Accident Analysis and Prevention 137 (2020) 105436

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

The importance of flow composition in real-time crash prediction

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

[a] School of Industrial Engineering, Pontificia Universidad Católica de Valparaíso, Chile
[b] Civil Engineering Department, Universidad de Chile, Chile
[c] Escuela de Ingeniería Industrial, Universidad Diego Portales, Chile
[d] Instituto Sistemas Complejos de Ingeniería (ISCI), Chile

ARTICLE INFO

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

ABSTRACT

Previous real-time crash prediction models have scarcely used data disaggregated by vehicle type such as light, heavy and motorcycles. Thus, little effort has been made to quantify the impact of flow composition variables as crash precursors. We analyze the advantages of having access to this data by analyzing two scenarios, namely, with aggregated and disaggregated data. For each case, we build Logistics Regressions and Support Vector Machines models to predict accidents in a major urban expressway in Santiago, Chile. Our results show that having access to disaggregated data by vehicle type increases the prediction power up to 30 % providing, at the same time, much better intuition about the actual traffic conditions that may lead to accidents. These results may be useful when evaluating technology investments and developments in urban freeways.

1. Introduction

1.1. The problem

Road accidents in cities are a significant externality caused by traffic, which then causes other costs such as congestion delays and, in many cases, fatalities. The number of accidents is increasing around the world, mainly due to an increase in the rate of motorization and distances traveled. For example, there were 94,879 road accidents in 2017 in Chile, the highest number ever. In that same year, 1483 people died in road accidents CONASET, 2018), with a sizeable financial impact. Previous studies calculate that safety measures designed to prevent one death on interurban highways could cost as much as US$1.3 M (Rizzi and Ortúzar, 2003; Iragüen and Ortúzar, 2004). Understanding the traffic and external conditions that increase the probability of a road accident could therefore have a major impact.

With traffic detection technology becoming increasingly common, a large amount of traffic data is being gathered by authorities and road and highway managers (Frez et al., 2019). The availability of such data has sparked research on a number of topics, one of them being road accidents. Multiple efforts have been deployed to identify the traffic conditions that lead to car crashes. Furthermore, some of these efforts have been aimed at the more difficult task of predicting car accidents based on real-time traffic conditions, given that the necessary communication technologies for the on-line availability of traffic data are already in place. Nevertheless, most studies are based on information obtained from inductive-loop detectors (ILD), which generally are not or cannot be used to obtain vehicle-by-vehicle (VBV) data or to classify vehicles by type due to its low rate of accuracy (Ki and Baik, 2006). Thus, typically, modeling variables – such as flow volume, lane-occupancy, and average speed – are aggregated over vehicle types and over time. Even though loop detectors usually collect data every 20 or 30 s, when an analysis is performed, the data is generally aggregated into longer intervals (5 − 10 min) to smooth fluctuations (Parsa et al., 2019) and to avoid inconsistencies about the precise time of crashes (Golob and Recker, 2003).

Furthermore, studies based on data from ILD are inherently subject to the unreliability of these devices, something that curbs the quality of predictive models. Ahmed and Abdel-Aty (2012) state that, due to environmental conditions, the failure rates of ILD vary between 24 % and 29 %. Ahmed and Abdel-Aty (2012) improved this using data from free-flow toll gates with Automatic Vehicle Identification (AVI). This system rarely fails because the detection is usually used to collect different fares for different vehicle types. Yet, even in this case, the authors do not seem to have access to vehicle classification. Consequently, little emphasis has been placed on the relevance of flow composition in (real-time) crash prediction, and even fewer studies have focused on the impact of the presence or absence of different types of vehicles on the driving behavior leading to accidents.

Even though other technologies may also be used to obtain

* Corresponding author at: School of Industrial Engineering, Pontificia Universidad Católica de Valparaíso, Chile.
E-mail address: francobasso@gmail.com (F. Basso).

https://doi.org/10.1016/j.aap.2020.105436
Received 27 May 2019; Received in revised form 6 January 2020; Accepted 7 January 2020
Available online 31 January 2020
0001-4575/

information at a microscopic level, the availability of traffic data by vehicles type is far from obvious. Recent research has used video images to collect VBV data; Van Beinum et al. (2018) use empirical trajectory data collected from a video camera mounted underneath a hovering helicopter camera to analyze the driving behavior at expressway ramps and weaving segments. Similarly, Gu et al. (2019) study the crash risk at interchange merging areas using an unmanned aerial vehicle (UAV). This technology provides individual VBV data, which enables conducting traffic analyses at a microscopic level, yet its application to real-time remains a challenge because of the time needed to process the video images and the rather expensive application on a continuous basis. Other technologies, such as cellular phone data has also been used. For example, Yuan et al. (2018) use Bluetooth data for real-time safety analysis on urban arterials. Even though VBV data seems to be available in this case, the impact of signal delay appears as a significant drawback in real-time application, together with partial penetration of the technology.

The objective of this paper is to analyze the importance of having access to disaggregated data per vehicle type for predicting crashes. In contrast to most of the previous efforts, this study uses a very rich dataset, provided by Autopista Central, an urban freeway in Santiago, Chile. This freeway is privately operated, and charges users through information obtained by AVI gates which communicate with transponders that are installed in all vehicles, as required by law. The fare charged depends on the distance traveled, which is obtained from the AVI gates crossed, and the vehicle type. Considering that the revenue of the freeway comes from these devices, the failure rate and classification errors are very small (less than 1 %). Nevertheless, due to the high cost of installing an AVI gate (which could exceed 500,000 USD), these devices are sparsely spaced within the study corridor. This feature is a drawback for the prediction performance. Still, the focus of our paper is on the relative improvement caused by flow composition data acquisition rather than absolute performance and, in that context, we still obtain that good prediction power is possible.

To obtain our results and conclusions, we build two logistic regression (LR) models and two Support Vector machine (SVM) models. The first LR and SVM models use aggregated data, that is, without information separated by vehicle type. In this case, the logic we pursue is to replicate similar previous studies (Abdel-Aty and Pande, 2005; Pande and Abdel-Aty, 2006a, b; Abdel-Aty et al., 2007; Xu et al., 2013; Theofilatos, 2017; Yang et al., 2018a), in which only aggregated data is available. The second LR and SVM models use variables separated by vehicle types. We quantify the improvement of having access to disaggregated data by analyzing the difference in the prediction power (sensitivity) of disaggregated vs aggregated models, while comparing across the two classification methods in order to assess if this matters for our conclusions. It is worth noting that the focus of this paper is not to advance in analytical methods in crash prediction –though we do use the current published state of the art– but to identify and quantify the impact of variables that have been rarely analyzed in previous studies (flow composition variables).

1.2. Literature review

There has been previous work on this area –some relevant references are reviewed below– however, there are two general differences concerning earlier efforts. First, even when there have been some studies that use flow composition variables, these works have mainly used aggregated crash frequency. To the best of our knowledge, this is the first research that quantifies the effect of having access to data disaggregated by vehicle type in a real-time setting. Second, most of the previous contributions have used sampling methods to balance the training data set. Yet, except for a handful of studies (Basso et al., 2018; Parsa et al., 2019; Yuan et al., 2019), the use of artificially balanced data was extended to the validation phase which does not show the actual, real pattern of accidents being rare events (Theofilatos et al., 2018). It is not simple to conjecture how the models calibrated but also tested on artificially-balanced data, would perform in a real-time environment, although they most likely will do worse. In this paper, we use the full unbalanced data set to analyze how the flow composition affect real-time crash prediction models. This, we think, allow us to assess if the flow composition matters for the performance of crash prediction models in real-time computational tools, like the one currently working in Autopista Central, Chile[1] .

We now review the literature focusing only on articles that attempted to predict or explain accidents using variables of flow composition. For a complete review of real-time crash prediction models, we refer the reader to Hossain et al. (2019).

Montella et al. (2008) use a generalized linear model with data coming from an unspecified technology for a stretch of a rural freeway in Southern Italy, to estimate how the increase in the percentage of heavy vehicles flow affects severe crashes frequency. The authors argue that when heavy vehicles circulate through the freeway, drivers might increase their risk perception and thus decrease their speed, reducing the crash severity, though this hypothesis is not tested.

Dong et al. (2014) use a multivariate Poisson-lognormal model to study car-truck interaction and their repercussion in crash frequency at urban signalized intersections using yearly data for Tennessee from 2005 to 2009. They established three types of crashes depending on the type of vehicles involved: (1) car crashes, (2) car -truck crashes and (3) truck - crashes. Besides geometric and environmental variables, they find that the truck percentage in the traffic stream is relevant, with car crash involvement rate decreasing as the number of trucks increases. They believe that this may be explained due to the fact that, for a constant vehicle density, a larger truck percentage implies fewer lane changing and overtaking movement by cars. Opposed to our work, the previous two papers, do not seek to predict accidents but to compute correlations between the explanatory variables and the crash frequency.

More recently, Theofilatos et al. (2018) find that, for a motorway located in Athens, Greece, the proportion of trucks in traffic does not affect crash occurrence. Nevertheless, the emphasis is put on the correction of the bias generated by the small number of crash records compared to non-crash events, rather than flow composition effects. Also, the data used comes from ILD and the issue of whether the proportion of trucks is accurate or not is not addressed. In terms of methodology, this study applies binary logistic regression models for the prediction, which has been used with relative success for real-time crash prediction (Abdel-Aty et al., 2004; Ahmed and Abdel-Aty, 2012; Xu et al., 2013; Theofilatos, 2017; Basso et al., 2018) and the machine learning algorithm, support vector machines, which was used by Yu and Abdel-Aty (2013).

Basso et al. (2018) also uses data from AVI gates as we do. However, the authors focus on a very central stretch of Autopista Central in Santiago, Chile, so the proportion of vehicles other than light is negligible. This does not allow the authors to determine the real impact of having access to disaggregated data. Contrarily, this paper study a section where the proportion of buses and trucks is considerably higher.

Dimitriou et al. (2018) assess rear-end crash potential in urban locations using ILD data. The authors analyze vehicle-by-vehicle interactions that take two types of vehicle into account, namely heavy goods vehicles (HGV) and passenger cars. As a behavior conclusion, they find that speeds were lower and headways higher when HGVs were ahead. As is common in the literature, and in contrast to our case, the models were not validated through an actual crash database.

Choudhary et al. (2018) use a multivariate Poisson lognormal regression to evaluate the impact of speed variations on crashes divided by severity and vehicle type. Contrary to what we do here, the authors analyze crash rates instead of proposing a real-time model. They also use the vehicle type as a dependent variable instead of independent

[1] www.autopistasegura.cl

variables as we do.

Finally, Wang et al. (2019) use traffic conditions along with vehicles' trajectories for real-time safety analysis. The authors find, as we do, that the change of truck percentage introduces turbulence increasing the crash likelihood. Opposed to our research, the authors use a matched-case control environment instead of a full unbalanced real dataset.

The rest of the article is organized as follows. Section 2 describes the data and preparation process. Section 3 presents the theoretical underpinnings of the methods used for both, variable selection and to fit the models (classification methods). Section 4 describes the Logistic regression models, first using aggregated data, then using disaggregated data and then comparing their performance. Section 5 contains similar analyses but for the SVM models. Section 6 concludes.

2. Data preparation and descriptive analyses

The study focuses on an 8.9 km portion of Autopista Central, an urban highway that spans 60.5 km crossing Santiago, Chile[2] in the north-south direction through two sections, General Velásquez and Ruta 5 as shown in Fig. 1. Every vehicle using the freeway is required by law to have a transponder which communicates with AVI gates, information that is then used by Autopista Central to charge the users based on the distance travelled and type of vehicle, classified in one of three categories: light (up to small commercial vehicles), heavy (all types of large trucks or buses) and motorcycle. Moreover, the AVI gate – transponder technology is able to capture the time and speed of each (type of) vehicle passing an AVI gate. This allows the Autopista Central (and us) to report flows and speed of each type of vehicle to the authorities. Autopista Central provided us with all this disaggregate detailed data from November 1 st, 2014 to April 30th, 2016 for every one of the thirty-one AVI gates present at the freeway, which enable us to calculate many different variables, averaged as we see fit. Note that, by focusing only on one segment, and not several segments, road geometry is kept fixed and, therefore we will not be using geometry variables in our model. This, in order to focus clearly on the importance of traffic flow composition.

Autopista Central also provided manually recorded crash data, which includes the time and exact location of every accident, determined using recording cameras, which completely cover the highway. Every crash record was then assigned to the sections used by Autopista Central for management. In this study we focus on the south direction of the lowest part of General Velásquez section, shown in Fig. 2. This section combines a sufficient number of accidents over the studied period (68 accidents) for empirics to be performed, and a high participation of vehicles other than light which, we think, allows to pursue our research question. This section spans for 8.9 km and has two AVI gates (PA-22 and PA-20 according to the direction of traffic; see Fig. 2), eight entry ramps (four of them in the section between the AVI gates) and six exit ramps (four of them in the section between the AVI gates). We also considered data from the upstream AVI gate, namely PA-24, which is only one kilometer away from the studied section.

We decided to focus on the afternoon rush hour, defined as the period between 5:30p and 8:30p during weekdays because it corresponds to the evening rush hour in which 19 of the 68 accidents of the section occurred. That is, 28 % of the accidents are concentrated in only 13 % of the day. The data was aggregated to 5-min averages, similar to what was done by Ahmed et al. (2012a) and Basso et al. (2018).

For the case of disaggregated data, we calculate 8 variables for each AVI gate and type of vehicle: flow, mean speed, standard deviation of the speed, percentage of such type of vehicle in total flow; and also, the change in all of those variables compared to the previous 5-min interval. Besides, we compute the traffic density for each 5-minute interval. Since we are studying mixed traffic, we follow the Highway Capacity Manual (Transportation Research Board, 2000) to calculate the density through a passenger-car equivalent flow rate approach using the most recent passenger-car equivalent factors for Chile (SECTRA, 2013). The average densities over the studied period are 29.7, 39.4, and 40.7 [pce/km] for gates 20, 22, and 24, respectively. Note that these calculations are only possible when vehicle classification is available.

On the other hand, for the aggregated data case, we compute the same 8 variables mentioned before but not differentiated by vehicle type. Also, as we are assuming that vehicle classification information is not available, we calculate a proxy of density for each 5-minute interval, simply defined as the quotient of total flow [veh/h] and average speed [km/h]. Thus, this approach assumes that traffic is homogeneous. Although this measure is only a proxy of the real traffic density, for the sake of simplicity, we will refer to it as density. Finally, for both disaggregated and aggregated case, we include the change of each density measure compared to the previous 5-min interval as explanatory variables.

A summary with some basic descriptive statistics for the variables considered is shown in Table 1 for AVI gate PA-20, Table 2 for AVI gate PA-22 and Table 3 for AVI gate PA-24.[3] The final variable considered was binary and indicates whether an accident occurred during the next five minutes' interval or not. This definition enables using the estimated model to predict crash occurrence using current traffic conditions. In the case of the model with aggregated data, we obviously calculate a single value for flow, mean speed, standard deviation of the speed and density for each 5-min interval and AVI gate, that is, we aggregate all types of vehicles into a unique category, just as if there was no classification information available. Evidently, we do not use the composition variable for each type of vehicle either.

After preparation, the dataset consists of 13,350 observations (5-min intervals), where only in 19 of them (0.14 %) an accident occurred.[4] This is known in the statistics literature as "rare events" and induce a number of difficulties, particularly for classification, since a model that always predict no accident would be a very good model almost always. From Tables 1 and 2 we can see a high share of heavy vehicles in flow composition, averaging a 12.4 % in the upstream AVI gate (PA-22) and 16.9 % in the downstream AVI gate (PA-20). This percentage is explained because Autopista Central offers reduced fares to this type of vehicles in this section, compared to the parallel North/South section where heavy vehicles account for only 4 % of the flow in the same period of the weekdays (Basso et al., 2018). It is possible to notice also a low participation of motorcycles, compared to the North/South section where this type of vehicle represents around 3 % of the flow, which is probably induced by the high participation of buses and trucks. In fact, there are many periods where no motorcycles passed through any of the AVI gates: 1304 and 574 observations of the AVI gate PA-20 and PA-22 respectively have zero motorcycles, representing the 9.8 % and 4.3 % of the total data of each gate.

[2] Santiago: Population: 5,822,316; Area: 641km^2; Motorization Rate: 177 [veh/1000 inh]; Motorized travels per day: 10,792,200; Modal split: Public 46.9%, Private 46.4%, Other 6.7%.

[3] As a referee correctly points out, traffic flow is sensitive to some road geometry characteristics, in particular entry and exit ramps. Table A1 in the Appendix shows that a non-negligible percentage of the flow enters or exits the highway between two AVI gates.

[4] Prediction of rare events such as crash accidents is indeed a major challenge and generally implies working with few accidents data. For example, the results of Theofilatos et al. (2018), were conducted using a data set that consists of 17 crashes and 91,118 non-crash cases (crash ratio: 0.019%). Parsa et al. (2019) data set consists of 32 crashes and 85,182 non-crash cases (crash ratio: 0.038%).

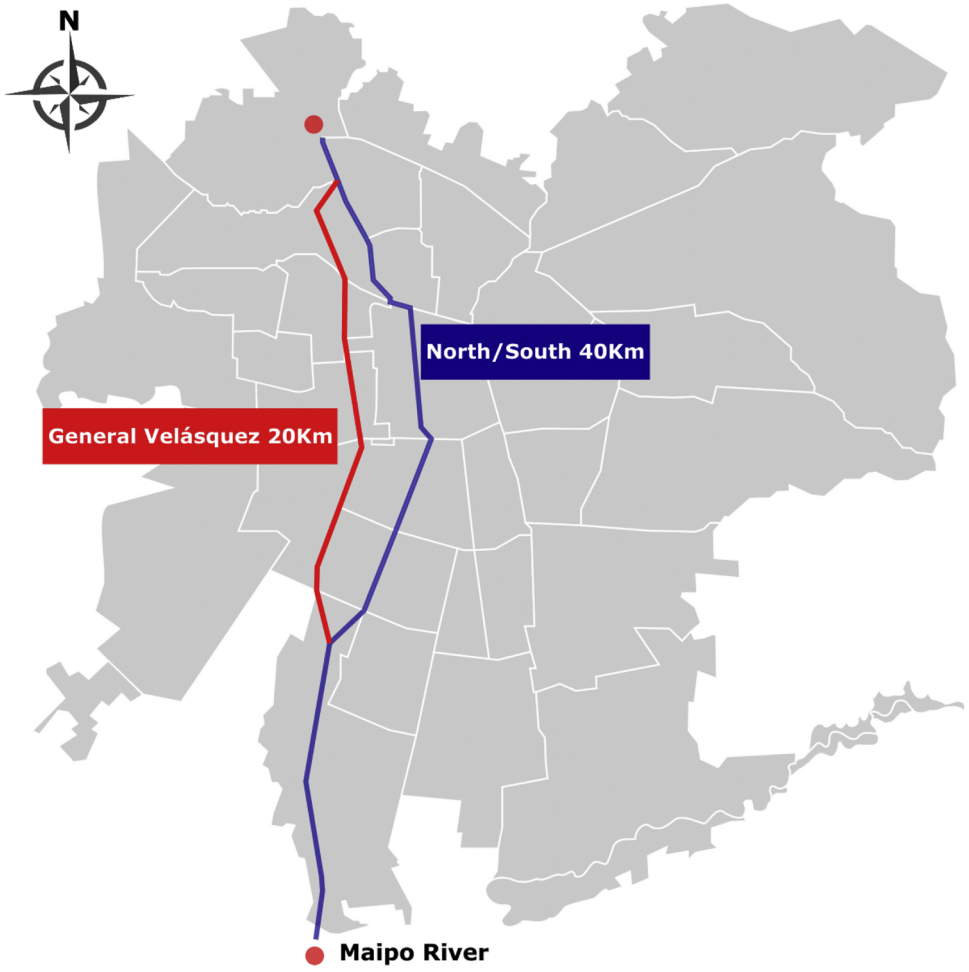

Fig. 1. Autopista Central, Santiago, Chile.

3. Methodology

In this Section, we briefly present the theoretical underpinnings of the methods we use, both for selecting relevant variables as for classification. We choose these methods because they have been widely and successfully used in previous studies with high predictive power, allowing us to keep the focus in the objective of the paper, namely, to determine the relevance of flow composition in real-time crash prediction. Nevertheless, for a recent comprehensive review of big data and machine learning methodologies applied to road safety, we refer the reader to Stylianou et al. (2019).

3.1. Random forest for variable relevance

In order to select the best variables for both models (with aggregated and disaggregated data), a Random Forest (RF) classifier is used. RF is an ensemble classifier that constructs many decision trees, where each tree cast a unit vote for the most popular class. Then, the class which has been assigned most times is selected (Breiman, 2001).

RF logic is based on two techniques, namely, bootstrapping aggregating and feature bagging. Every one of the T trees used is trained using a bootstrap sample of the original data and m random features. Then, this tree is used for test purposes only in the instances not contained in the bootstrap sample used to train, called out-of-bag (OBB) data, which should account for around one third of the total data. The use of the OBB data allows RF to compute an unbiased estimator of the classification error. Nevertheless, this estimator depends on the number of features m used for each tree. A large value of m leads to high correlation between different trees, increasing the OBB error. On the other hand, it also increases the strength of each tree which implies a lower error. The use of a small value of m produces the opposite effect. For the purpose of this study, we use the suggested value $m = log_2(M + 1)$, where M is the total number of available features.

The measure used to rank the variables importance will be the mean decrease in Gini. The Gini impurity index (Breiman et al., 1984) is defined for the node t as:

$$i(t) = 2p(1/t)p(2/t)$$

where $p(i/t)$ is the probability of class i given node t. Every time we split a tree using the variable u, the sum of the impurity of the children nodes is less than the impurity of the parent node. The decrease in impurity is then defined as this difference. The mean decrease Gini is defined as the average of each variable over all the trees.

RF for feature selection has been widely used in the context of real-time crash prediction (Abdel-Aty et al., 2008; Ahmed and Abdel-Aty, 2012; Xu et al., 2013; Wang et al., 2015; Lin et al., 2015; Basso et. al, 2018).

3.2. Logistic regression for classification

Logistic regression is a method that can be used to classify binary outcomes. In particular, given a set of predictors x, the probability of a crash occurrence p(x) is defined in the classical fashion:

$$\text{logit}(\text{p}(\text{x})) = \beta_0 + \beta^T \text{x}$$

where β_0 and β are estimated by maximum likelihood and

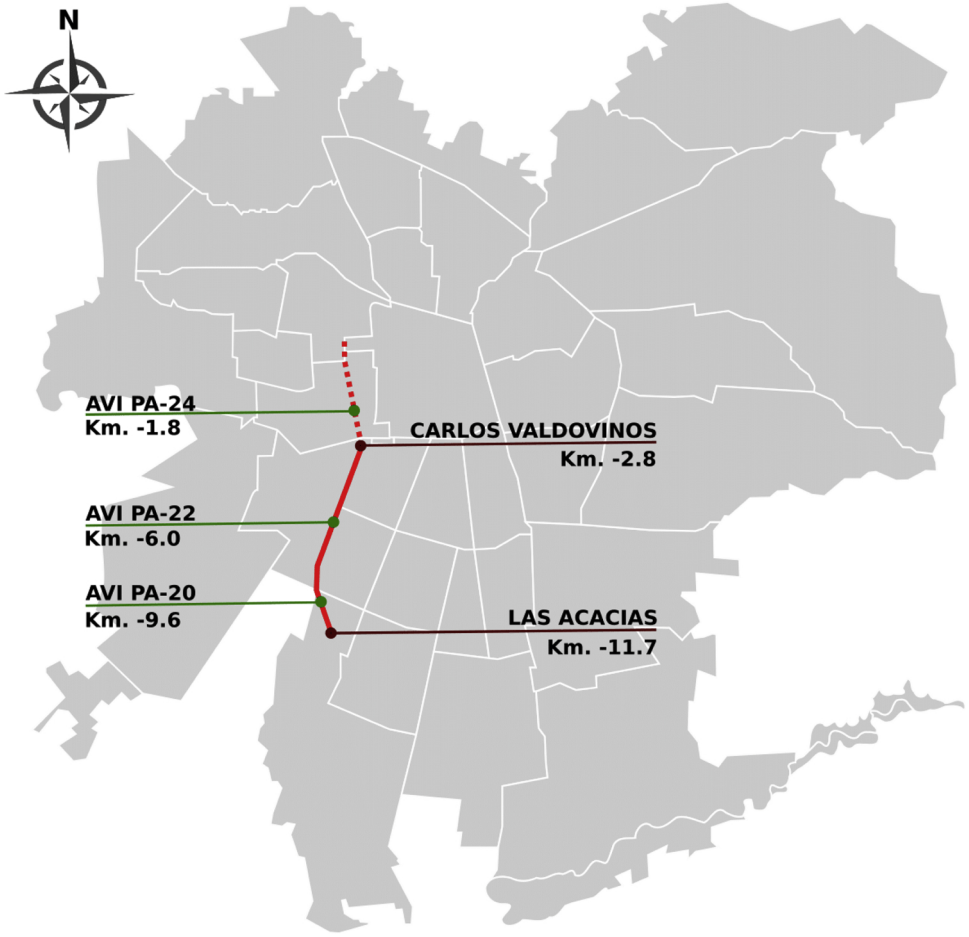

Fig. 2. Studied section of Autopista Central.

Table 1
Descriptive statistics of AVI gate PA-20.

AVI PA-20					
		Average	Std. Dev.	Minimum	Maximum
Light	Speed [km/h]	91.4	11.8	9.1	108.4
	Flow [veh/5 min]	157	34.6	44	281
	% Composition	81.7 %	4.0 %	59.9 %	99.0 %
Heavy	Speed [km/h]	82.3	10.4	6.5	102.5
	Flow [veh/5 min]	32.5	10.1	0	71
	% Composition	16.9 %	3.9 %	0.0 %	39.0 %
Motorcyles	Speed [km/h]	85.1	16.9	5.5	186.9
	Flow [veh/5 min]	2.7	1.9	0	15
	% Composition	1.4 %	0.9 %	0.0 %	7.4 %

Table 2
Descriptive statistics of AVI gate PA-22.

AVI PA-22					
		Average	Std. Dev.	Minimum	Maximum
Light	Speed [km/h]	88.5	12.6	12.9	107.0
	Flow [veh/5 min]	228.4	51.5	62	418
	% Composition	86.2 %	2.8 %	73.5 %	100 %
Heavy	Speed [km/h]	82.7	12.1	7.1	102.0
	Flow [veh/5 min]	32.8	9.8	0	68
	% Composition	12.4 %	2.7 %	0.0 %	25.7 %
Motorcyles	Speed [km/h]	87.2	13.5	3.3	187.6
	Flow [veh/5 min]	4.0	2.4	0	18
	% Composition	1.5 %	0.8 %	0.0 %	7.7 %

Table 3
Descriptive statistics of AVI gate PA-24.

AVI PA-24					
		Average	Std. Dev.	Minimum	Maximum
Light	Speed [km/h]	74.1	13.8	7.8	98.8
	Flow [veh/5 min]	239.8	62.0	36	393
	% Composition	86.8 %	2.8 %	57.6 %	100 %
Heavy	Speed [km/h]	70.4	11.7	6.8	100.5
	Flow [veh/5 min]	31.8	10.1	0	62
	% Composition	11.5 %	2.6 %	0.0 %	30.3 %
Motorcyles	Speed [km/h]	72.7	20.4	11.4	193.2
	Flow [veh/5 min]	4.5	2.7	0	20
	% Composition	1.6 %	1.2 %	0.0 %	12.1 %

$$\text{logit}(p(x)) = \log\left(\frac{p}{1-p}\right)$$

so,

$$p(x) = \frac{1}{\exp(-\beta_0 - \beta^T x) + 1}$$

This classification method has been used by many previous studies (Ahmed et al., 2012a; and 2012b; Theofilatos, 2017; Theofilatos et al., 2018; Basso et al., 2018) and, as opposed to non-parametric classification methods, it has the upside that parameters are very easy to interpret: a positive parameter β_i for the predictor x_i indicates a positive effect on the crash probability. Moreover, other useful indicators such as (point) elasticities may be calculated.

3.3. Support vector machine for classification

In the context of binary classification, Support Vector Machines (SVM) seek to find a separator hyperplane $f(\boldsymbol{z}) = w\boldsymbol{z} + b$ that maximizes the distance between the hyperplane and the data points in the two classes. For the case where data is not linearly separable, Cortes and Vapnik (1995) introduced the so-called *soft-margin* separator hyperplane, which allows misclassification by using a penalization parameter C. If we consider the labeled data $(\boldsymbol{x}_1, y_i), \ ..., (\boldsymbol{x_n}, y_n) y_i \in \{-1,1\}, \boldsymbol{x_i} \in \mathbb{R}^d, i = 1, \ ..., n$, such hyperplane can be found by solving the optimization problem:

$$min \frac{1}{2} ||\boldsymbol{w}||^2 + C \sum_i \xi_i$$

$$s.\ a.\ y_i(\boldsymbol{x_i} \cdot \boldsymbol{w} + b) \geq 1 - \xi_i, \ i = 1, \ ..., n$$

$$\xi_i \geq 0, \ i = 1, \ ..., n$$

The intuition for the parameter C is then straightforward, because it controls the tradeoff between misclassification and the classifier's margin. A small value of C will cause a separator hyperplane with a large margin, at the cost of possibly more points being misclassified. On the other hand, a large value of C will produce a separator hyperplane with a smaller margin, but with more training points being correctly classified.

Cortes and Vapnik (1995) also showed that the hyperplane mentioned has the form

$$f(\boldsymbol{z}) = \sum_S \alpha_i \boldsymbol{x_i} \cdot \boldsymbol{z} + b$$

with S the set of support vectors, that is, the data points x_i that either attain the minimum margin or violate it.

The values of $\Lambda = (\alpha_i)_{x_i \in S}$ can be found solving the maximization problem

$$\max W(\Lambda, \delta) = \Lambda^T 1 - \frac{1}{2}\left(\Lambda^T \boldsymbol{D} \Lambda + \frac{\delta^2}{C}\right) (P_2)$$

$$\text{s.t. } \Lambda^T Y = 0$$

$$\delta \geq 0$$

$$0 \leq \Lambda \leq \delta 1$$

where $Y = (y_i)_{x_i \in S}$ correspond to the labels of the support vectors and $\boldsymbol{D}$ is a symmetric matrix with elements

$$D_{ij} = y_i y_j \boldsymbol{x_i} \cdot \boldsymbol{x_j} \ \forall \ i, j \text{ such that } \boldsymbol{x_i}, \boldsymbol{x_j} \in S$$

By replacing the dot product for a general form of the dot product in a Hilbert space, $K(\boldsymbol{u}, \boldsymbol{v}) = \phi(\boldsymbol{u}) \cdot \phi(\boldsymbol{v})$, ϕ: $\mathbb{R}^n \rightarrow \mathbb{R}^N$, the separator hyperplane is then expressed as

$$f(\boldsymbol{z}) = \sum_S \alpha_i \phi(\boldsymbol{x_i}) \cdot \phi(\boldsymbol{z}) + b$$

mimicking the case where the original feature space is mapped into a higher dimensional space using the vector function ϕ

$$\phi(\boldsymbol{x_i}) = \phi_1(\boldsymbol{x_i}), \phi_2(\boldsymbol{x_i}), \ ..., \phi_N(\boldsymbol{x_i}) \ \forall \ i \text{ such that } x_i \in S$$

allowing for non-linear decision boundaries in the original feature space. The values of $\Lambda = (\alpha_i)_{x_i \in S}$ can be found solving the same maximization problem (P_2), but where now $\boldsymbol{D}$ is a symmetric matrix with elements

$$D_{ij} = y_i y_j \phi(x_i) \cdot \phi(x_j) \ \forall \ i, j \text{ such that } x_i, x_j \in S$$

As remarked by Cortes and Vapnik (1995), the function $K(\boldsymbol{u}, \boldsymbol{v})$ used must satisfy the condition

$$\int K(\boldsymbol{u}, \boldsymbol{v}) g(\boldsymbol{u}) b(\boldsymbol{v}) d\boldsymbol{u} d\boldsymbol{v} > 0$$

For all g such that $\int g^2(\boldsymbol{u}) d\boldsymbol{u} < \infty$. The general dot products used in this paper are the classical ones that satisfy the given condition:

1 Radial: $K(\boldsymbol{u}, \boldsymbol{v}) = \exp(-\gamma \ ||\boldsymbol{u} - \boldsymbol{v}||^2)$
2 Polynomial: $K(\boldsymbol{u}, \boldsymbol{v}) = (\gamma \boldsymbol{u} \cdot \boldsymbol{v} + 1)^q$, with $q = 2,3$
3 Sigmoid: $K(\boldsymbol{u}, \boldsymbol{v}) = \tanh(\gamma \boldsymbol{u} \cdot \boldsymbol{v} + 1)$

3.4. Oversampling techniques

As discussed in Section 2, accidents are infrequent events, accounting for only 0.14 % of the data in this study. This can be troublesome because the simplest and most effective classifier (in terms of accuracy) will be the one that classifies almost all instances as negative (Akbani et al., 2004). There are multiple ways to tackle this issue, usually called class-imbalance, see e.g., Weiss (2004) for a review. In this paper, we explore two techniques. The first one simply oversamples the minority class; the second, more advanced, use the Synthetic Minority Oversampling Technique (SMOTE).

The basic oversampling method deals with class-imbalance by duplicating minority class examples. In other words, to balance the dataset this method keeps all the majority class observations, while making n exact copies of every minority class observation. In this study, we try different values of n, in order to oversample the minority class (for training purposes only) up to a proportion ranging from 60 % to 100 % of the majority class. The exact value of n is chosen via cross-validation. For this oversampling method, all of the majority class (non-crashes) cases are considered, i.e. no undersampling method is used. Even though this method does not increase information, it increases the misclassification cost of the minority class (Mani and Zhang, 2003). This method has been used by Mussone et al. (2017) in a related study, in order to balance a dataset of crashes and its respective injury severity.

Since the basic oversampling method involves making exact copies of the minority class data, it could lead to overfitting, and thus, may not significantly improve minority class recognition (Chawla et al., 2002). To overcome this problem, we use the Synthetic Minority Oversampling Technique (SMOTE), introduced by Chawla et al. (2002). This method oversamples the minority class by randomly creating synthetic data points among the minority class data points and their k nearest neighbors, while the majority class is undersampled by randomly removing samples from the majority class population until the minority class becomes some specified percentage of the majority class. This method has been used to balance datasets for real-time crash prediction (Basso et al., 2018) and real-time accident detection (Parsa et al., 2019). Parsa et al. (2019) also use two modifications of the SMOTE method, namely, borderline SMOTE and SVM SMOTE. Although in their case, regular SMOTE proved to achieve the highest performance. Thus, we decided to use the regular SMOTE technique in this study.

3.5. Validation

To assess the performance of the calibrated models we do not use a matched-case control methodology as other papers in the literature; instead, we use a 5-fold cross-validation procedure in the full unbalanced data set. Breiman and Spector (1992) showed that this procedure helps reducing bias in a regression setting. The crash to non-crash ratio is 0.14 % showing the rare event pattern discussed above.

The 5-fold cross-validation procedure is as follows: we first partition the data set randomly into 5 subsets of equal size, and then calibrate using 80 % of the sample (four subsets, 10,680 rows in this case) while validation is done using the remaining 20 % of the sample one subset, 2670 rows in this case where no balance technique takes place. In other words, we validate our models over the full unbalanced validation dataset. The procedure is repeated five times by choosing different subsets for validation. The overall sensitivity for the 5-fold cross-validation process is given simply by the average of the five sensitivities obtained.

Considering that in our context the minority class has very few observations (19 accidents), to ensure that the validation data set has an appropriate number of accidents, we use a stratified approach to select the five random subsets. That is, we choose the random subsets in a way that the class distribution in each fold is approximately the same as in the initial data set (Diamantidis et al., 2000). This is done by randomly dividing each of the two classes into groups of about 20 percent each and then combining (Breiman et al., 1984). Empirical studies show that stratified cross-validation gives better estimates for both bias and variance (Kohavi, 1996).

The mean sensitivity obtained through the 5-fold cross-validation process may be an inaccurate estimator of the actual predictor power of the model, because of the random nature of the five partitions. In order to obtain a more robust sensitivity value for the model, it is advised to perform many repetitions (Kim, 2009). We, therefore, performed 500 repetitions of the 5-fold validation process for each selected model. Thus, overall, we believe we have a very robust and strong methodology for validation, well-rooted in best practices.

4. Logistic regression models

4.1. Aggregated LR model

We start analyzing the case where only aggregated data is used, that is when all variables –flow, mean speed, standard deviation of the speed, density, and the changes in all of those variables compared to the previous 5′ interval– are calculated using all data regardless of the type of vehicle. By doing so we intend to replicate previous studies where vehicle classification is either not available or unreliable.

First, as explained in Section 3, we apply a Random Forest classifier to rank available variables according to their importance as crash precursors. From this analysis, synthesized in Fig. 3, the speed and density registered 5 min before an accident in the upstream AVI gate (PA22) seems to be the most important variables for crash prediction. Also, speed, density and speed change in the downstream AVI gate (PA20) are also relevant. This last variable has been found to be important as a crash precursor in previous studies (Abdel-Aty et. al, 2004; Abdel-Aty and Pande, 2005; Ahmed and Abdel-Aty, 2012).

With this information, multiple models were adjusted. We started by including the two most important variables, namely the speed and density registered 5 min before an accident in the upstream AVI. Fig. 4 shows all data points, with accidents in red, graphed according to those two variables. As it seems obvious, these two variables are not really useful to separate crashes from non-crash points. We can also observe that the hyper congestion part of the density-speed graph has very few points, implying that most of the time traffic conditions correspond to the congestion branch and, moreover, accidents mostly happen in those times.

Because using the two most important variables was not enough to separate safe from hazardous traffic conditions, we then searched for a third variable to include. For this we used the mean sensitivity over a 5-fold cross-validation as the performance indicator when each other possible variable was added. In each of the five repetitions (and for each variable tested) we set a threshold probability such that the model achieves a false positive rate (FPR) close to 20 % over the training dataset. This value for the FPR was chosen to make sensitivity comparable to what Basso et al. (2018) found, since their FPR was about 20 % in their best model. The overall procedure led to select the variable speed at the AVI gate PA20 as the third variable, which coincidentally is exactly the third most important variable according to the RF procedure. We settled with three explicative variables due to our small number of crash cases (19), and considering that a higher number of predictors could lead to biased regression coefficients and overfitting, which here degraded the predictive power of the model, as explained by Peduzzi et al. (1996). The degradation of the predictive power when including more than three variables is consistent with the fact that we could not find a model with four variables with estimated parameters different from zero at 90 % confidence level. Finally, the three variables model attains a mean sensitivity of 47.4 % and a mean FPR of 20.1 % in the 5-fold cross-validation.

The estimated parameters of the logistic regression over the full dataset are shown in Table 4; two of the parameters are different from zero at 95 % confidence level (but not 99 %), the third variable is different from zero at 90 % confidence level. The estimated values allow us to conjecture what are the conditions that lead to high-risk situations: lower than average speeds in both AVI gates together with a low density in the upstream gate increase the risk of a crash in the next five minutes. but other than that, the intuition of crash occurrence is not clear. We conjecture that an element is missing from this model, which makes difficult to fully understand the phenomenon. In the next section, we will explore this issue.

The decision frontier of the LR three variables model is shown in Fig. 5, with a threshold probability set at $p_0 = 0.177\%$. The model predicts that a crash will occur in the next five minutes for any point to the left of the plane. Changing the threshold probability moves the decision frontier in parallel fashion, moving to the right if the p_0 is set at a smaller value, creating a smaller “high-risk” zone, with a lower FPR but with also a lower sensitivity (Yang et al., 2018b). A larger threshold probability induces the opposite effect, causing a larger FPR and sensitivity. Considering that in a real-life context, taking a decision associated with this trade-off of FPR/Sensitivity is not easy, we will present in the next section the receiver operating characteristic (ROC) curve for

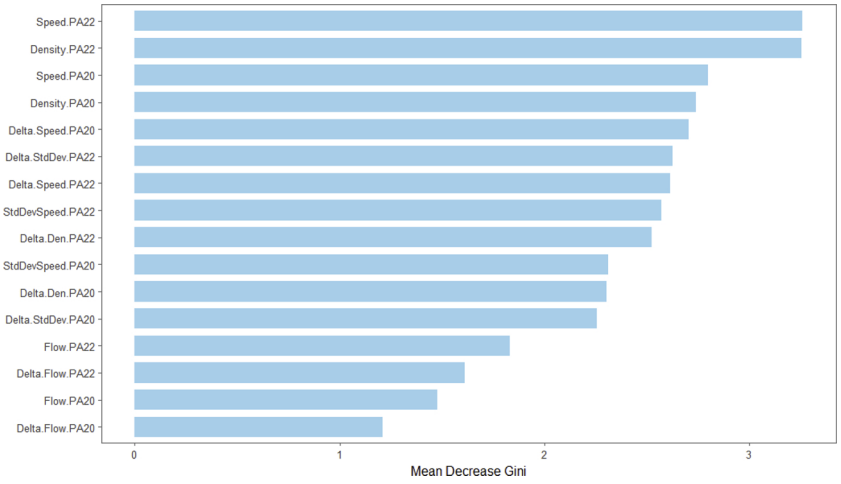

Fig. 3. Variable importance according to Mean Decrease Gini.

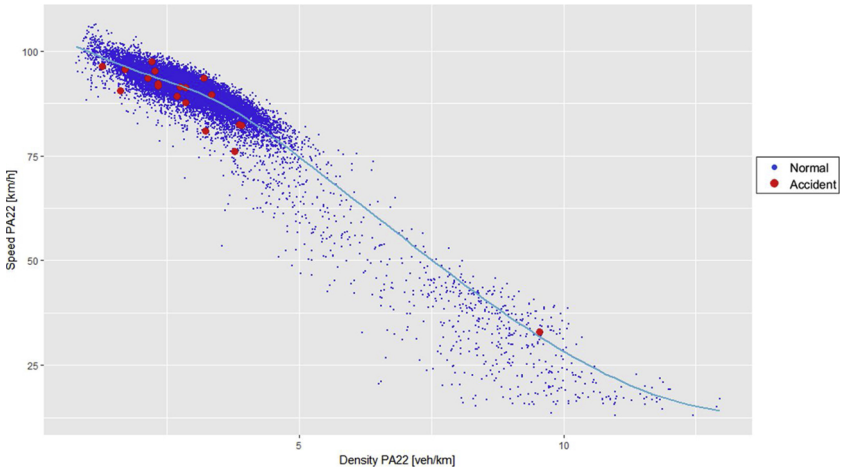

Fig. 4. Speed [km/h] versus density [veh/km] in PA22.

Table 4
Maximum likelihood parameters for the aggregated logistic regression model.

Variable	Estimate	Std. Error	p-value
Speed.PA22	−0.097	0.040	0.015
Density.PA22	−0.975	0.414	0.018
Speed.PA20	−0.019	0.012	0.097

both the aggregated and disaggregated model. This curve shows the prediction power under different FPR, and thus, under different strategic decisions.

Fig. 6 shows the evolution of the mean sensitivity estimate with the increase of 5-fold validation repetitions, when fixing the FPR at around 20 %. It can be seen that these 500 repetitions seem to be enough to obtain a stable sensitivity estimate: the mean sensitivity reaches 49.9 %, which we take as the most likely value of performance for the model with aggregated data. This value is low in comparison to previous studies (see Table 8 in Basso et al., 2018). The conjecture we have is that the model fails to predict better because the stretch study has a high share of other than light-duty vehicles, whose behavior and interactions matter, but are lost when using aggregate data. For example, Basso et al. (2018), who obtain a mean sensitivity of 67 %, use disaggregated data but in a section of the highway where light vehicles account for around 93 % of the traffic, and indeed they find that the most relevant variables are the ones related to light vehicles. The poor performance of the aggregated model is also reflected, quantitatively, in the fact that parameters are significant only at 10 % and, qualitatively, in the fact that it does not seem obvious, from a traffic theory point of view, what is exactly leading to accidents.

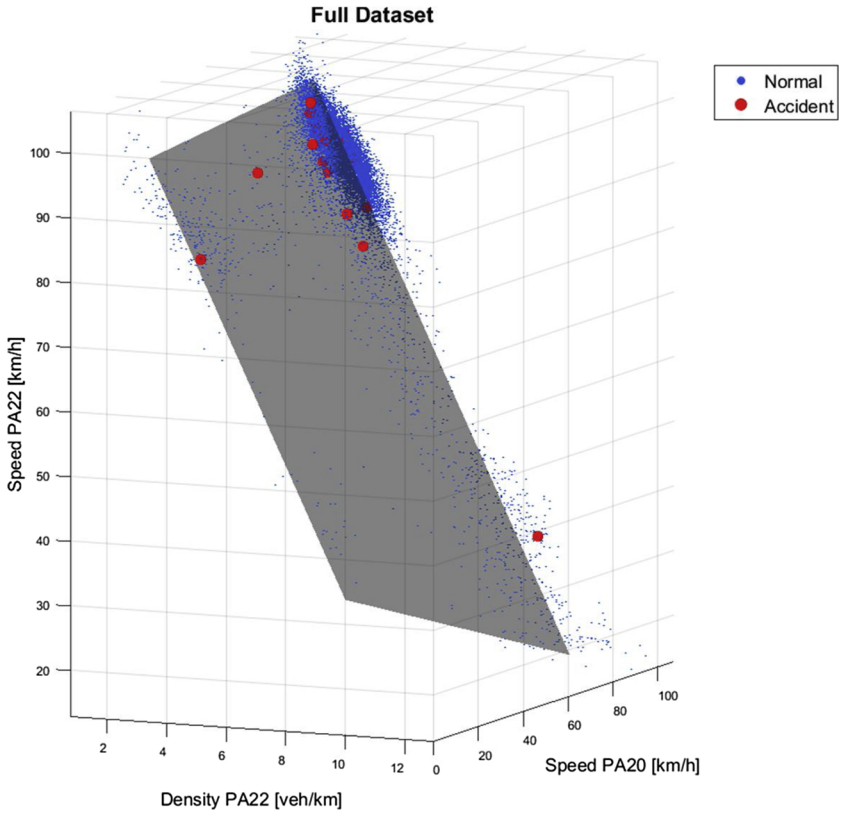

Fig. 5. Decision frontier for the three variables logistic regression for the aggregated data.

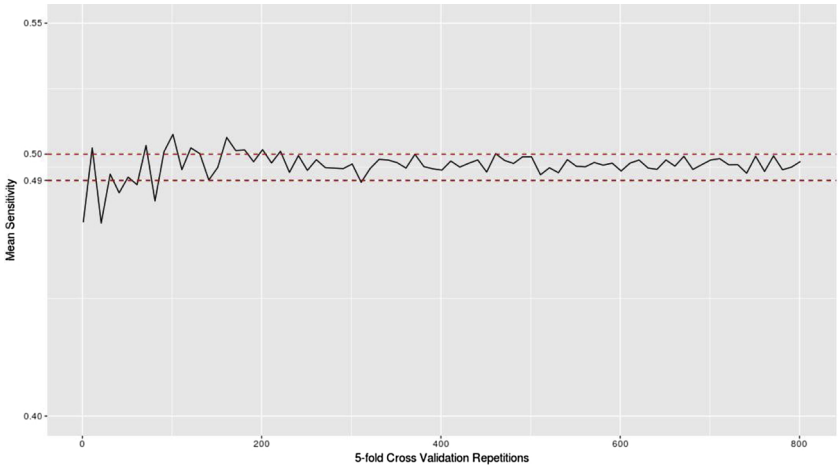

Fig. 6. Mean sensitivity for different number of cross validation repetitions.

4.2. Disaggregated LR model

We now turn to the model that uses disaggregated information, i.e. including the type of vehicle (light, heavy or motorbike) for each data point. We calculate 8 variables for each AVI gate and vehicle type: flow, mean speed, standard deviation of the speed, percentage of such type of vehicle in total flow; and also the change in all of those variables compared to the previous 5-min interval for each vehicle type. The final two variables are average traffic density (over all vehicle types) and its change. The 10 most relevant variables, according to the RF procedure, are presented in Fig. 7, which presents us with the first evidence that recognizing types of vehicles matter: the most important crash predictor is the flow change of heavy-duty vehicles at the AVI gate PA22. Note that, since the flow of light-duty vehicles dominate all others, there was no proxy for this variable in the aggregated model. Moreover, 9 of the 10 most important variables correspond to values associated to motorcycles or heavy vehicles, even though some of them are highly correlated (e.g. heavy vehicles flow change and heavy vehicles composition change). The only variable related to light-duty vehicles that appear is their change in speed at the downstream AVI gate PA20.

We also conducted a graphical analysis to asses to the importance of the variables as accident precursors. We compare the mean value of each variable five minutes before the accident against the mean values registered in non-accident conditions. In Figs. 8 and 9, we show this for the change in flow for heavy-duty vehicles upstream (PA22), the most relevant variable, and for the change in speed of light-duty vehicles downstream, (PA20), the only variable related to light-duty vehicles that the RF procedure selected. What Figs. 8 and 9 unveil is striking: the global minimum of the mean value for both variables is registered exactly five minutes before the accidents.

In addition to these two variables we included in the model a third variable, using again the mean sensitivity over a 5-fold cross validation as the performance indicator. The chosen variable was the speed of heavy vehicles at the upstream gate PA24; note that this gate is before the stretch being analyzed. The estimated parameters of the logistic regression over the full data set are presented in Table 5 while the decision frontier is shown in Fig. 10, with a threshold probability set at $p_0 = 0.207\%$. Any point falling to the left of the plane shown in Fig. 10 will be predicted as a crash under this model.

The first relevant observation is that two of the variables are significantly different from zero at 99 % confidence level, while the third variable reaches 95 %, an improvement over the disaggregated model. The initial reading of the signs of the parameters indicate that the probability of crashes increases when: (i) heavy vehicles drive faster at gate PA24 (ii) their number at gate PA22 diminishes and (iii) downstream, in gate PA20, light vehicles are slowing down. A more involved explanation goes as follows. If upstream, at gate PA24, heavy vehicles pursue high speeds, so do light vehicles. This is because light vehicles are sometimes slowed by heavy vehicles, but the opposite rarely occurs (Johnson and Murray, 2010). Furthermore, heavy vehicles speeds have lower standard deviation than light vehicles speeds (Table 3), so the inclusion of the former in the model might be preferable. Moreover, if at the middle gate PA22, the number of heavy vehicles decreases, light vehicles will increase even more their speed, most likely in a non-

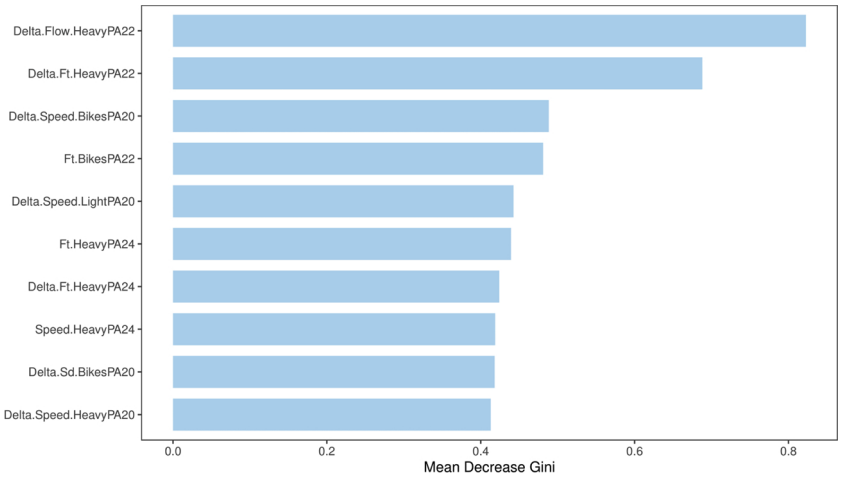

Fig. 7. Variable importance according to Mean Decrease Gini – 10 most important variables.

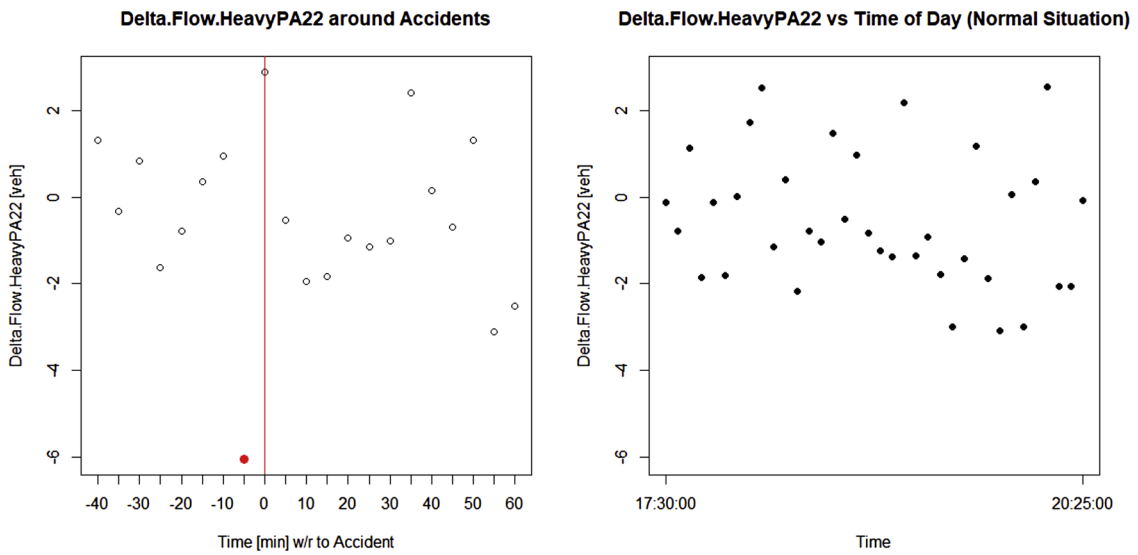

Fig. 8. Mean values of Delta.Flow.HeavyPA22 around the time of accidents.

uniform way. But if this happens in addition to slowing cars downstream, at gate PA20, then the likelihood of accidents –possible rear-end crashes– increases sizably.

To test this possible explanation, we graph the change in speed of light vehicles against the change in flow of heavy vehicles at gate PA22; this is shown in Fig. 11. An evident correlation as conjectured is indeed observed (Pearson's $r = 0.23$), although there is observable variability, indicating that the acceleration of light vehicles is far from uniform after a reduction in the number of heavy vehicles. Indeed, the standard deviation of light vehicle speed at the same gate increases in average when heavy vehicles exit the freeway. These analyses show that, in fact, light vehicles drivers modify their behavior when the traffic composition changes; when less heavy vehicles circulate through the freeway, drivers might decrease their risk perception and thus increase their speed, much like Montella et al. (2008) suggested.

Now, the fact that the correlations are low –as it is evident from Fig. 11– implies that the change in flow of heavy vehicles will be poorly represented by the light-vehicle variables (change in speed and standard deviations of speed), which suggest that there are other effects related to this driving behavior that cannot be captured by the measured variables. One of these effects could be, for example, a reduction in the headway as a consequence of the exit of heavy vehicles, which could increase the risk of rear-end collision.

Table 5
Maximum likelihood parameters for the disaggregated logistic regression model.

Variable	Estimate	Std. Error	p-value
Delta.Flow.HeavyPA22	−0.080	0.028	0.004
Speed.HeavyPA24	0.067	0.032	0.039
Delta.Speed.LightPA20	−0.073	0.026	0.005

Turning to robust validation, we use the same methodology described in the previous section: 500 repetitions of a 5-fold cross validation with a fixed FPR of 20 % in the training set. The distribution of the sensitivities obtained is shown in Fig. 12, while the mean sensitivity over the 500 repetitions is 65.90 %. This value represents our best assessment of the performance of the disaggregated model. It achieves 16 percentage points more of prediction power than the aggregated model, in addition to providing better intuition from a traffic engineering point of view. Thus, traffic composition does improve significantly real-time crash prediction accuracy through the use of separated traffic flows

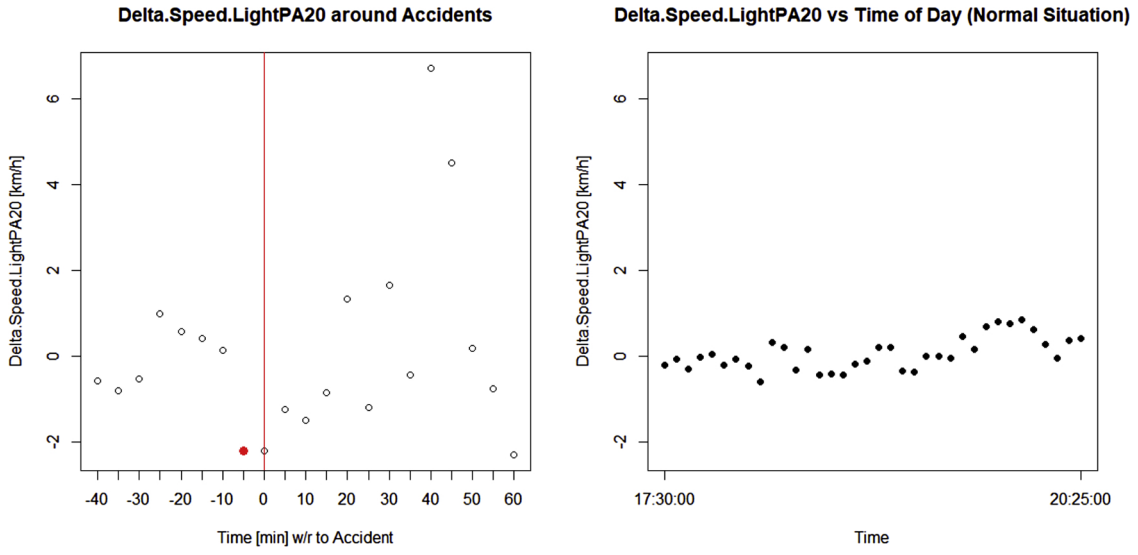

Fig. 9. Mean values of Delta.Speed.LightPA20 around the time of accidents.

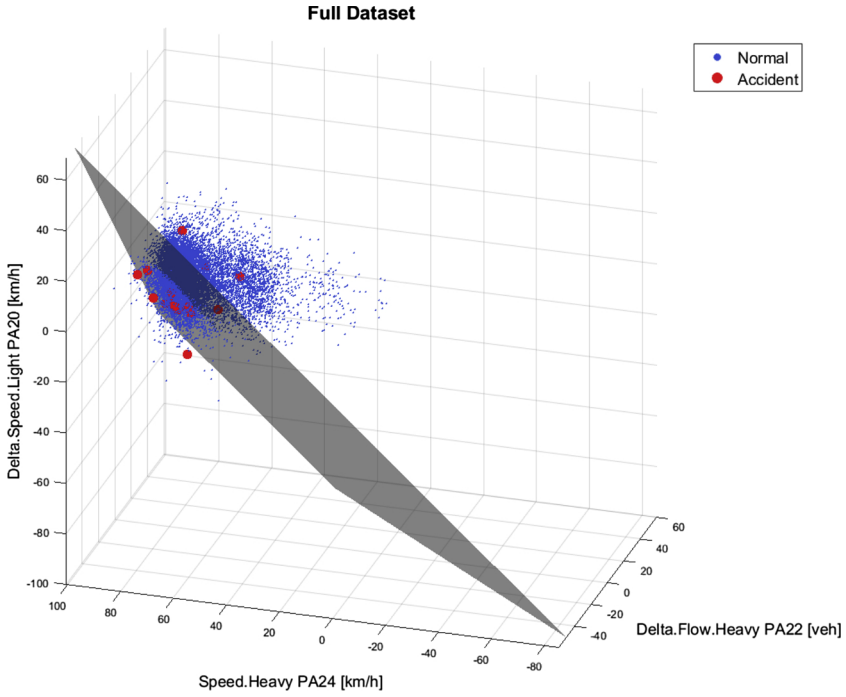

Fig. 10. Decision frontier for the three variables logistic regression for the disaggregated data.

rather than passenger-car equivalent density.

4.3. ROC comparison

As explained in the previous sections, a fixed FPR of around 20 % was used in the training datasets to find the threshold probability of the logistic models. We showed that the use of disaggregated data produces an increase of 16 percentage points in the prediction power of the model. Nevertheless, the choice of such FPR is arbitrary, and there may be scenarios where, for example, only a 10 % of FPR is acceptable. To address the issue of whether the use of disaggregate data is indeed a sizable advantage for other FPR, we present the comparison of the receiver operating characteristic curve, or ROC curve. This curve shows the sensitivity achieved for different values of the maximum FPR allowed: in this case, values from 0 % to 100 %, with increments of 2.5 %. Every point of the curves was obtained with the same methodology proposed, using 500 repetitions of a 5-fold cross-validation.

From Fig. 13, it can be seen that the model that used disaggregated information always dominate the model that only use aggregated data, as expected. Nevertheless, this difference is more important for relatively low values of FPR, reaching a maximum of 30 percentage points in sensitivity when a 15 % of FPR is allowed. Given the nature of the studied phenomenon, it does not seem audacious to argue that low values of FPR are the only ones that matter, because if a tool for real-time crash prediction is to be useful, it should not often deliver false alarms.

5. Support vector machine models

We now turn to the use of SVM as the classification method, instead of Logistic regressions. The goal here is to see if any improvement or degradation in predictor power affects both, disaggregated and aggregated models in the same way, or their relative performance changes. We reuse the variable selection results from the previous section, so we present the best result found using the same three variables for each case: Speed.PA22, Density.PA22 and Speed.PA20 for the case with aggregated data; Delta.Flow.HeavyPA22, Speed.HeavyPA24 and Delta.Speed.LightPA20 for the case with disaggregated data.

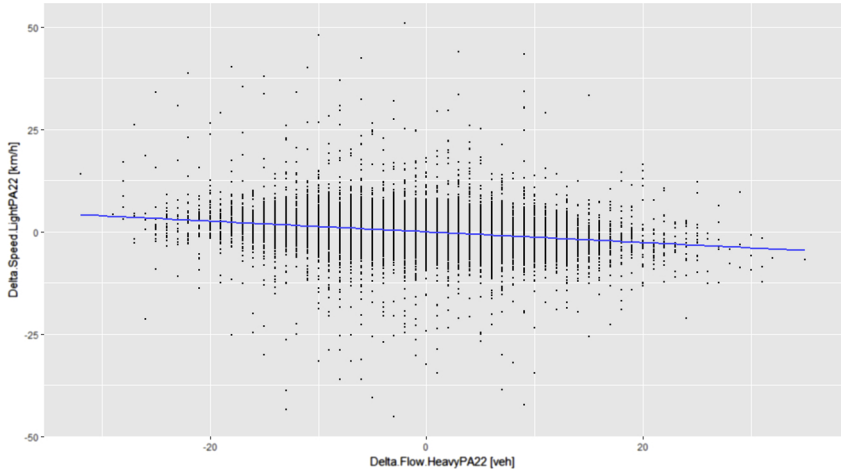

Fig. 11. Delta.Speed.LightPA22 [km/h] versus Delta.Flow.HeavyPA22 [veh].

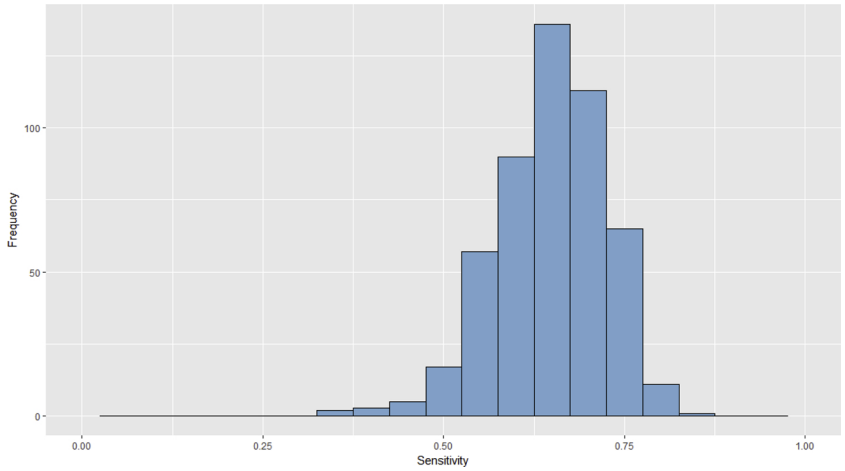

Fig. 12. Distribution of mean sensitivity of 5-fold CV over 500 repetitions – disaggregated model.

5.1. Aggregated SVM model

As explained in Section 3, the optimization problem (P_2) solved to obtain the parameters $\Lambda = (\alpha_i)_{x_i \in S}$ of the decision function f depends on the values of C and γ. To find the values of these parameters that allows us to find the best possibly classifier, we used a grid search. Particularly, we considered values $C = 2^{-15}, 2^{-13}, ..., 2^{15}$ and $\gamma = 2^{-15}, 2^{-13}, ..., 2^{15}$. That is, we used exponentially growing sequences of C and γ, as suggested by Hsu et al. (2004). We performed a 5-fold cross-validation for each combination of values and kernels presented in Section 3.3. Overall, the training data set in each of the five repetition consists of 10,680 rows and 2670 for the validation data-set. The best results found are presented in Table 6.

From Table 6, it can be seen that both types of oversampling techniques achieve a similar result, with a mean sensitivity of around 42 % over the 500 repetitions of a 5-fold cross validation. This means that the SVM model with aggregate data performs even worse than the logistic regression model, as it losses 8 percentage points of sensitivity.

Table 6
Results of SVM model, aggregated data case.

Type of Oversampling	Basic	SMOTE
Oversampling parameters	100 % of majority class	perc.under = 450, perc.over = 750
Kernel	Radial	Radial
C	32	8192
γ	$7.8 \cdot 10^{-3}$	2
Mean Sensitivity (500 repetitions 5-fold CV)	42.6 %	42.1 %
Mean FPR (500 repetitions 5-fold CV)	20.1 %	20.4 %

5.2. Disagreggated SVM model

Just as in the case of the Logistic Regression, the variables used here are the change in flow of heavy vehicles in PA22, the change in speed of light vehicles at gate PA20, and the speed of heavy vehicles at gate PA24. The values of C and γ found using a grid search are presented in Table 7.

Performing 500 repetitions of a 5-fold cross-validation, we obtain a

Table 7
Results of SVM model, disaggregated data case.

Type of Oversampling	Basic	SMOTE
Oversampling parameters	90 % of majority class	perc.under = 200, perc.over = 150
Kernel	Radial	Radial
C	2	0.125
γ	$1.22 \cdot 10^{-4}$	$1.95 \cdot 10^{-3}$
Mean Sensitivity (500 repetitions 5-fold CV)	62.1 %	64.2 %
Mean FPR (500 repetitions 5-fold CV)	20.1 %	21.5 %

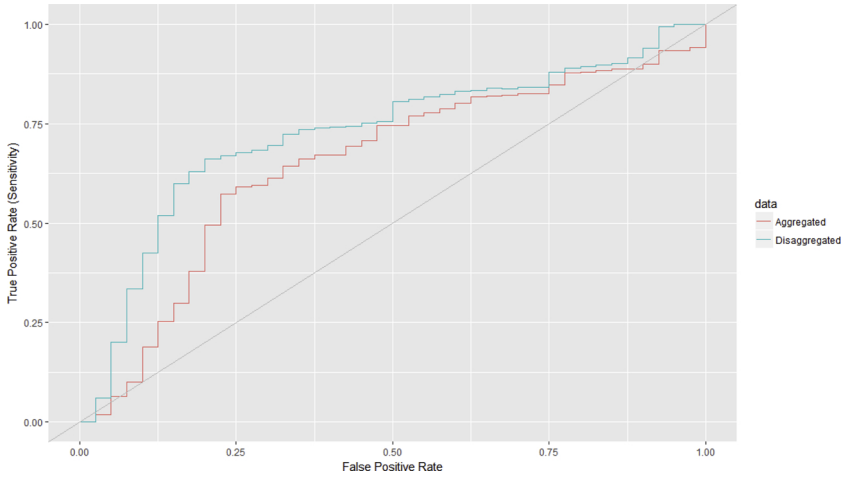

Fig. 13. ROC curves for the aggregated data and disaggregated data models.

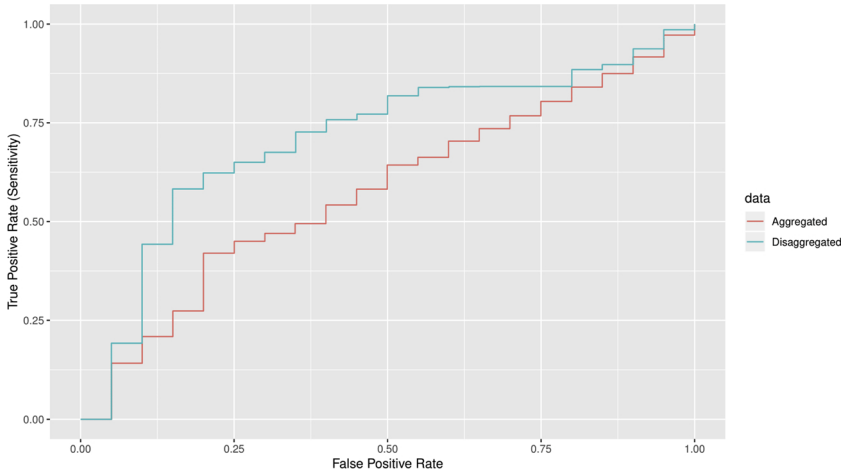

Fig. 14. ROC curves for the aggregated data and disaggregated data models, SVM with basic oversampling.

mean sensitivity of 62.1 % for the model with the basic oversampling technique, and a 64.2 % for the model with SMOTE, although with a slightly higher FPR. Comparing the mean sensitivity of the disaggregated model with the aggregated model, we observe an improvement, this time of 20 percentage points, which seems to confirm that the gain from observing and using disaggregated data does not depend on the classification method.

Note that in the case of SVM models, as discussed by Platt (1999), the result produces an uncalibrated value that is not a probability, and therefore, is not possible to directly set an FPR. Nevertheless, to overcome this issue, we follow Platt (1999) to train a sigmoid function to map the SVM output into probabilities. The results of the ROC curves for both SVM with basic oversampling and SVM with SMOTE are depicted in Figs. 14 and 15. Except for one point, the disaggregated model performed strictly better than the aggregated model for both oversampling techniques confirming the advantage of having access to disaggregated data.

6. Concluding remarks

Our results show that models for crash prediction that use variables separated by vehicle type have prediction power that is sizably larger, while also providing much better intuition about the actual traffic conditions that may lead to accidents; in a nutshell, for the stretch of the highway we analyze, most crashes occur when, following a fall in the number of heavy-duty vehicles, light-duty vehicles accelerate at different paces, inducing a large dispersion of speeds, but encounter slowing traffic down the road. We show that this phenomenon simply cannot be captured when data is aggregated. Learning that much is gained by having access to disaggregated vehicle-type information in crash prediction and, eventually, crash prevention, is important: we believe, as a policy conclusion, that our results should be an important part for assessing technology investments on urban freeways and may be informative on what technology improvements and innovations should be pursued (Chung and Hensher, 2018; Song et al., 2018).

The results of this paper suggest that new road infrastructure should incorporate devices able to determine reliable flow composition data. This last may have a sizeable impact on the Active Traffic Management (Mirshahi et al., 2007) performance that has proliferated last years with the objective of reducing congestion and increasing security. The majority of the highways worldwide are equipped with loop detectors which, usually, are not able to properly distinguish the vehicle types. In this research, AVIs data has been used for determining the impact of having access to disaggregated data. In Santiago, Chile transponders are mandatory for using the urban expressways and, therefore, all vehicles are equipped with this kind of devices; nevertheless, we are aware that Chile is an exception in that way. Moreover, the cost of an AVI gate such as the ones used in Chilean urban highways may easily exceed USD 500,000, something that limits their use. In a more massive setting, new technologies, such as video devices; microwave and laser radar; and passive infrared, ultrasonic, and acoustic sensors should be considered in the near future (Yang et al., 2018a). Nowadays, this is possible due to the technology's advances and cost reduction.

Even though we do not believe that the findings of this paper would

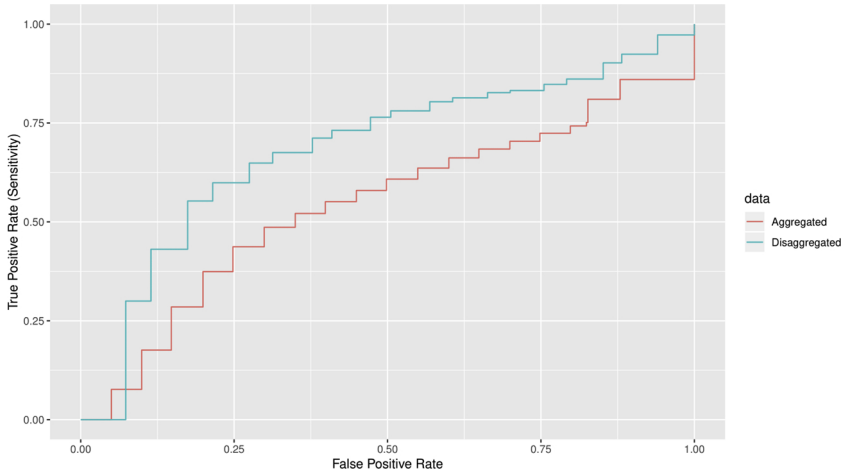

Fig. 15. ROC curves for the aggregated data and disaggregated data models, SVM with SMOTE.

change if other classifications techniques were used instead, we do think that crash-prediction accuracy could be improved if new, more advanced techniques are employed. In particular, there is a recent trend of using deep learning methods in intelligent transportation systems (Nguyen et al., 2018). Nevertheless, its application to crash-prediction is at an early stage. Recent contributions on the subject are not directly applicable to a real-time rare event environment as the one we considered here, because they intend to predict crash frequency in aggregated terms (Dong et al., 2018; Cai et al., 2019), or they do not use the full data set (Theofilatos et al., 2019) or they use spatially aggregated models (Bao et al., 2019). Thus, more research is required, something that we are indeed pursuing.

CRediT authorship contribution statement

Franco Basso: Conceptualization, Methodology, Validation, Formal analysis, Writing - original draft, Writing - review & editing, Project administration, Funding acquisition. **Leonardo J. Basso:** Methodology, Validation, Formal analysis, Writing - original draft, Writing - review & editing, Supervision. **Raul Pezoa:** Methodology, Software, Validation, Formal analysis, Writing - original draft, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowlegments

We thank Autopista Central, and particularly Cindy Carmona, Elena Hurtado, Victor Montenegro and Christian Barrientos for providing us with data and collaborating enthusiastically with the project. We gratefully acknowledge financial support from Fondecyt1191010 and CONICYT PIA/BASAL AFB180003.

Appendix A

Table A1
Traffic flows from entry and exit ramps in the portion studied.

Month	PA24-PA22		PA22-PA20	
	Entries (%)	Exits (%)	Entries (%)	Exits (%)
nov-14	22.61	26.62	25.79	43.4
dec-14	23.93	26.1	26.69	43.29
jan-15	23.49	25.43	26.95	41.75
feb-15	23.59	24.99	27.65	40.77
mar-15	22.93	26.23	25.05	43.52
apr-15	23.12	26.35	25.48	42.92
may-15	23.45	26.22	25.68	42.99
jun-15	23.21	25.75	24.96	43.56
jul-15	23.91	25.24	25.6	43.61
aug-15	23.19	25.95	24.52	44.13
sept-15	23.84	25.68	24.2	44.01
oct-15	23.08	25.67	24.36	43.5
nov-15	22.79	26.03	24.38	43.97
dec-15	24.07	24.53	25.31	43.17
jan-16	22.86	25.19	25.39	42.81
feb-16	23.73	25.11	26.44	41.26
mar-16	22.47	26.71	24.08	42.8
apr-16	21.74	27.81	23.74	43.22

Appendix B. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.aap.2020.105436.

References

Abdel-Aty, M., Pande, A., 2005. Identifying crash propensity using specific traffic speed conditions. J. Saf. Res. 36, 97–108.
Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, F., Hsia, L., 2004. Predicting freeway crashes from loop detector data by matched case-control logistic regression. Transp. Res. Rec.: J. Transp. Res. Board 1897, 88–95.
Abdel-Aty, M., Pande, A., Lee, C., Gayah, V., Santos, C.D., 2007. Crash risk assessment using intelligent transportation systems data and real-time intervention strategies to improve safety on freeways. J. Intell. Transp. Syst. Technol. Plan. Oper. 11 (3), 107–120.
Abdel-Aty, M., Pande, A., Das, A., Knibbe, W., 2008. Assessing safety on Dutch freeways with data from infrastructure-based intelligent transportation systems. Transp. Res. Rec.: J. Transp. Res. Board 2083, 153–161.
Ahmed, M.M., Abdel-Aty, M.A., 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. Bayesian updating approach for real-time safety evaluation with automatic vehicle identification data. Transp. Res. Rec.: 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.
Akbani, R., Kwek, S., Japkowicz, N., 2004. Applying support vector machines to imbalanced datasets. European Conference on Machine Learning. Springer, Berlin Heidelberg, pp. 39–50.
Bao, J., Liu, P., Ukkusuri, S.V., 2019. A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data. Accid. Anal. Prev. 122, 239–254.
Basso, F., Basso, L.J., Bravo, F., Pezoa, R., 2018. Real-time crash prediction in an urban expressway using disaggregated data. Transp. Res. Part C Emerg. Technol. 86, 202–219.
Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.
Breiman, L., Spector, P., 1992. Submodel Selection and Evaluation in Regression. The X-random Case. International statistical review/revue internationale de Statistique, pp. 291–319.
Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth & Brooks, Monterey, CA.
Cai, Q., Abdel-Aty, M., Sun, Y., Lee, J., Yuan, J., 2019. Applying a deep learning approach

for transportation safety planning by using high-resolution transportation and land use data. Transp. Res. Part A Policy Pract. 127, 71–85.
Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357.
Choudhary, P., Imprialou, M., Velaga, N.R., Choudhary, A., 2018. Impacts of speed variations on freeway crashes by severity and vehicle type. Accid. Anal. Prev. 121, 213–222.
Chung, D., Hensher, D.A., 2018. Public private partnerships in the provision of tolled roads: shared value creation, trust and control. Transp. Res. Part A Policy Pract. 118, 341–359.
CONASET, 2018. Estadística De Accidentes Fatales En El Gran Santiago año, 2017. Comisión Nacional de Seguridad de Tránsito, Santiago (in Spanish).
Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297.
Diamantidis, N.A., Karlis, D., Giakoumakis, E.A., 2000. Unsupervised stratification of cross-validation for accuracy estimation. Artif. Intell. 116 (1-2), 1–16.
Dimitriou, L., Stylianou, K., Abdel-Aty, M.A., 2018. Assessing rear-end crash potential in urban locations based on vehicle-by-vehicle interactions, geometric characteristics and operational conditions. Accid. Anal. Prev. 118, 221–235.
Dong, C., Clarke, D.B., Richards, S.H., Huang, B., 2014. Differences in passenger car and large truck involved crash frequencies at urban signalized intersections: an exploratory analysis. Accid. Anal. Prev. 62, 87–94.
Dong, C., Shao, C., Li, J., Xiong, Z., 2018. An improved deep learning model for traffic crash prediction. J. Adv. Transp.
Frez, J., Baloian, N., Pino, J.A., Zurita, G., Basso, F., 2019. Planning of urban public transportation networks in a Smart City. J Univers. Comput Sci 25 (8), 946–966.
Golob, T.F., Recker, W.W., 2003. Relationships among urban freeway accidents, traffic flow, weather, and lighting conditions. J. Transp. Eng. 129 (4), 342–353.
Gu, X., Abdel-Aty, M., Xiang, Q., Cai, Q., Yuan, J., 2019. Utilizing UAV video data for in-depth analysis of drivers' crash risk at interchange merging areas. Accid. Anal. Prev. 123, 159–169.
Hossain, M., Abdel-Aty, M., Quddus, M.A., Muromachi, Y., Sadeek, S.N., 2019. Real-time crash prediction models: state-of-the-art, design pathways and ubiquitous requirements. Accid. Anal. Prev. 124, 66–84.
Hsu, C.W., Chang, C.C., Lin, C.J., 2004. A Practical Guide to Support Vector Classification. Technical Report. Department of Computer Science and Information Engineering, National Taiwan University.
Iragüen, P., Ortúzar, J.D., 2004. Willingness-to-pay for reducing fatal accident risk in urban areas: an Internet-based Web page stated preference survey. Accid. Anal. Prev. 36 (4), 513–524.
Johnson, S., Murray, D., 2010. Empirical analysis of truck and automobile speeds on rural interstates: impact of posted speed limits. Transportation Research Board 89th Annual Meeting (No. 10-0833).
Ki, Y.K., Baik, D.K., 2006. Vehicle-classification algorithm for single-loop detectors using neural networks. IEEE Trans. Veh. Technol. 55 (6), 1704–1711.
Kim, J.H., 2009. Estimating classification error rate: repeated cross-validation, repeated hold-out and bootstrap. Comput. Stat. Data Anal. 53 (11), 3735–3745.
Kohavi, R., 1996. Wrappers for Performance Enhancement and Oblivious Decision Graphs. PhD thesis. Stanford University, Stanford, CA, USA.
Lin, L., Wang, Q., Sadek, A.W., 2015. A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction. Transp. Res. Part C Emerg. Technol. 55, 444–459.
Mani, I., Zhang, I., 2003. kNN approach to unbalanced data distributions: a case study involving information extraction. August. Proceedings of Workshop on Learning from Imbalanced Datasets 126.
Mirshahi, M., Obenberger, J.T., Fuhs, C.A., Howard, C.E., Krammes, R.A., Kuhn, B.T., Mayhew, R.M., Moore, M.A., Sahebjam, K., Stone, C.J., Yung, J.L., 2007. Active Traffic Management: The Next Step in Congestion Management. Federal Highway Administration, United States.
Montella, A., Colantuoni, L., Lamberti, R., 2008. Crash prediction models for rural motorways. Transp. Res. Rec.: J. Transp. Res. Board 2083, 180–189.
Mussone, L., Bassani, M., Masci, P., 2017. Analysis of factors affecting the severity of crashes in urban road intersections. Accid. Anal. Prev. 103, 112–122.
Nguyen, H., Kieu, L.M., Wen, T., Cai, C., 2018. Deep learning methods in transportation domain: a review. Iet Intell. Transp. Syst. 12 (9), 998–1004.
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.
Parsa, A.B., Taghipour, H., Derrible, S., Mohammadian, A.K., 2019. Real-time accident detection: coping with imbalanced data. Accid. Anal. Prev. 129, 202–210.
Peduzzi, P., Concato, J., Kemper, E., Holford, T.R., Feinstein, A.R., 1996. A simulation study of the number of events per variable in logistic regression analysis. J. Clin. Epidemiol. 49 (12), 1373–1379.
Platt, J., 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv. Large Margin Classifiers 10 (3), 61–74.
Rizzi, L.I., Ortúzar, J.D., 2003. Stated preference in the valuation of interurban road safety. Accid. Anal. Prev. 35 (1), 9–22.
SECTRA, 2013. Manual De Evaluación Social De Proyectos De Vialidad Urbana. Secretaría de Planificación de Transporte, Santiago (in Spanish).
Song, J., Zhao, Y., Jin, L., Sun, Y., 2018. Pareto optimization of public-private partnership toll road contracts with government guarantees. Transp. Res. Part A Policy Pract. 117, 158–175.
Stylianou, K., Dimitriou, L., Abdel-Aty, M., 2019. Big data and road safety: a comprehensive review. Mobility Patterns, Big Data and Transport Analytics. Elsevier, pp. 297–343.
Theofilatos, A., 2017. Incorporating real-time traffic and weather data to explore road accident likelihood and severity in urban arterials. J. Safety Res. 61, 9–21.
Theofilatos, A., Yannis, G., Kopelias, P., Papadimitriou, F., 2018. Impact of real-time traffic characteristics on crash occurrence: preliminary results of the case of rare events. Accid. Anal. Prev.
Theofilatos, A., Chen, C., Antoniou, C., 2019. Comparing machine learning and deep learning methods for real-time crash prediction. Transp. Res. Rec 0361198119841571.
Transportation Research Board, 2000. Highway Capacity Manual. National Research Council, Washington, DC.
van Beinum, A., Farah, H., Wegman, F., Hoogendoorn, S., 2018. Driving behaviour at motorway ramps and weaving segments based on empirical trajectory data. Transp. Res. Part C Emerg. Technol. 92, 426–441.
Wang, L., Shi, Q., Abdel-Aty, M., 2015. Predicting crashes on expressway ramps with real-time traffic and weather data. Transp. Res. Rec.: J. Transp. Res. Board 2514, 32–38.
Wang, L., Abdel-Aty, M., Ma, W., Hu, J., Zhong, H., 2019. Quasi-vehicle-trajectory-based real-time safety analysis for expressways. Transp. Res. Part C Emerg. Technol. 103, 30–38.
Weiss, G.M., 2004. Mining with rarity: a unifying framework. Acm Sigkdd Explor. Newsl. 6 (1), 7–19.
Xu, C., Wang, W., Liu, P., 2013. A genetic programming model for real-time crash prediction on freeways. Ieee Trans. Intell. Transp. Syst. 14 (2), 574–586.
Yang, K., Wang, X., Yu, R., 2018a. A Bayesian dynamic updating approach for urban expressway real-time crash risk evaluation. Transp. Res. Part C Emerg. Technol. 96, 192–207.
Yang, K., Yu, R., Wang, X., Quddus, M., Xue, L., 2018b. How to determine an optimal threshold to classify real-time crash-prone traffic conditions? Accid. Anal. Prev. 117, 250–261.
Yu, R., Abdel-Aty, M., 2013. Utilizing Support Vector Machine in Real-time Crash Risk Evaluation.
Yuan, J., Abdel-Aty, M., Wang, L., Lee, J., Yu, R., Wang, X., 2018. Utilizing bluetooth and adaptive signal control data for real-time safety analysis on urban arterials. Transp. Res. Part C Emerg. Technol. 97, 114–127.
Yuan, J., Abdel-Aty, M., Gong, Y., Cai, Q., 2019. Real-time crash risk prediction using long short-term memory recurrent neural network. Transp. Res. Rec 0361198119840611.