-Aty et al., 2012; Xu et al., 2013; Yu et al., 2014; Shi et al., 2016). Disaggregated data differentiated by vehicle type allows us to explore a rather understudied issue: the influence of vehicle composition, and the corresponding speed differences, on the crash likelihood.

The second main difference with the previous literature is how the performance of the resulting models is tested. We improve on this issue on two aspects. First, all the papers reviewed below discarded some of the non-accident observations in order to 'balance' the data set and, then, calibrated the model using a fraction of the adjusted data set (typically 70% or 80%) while using the remaining observations for validation. This calibration technique, however, was extended to validation/prediction: to the best our knowledge all previous papers tested the model using the same artificially balanced data, that is, on data that does not show the actual, real pattern of accidents being rare events (Theofilatos et al., 2016). While for the calibration of one of our classification models we do balance the data set (the SVM case), in all cases the performance was tested by attempting to predict accidents using real data, where accidents are indeed very rare events. It is hard to say with certainty how the models calibrated on artificially balanced data would perform on a real-time environment yet, our conjecture is that they necessarily will do worse. Our second improvement is on the robustness of the models. As far as we know, in all papers calibration is made for just one partition of the data which raises the question of robustness: would the parameters of the model be the same if a different partition were used? And would predictive power (also called sensitivity) remain the same? To answer these questions, we created the additional 300 repetitions explained above, in order to calculate 300 values for sensitivity and false positive rates, obtaining then the averages, maximums, minimums and standard deviations. Hence, it is important to keep in mind that, while some papers reviewed below may present performances similar to ours, that performance was achieved –in contrast to our case– in a non-real environment and using just one partition of the data. As we explicitly show, it is quite possible that for that one partition, results end up being much better than for others. The power of the calibrated model was also tested on traffic and crash data that was collected by Autopista Central on a period of time later than the one we had at hand. This test is what comes close to learn what would have been the result should a real-time model been working. The sensitivity was actually better than before: we are able to predict 75.03% of the accidents.

We now briefly review some important references. Golob and Recker (2004) used k-clustering techniques looking at 1000 crashes occurred in 1999 in Southern California, in order to define taxonomies for the flow regimes previous to an accident. Note the emphasis here is on identifying flow regimes that make more likely that an accident will occur, rather than on attaching an actual probability of accident to a particular traffic condition. In the beginning of this project we tried to use k-clustering techniques but its performance was evidently inferior so we did not pursue this more. For a recent review of the effect of flow regimes and climate conditions see Theofilatos and Yannis (2014).

Abdel-Aty et al. (2004) used a logistic model, as we do, but in a matched case-control setting, implying that not all the non-accident data is used, as opposed to what we do. They looked at data from the Interstate 4 in 1999 obtained from the Orlando Police Department and loop detectors installed approximately 0.5 miles apart. This model has a predictive power of 67%. The false alarm

[1] www.autopistacentral.cl.

[2] Failure means rate that the loop does not capture well the speed or type of vehicle. Autopista Central also has loop detectors installed. Starting in 2016, the operator started improving all loop detectors to a more reliable one and, in the future, will provide much more reliable data.