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

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.

rate it not included in the paper.

More recently, SVM has been studied as the classification method for prediction. For instance, Lv et al. (2009) used simulated data obtained from the software TSIS to identify traffic conditions which increase the probability of accidents. Yu and Abdel-Aty (2013) used data from I-70 highway in Colorado to measure the risk of crash-accident in real time using SVM. As opposed to our case, the authors only select some of the observations with no accidents in order to avoid unbalanced data and facilitate calibration. We tackle the issue in a different way when calibrating SVM: a Synthetic Minority Over-sampling Technique (SMOTE) was used, discussed in detail in Section 4.

Hossain and Muromachi (2012) used random multinomial logit model to identify the most important predictors and applied a Bayesian belief net procedure to predict crashes. The authors used data collected from Shibuya 3 and Shinjuku 4 expressways under the jurisdiction of Tokyo Metropolitan Expressway Company Limited in Japan. They obtained a mean sensitivity (predictive power) of 66% with a 20% false alarm rate. Recently, based on 551 crashes and corresponding speed information collected on expressways in Shanghai, China, Sun and Sun (2015) calibrate a dynamic Bayesian network with time series. They obtained maximum of 76.4% sensitivity and a false alarm rate of 23.7%.

The rest of the article is organized as follows. In Section 2 the data is described, explaining how it was processed and providing some descriptive statistics. In Section 3 we tackle the variable selection problem, in order to identify what are the strongest precursors of car accidents. In Section 4 we present and calibrate the Support Vector Machine classification model for the initial partition of data (first 80% for calibration, last 20% for validation) while in Section 5 the same is done with the logistic regression model. In Section 6 we test for robustness and compare the performance of models, ours and the ones presented in the literature. Section 7 concludes.

2. Data set and preparation

Autopista Central is an expressway in Santiago, Chile,[3] which is 60.5 km long and has a north-south orientation (see Fig. 1). The raw traffic data set they provided us with has traffic information from November 1st 2014 to April 30th 2016. A data point is the time and speed at which a certain vehicle (fully identified by its transponder) passed an AVI gate using any of the available lanes; in other words traffic per lane cannot be distinguished. Vehicles are classified as light (this include SUVs and smaller commercial vehicles), heavy (including trucks and buses) or motorcycle. On the other hand, Autopista Central also provided us with their accident information. This information is recorded manually: when any incident happens, they track it and store their type (accident, broken down car, roadworks, etc.), date, time and exact location. In this work we are only interested in the accidents.

The highway is divided in Sections for managerial reasons. We decided to focus on the section of the expressway that has the largest accident rate per kilometer per unit of time, namely 3.41 accidents per kilometer per month. We studied the north-south direction of this section which spans for 4.7 km between the Mapocho River and Carlos Valdovinos street, and has six entry ramps and three exit ramps. It has two AVI gates from where traffic information is obtained (see Fig. 2). We consider the afternoon rush hour, that is, Monday to Friday from 5:30p to 8.30p, which left us with 10,745,766 observations, of which 5,298,683 correspond to the AVI gate AC-09 and 5,447,083 to the AVI gate AC-08. This difference is due to existence of two entry ramps between those AVI gates. By choosing a specific section and period, we think we can avoid the influence of, for example geometry or changes in driving behavior, thus helping us to better predict. This, we think, does not decrease the applicability of the overall approach, since a functioning real-time accident prediction tool may have different models running for different times of the day and different sections of the road (Kwak and Kho, 2016).

The raw data was used to calculate 17 variables, averaged over periods of five minutes, for each of the two gates, giving us a total of 34 variables. They are, for each type of vehicle: flow, speed, standard deviation of the speed, density (that is, average flow divided in average speed) and density change, simply calculated as the difference with its value in the previous five minutes. A *composition* variable was also considered, defined as the proportion of each type of vehicle compared with the total flow. The right-hand side variables, for light vehicles and gate, are defined in Table 1

The remaining variables are defined analogously, changing the type of vehicle and/or the AVI gate. The accident data was then used to create the 35th variable, namely, whether there was an accident during the next period of five minutes or not. The variable takes a value of zero if there was no accident and a value of one if there was one. This data set was complemented with –for each period of five minutes– temperature, atmospheric pressure and rainfall. The weather data comes from a station installed by the Department of Geophysics of University of Chile, only 1 km away from the north AVI (AC-09). The final data set has 13,029 observations (5 min periods) of which only 39, i.e. 0.30% had an accident, confirming the rare event feature discussed above. Tables 2 and 3 provide descriptive statistics which enable to show the main features of the traffic conditions of the section of the expressway studied. Fig. 3 shows the evolution of the average speed of the lights vehicles during the studied period.

From Tables 2 and 3 is possible to see that the composition of the traffic in the studied section and period is given by an extremely high percentage of light vehicles, which follows from the fact that heavy vehicles have lower fares in the parallel west section of the freeway (General Velásquez; see Fig. 1). The average participation of heavy vehicles in General Velásquez is around 11% for the same period and the equivalent section, much higher than the 4% showed here.

We also see that the average speed in the AVI gate AC-09 is much lower than the one in the AVI gate AC-08 (for all types of vehicles). This is probably due to the existence of an exit ramp just ahead (less than 200 meters) of the AVI gate AC-09, which

[3] Santiago: Population: 5,822,316; Area: 641 km^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%.

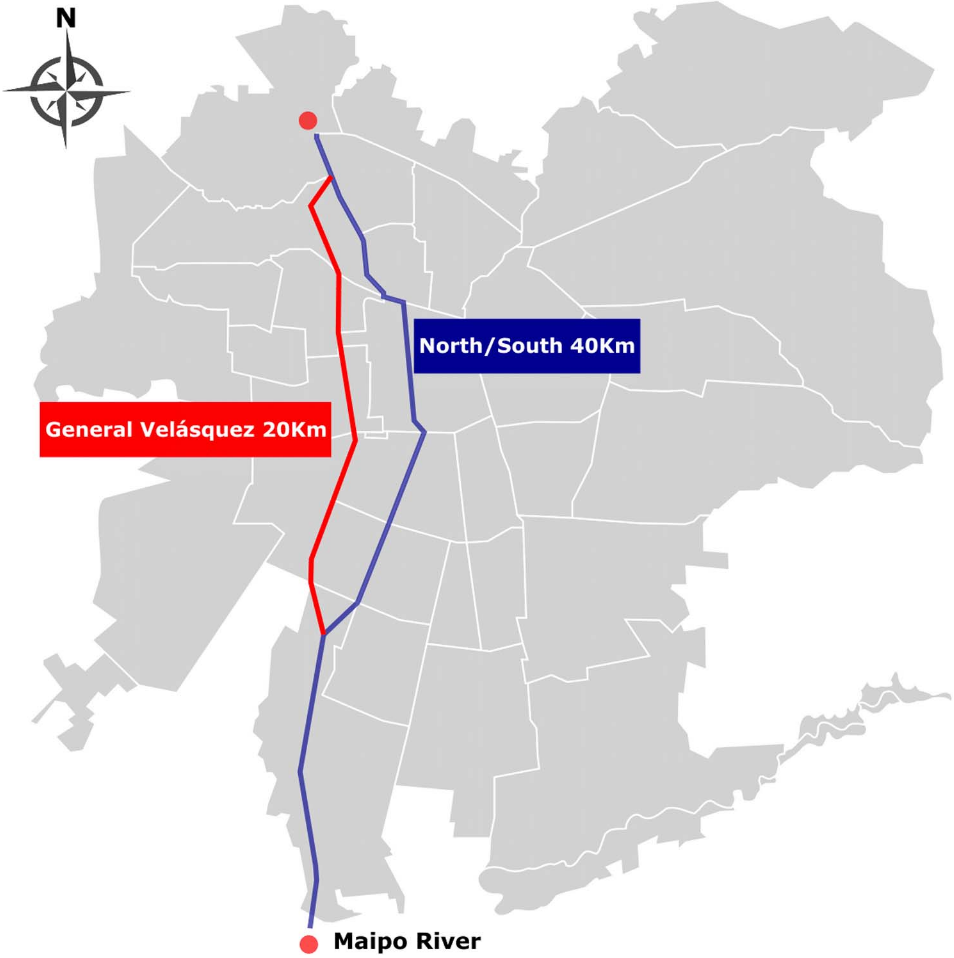


Fig. 1. Autopista Central, Santiago, Chile.

connects nicely with Avenida Libertador Bernardo O'Higgins, the main avenue of Santiago. Thus, near this AVI gate at rush time, one of the lanes is nearly blocked by the vehicles leaving the freeway, leaving only two tracks for the rest of the traffic.

Finally, the average flow of light vehicles is higher in the AVI gate AC-08 than in the AVI gate AC-09, which is explained by the same remark made before: there exists two entry ramps between these AVI gates, and this area, located in the center of the city, is a very congested one in the rush time, thus we have many vehicles entering in this section of the freeway that aim to travel to the suburbs.

In Fig. 4, the distribution of the accidents recorded over the studied period is provided. We can see that the most dangerous time interval is from 7.30p to 8.00p with 11 accidents. That represents 28% of our accident information. We can also note that this sub period coincides with the lowest average speed of light vehicles recorded in AVI gate AC-08, and it has some of the highest average speeds in AVI gate AC-09 (Fig. 3).

3. Variable selection

Having access to a large data set is, undoubtedly, a plus in our goal to predict accidents. But it also brings in the problem of variable selection. Directly including a large number of variables in a classification or regression model may cause over adjustment of the model (Sawalha and Sayed, 2006) which, in turn, may affect both the interpretation of the interrelation between variables and, more importantly, the use of the model in prediction phase. It thus becomes important to analyze our data in order to identify what are the variables that appear to be strong precursors of car accidents. In order to do this, Pearson correlations, Random Forest techniques (also used by Ahmed and Abdel-Aty, 2012) and graphical analyses are used.

The Pearson correlation $\rho_{X,Y}$ was computed for each pair of variables, in order to test for linear dependence; see Fig. 5. We discarded variables with $|\rho_{X,Y}| > 0.95$ to avoid multicollinearity issues. For instance, this procedure removed Den.Light.08 since it is highly correlated to Speed.Light.08. In the same way, the variables Composition.Light.08 and Composition.Light.09 are discarded because they are correlated to Composition.Heavy.08 + Composition.Bike.09 and Composition.Heavy.08 + Composition.Bike.09, respectively.

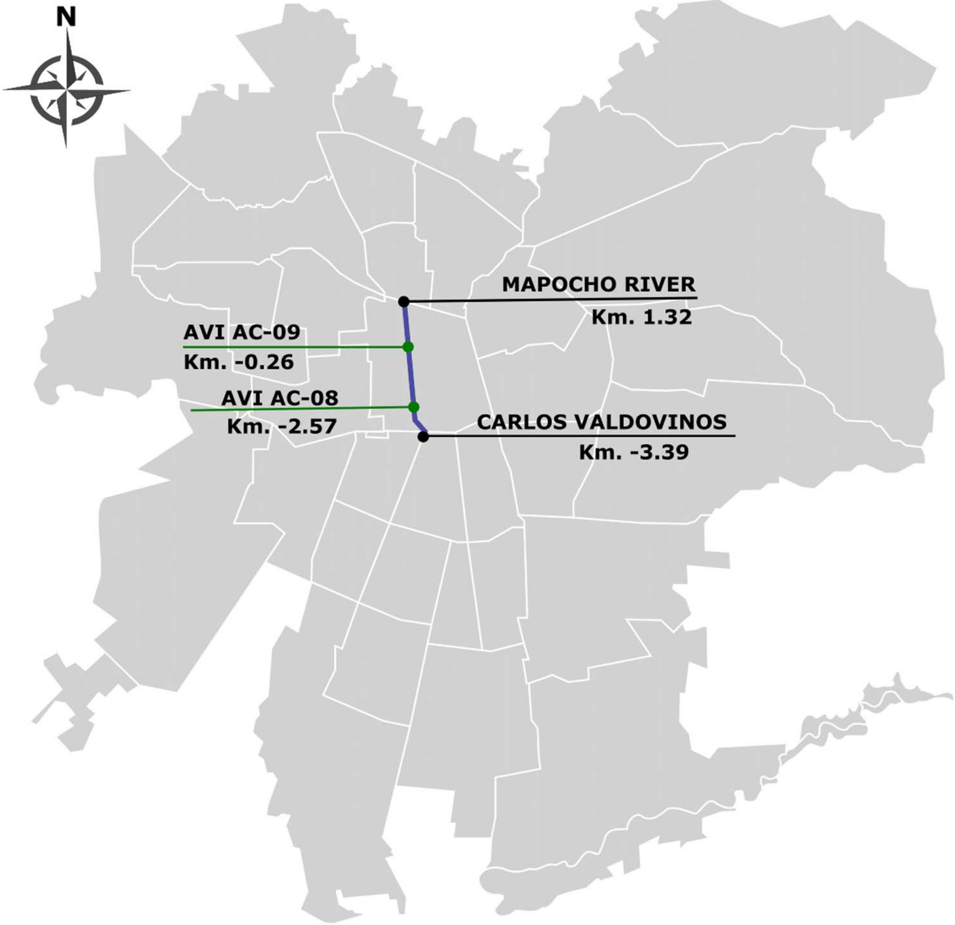


Fig. 2. Section of the expressway studied.

Table 1
Variables used for light vehicles crossing AVI gate 08.

Variable	Definition
Flow.Light.08	Total flow of light vehicles
Speed.Light.08	Mean speed of light vehicles
StdDev.Speed.Light.08	Standard deviation of the speed of light vehicles
Dens.Light.08	Temporal density of light vehicles, defined as Flow.Light/Speed.Light
Composition.Light.08	Percentage of light vehicles in total flow
Delta.Den.Light.08	Change in Dens.Light compared to the previous five minutes
Delta.Speed.Light.08	Change in Speed.Light compared to the previous five minutes

We not turn to the Random forest (RF) procedure. RF was used in our study because it is a widely used method to determine variable importance (Lin et al., 2015). In the crash prediction context, this technique was also applied in similar fashion by Abdel-Aty et al. (2008), Ahmed and Abdel-Aty (2012), Xu et al. (2013).

RF is a machine learning classification method composed by a collection of decision trees. RF classifies an entry in the class which has been assigned most times by the trees (Breiman, 2001). The construction of each tree of the RF is made through two random processes. First, a random sample with replacement of cases is performed, which serves to grow the tree. Second, a sample is selected among all the variables, which is then used to split the nodes (see Fig. 6). The unused data is called out-of-bag (OOB) data. The OOB data could be used to determine an unbiased estimation of classification error.

In this paper, the RF is used to estimate each variable's importance. The importance of a variable in a decision tree is estimated in its ability to reduce an impurity index of nodes when used as a split variable. We use the Gini index as a measure of impurity. For a binary tree (i.e. with two classes as is the case in this study: accident/non accident), the Gini impurity index (Breiman et al., 1984) is defined for node t as:

Table 2
Descriptive statistics of AVI gate AC-08.

AVI AC-08					
		Average	Std. Dev.	Minimum	Maximum
Light	Speed [km/h]	76.6	15.8	8.8	102.4
	Flow [veh]	386.6	50.8	118	527
	% Composition	92.9%	1.9%	83.4%	99.1%
	Density [veh/km]	5.5	2.2	1.4	18.9
Heavy	Speed [km/h]	71.2	14.0	7.9	104.5
	Flow [veh]	18.0	6.1	1	43
	% Composition	4.3%	1.3%	0.3%	10.5%
	Density [veh/km]	0.3	0.1	0.0	1.6
Motorcycle	Speed [km/h]	79.7	14.9	18.8	134.1
	Flow [veh]	11.9	5.0	1	39
	% Composition	2.9%	1.1%	0.2%	10.3%
	Density [veh/km]	0.2	0.1	0.0	0.8

Table 3
Descriptive statistics of AVI gate AC-09.

AVI AC-09					
		Average	Std. Dev.	Minimum	Maximum
Light	Speed [km/h]	51.6	17.9	10.8	95.1
	Flow [veh]	379.0	54.3	6	510
	% Composition	93.4%	1.9%	46.2%	99.7%
	Density [veh/km]	8.2	2.7	0.1	17.1
Heavy	Speed [km/h]	50.9	16.0	7.0	104.2
	Flow [veh]	15.2	5.9	1	42
	% Composition	3.7%	1.3%	0.3%	46.2%
	Density [veh/km]	0.3	0.2	0.0	1.0
Motorcycle	Speed [km/h]	58.1	16.0	13.9	117.4
	Flow [veh]	11.8	5.1	1	42
	% Composition	2.9%	1.5%	0.2%	10.6%
	Density [veh/km]	0.2	0.1	0.0	0.9

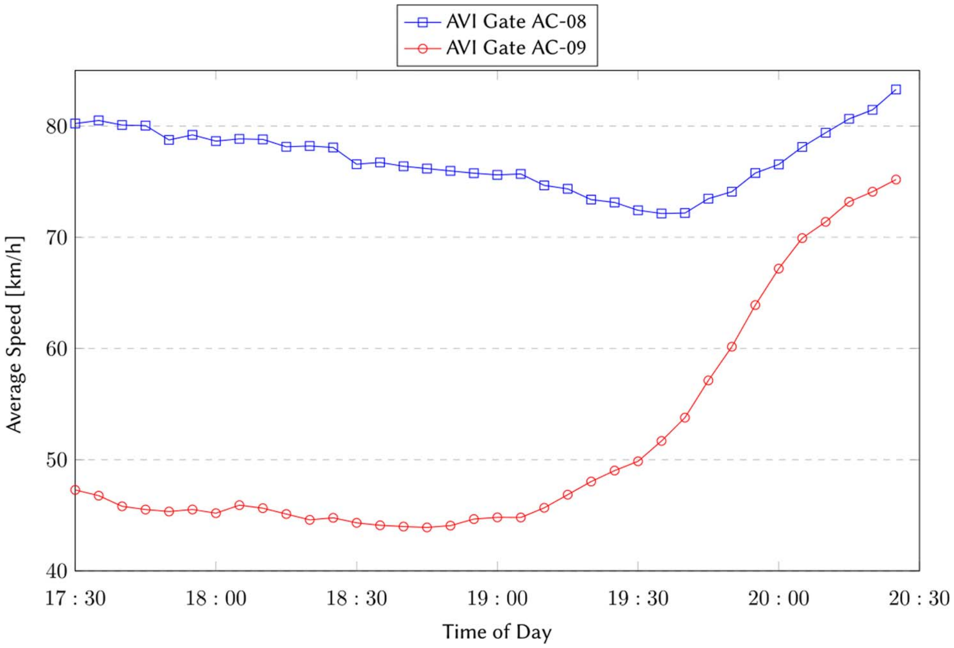


Fig. 3. Evolutions of average speed of light vehicles in the studied period.

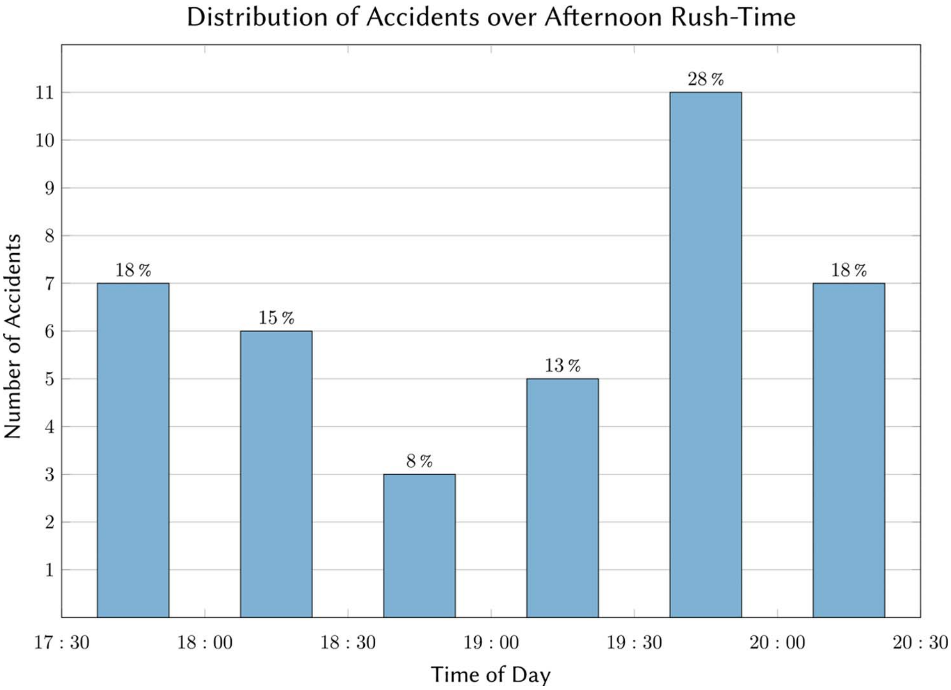


Fig. 4. Distribution of accidents over afternoon rush time.

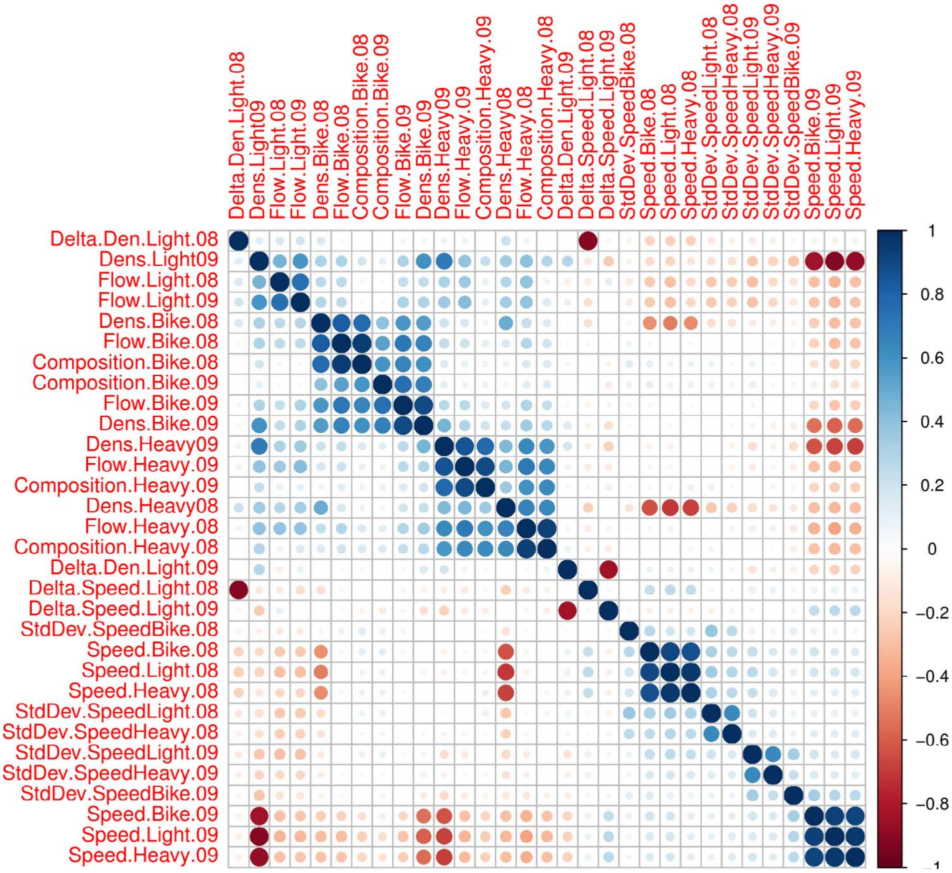
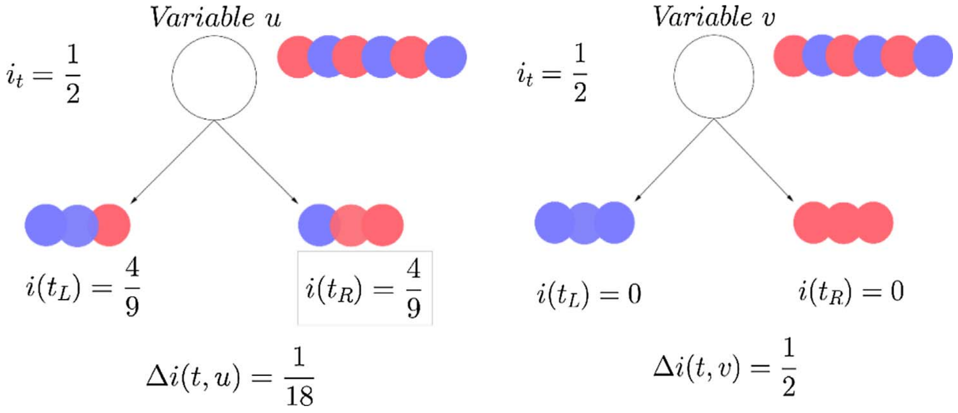


Fig. 5. Correlation matrix.



(a) Case 1, Variable u. (b) Case 2, Variable v.

Fig. 6. Example of two splittings, with case 2 (variable v) preferred.

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

where $p(i/t)$ is the probability of case in class i given node t.

Then, after splitting the tree using the variable u, the decrease in impurity is defined as:

$$\Delta i(t,u) = i(t) - \frac{N_L}{N} i(t_L) - \frac{N_R}{N} i(t_R)$$

where N_L and N_R are the number of observations falling into the left and right children of the split, respectively, while $N = N_L + N_R$ is the total number of observations. $i(t_L)$ and $i(t_R)$ are the Gini's impurity index for the left and right children.

Thus, the larger the value of $\Delta i(t,u)$, the more important variable u is. Our RF has 500 trees which include 4 variables randomly chosen. Then, the average decrease in impurity is computed for those 500 trees for each variable. The results are shown in Fig. 7. We observe that the most important variables are related mainly to the light class, something reasonable given the high proportion of the light class compared to others ones (see Tables 2 and 3).

Finally, a graphical analysis was conducted to determine accidents precursors. To do so, we compare the behavior of the mean of each variable around the time of the accident. Our most important findings are (see Figs. 8 and 9) that the global minimum for the

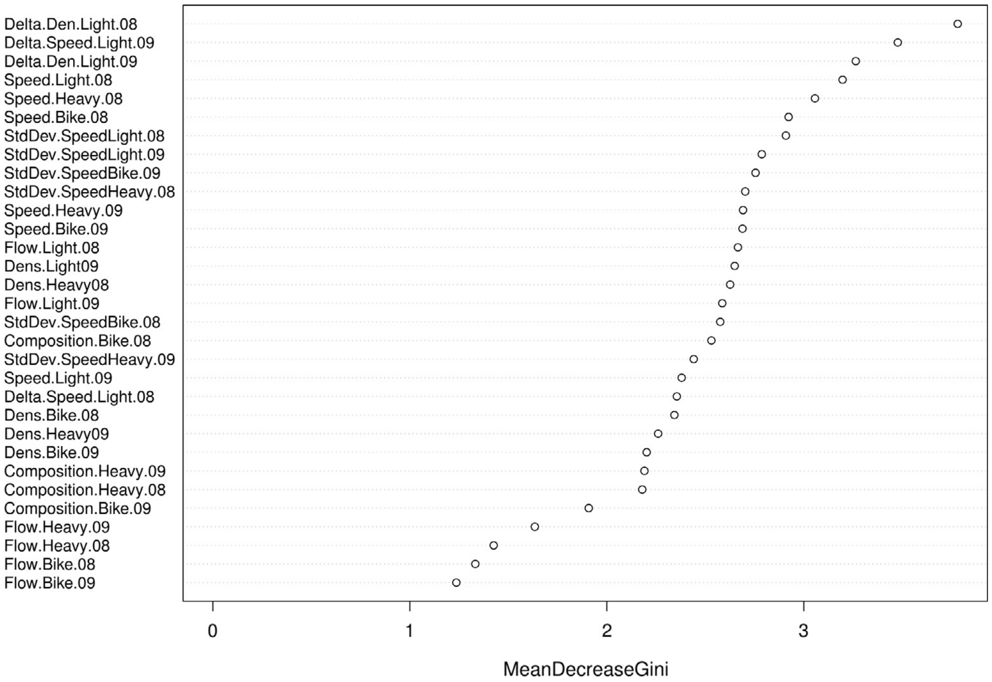


Fig. 7. Change in Gini impurity index to determine variable importance.

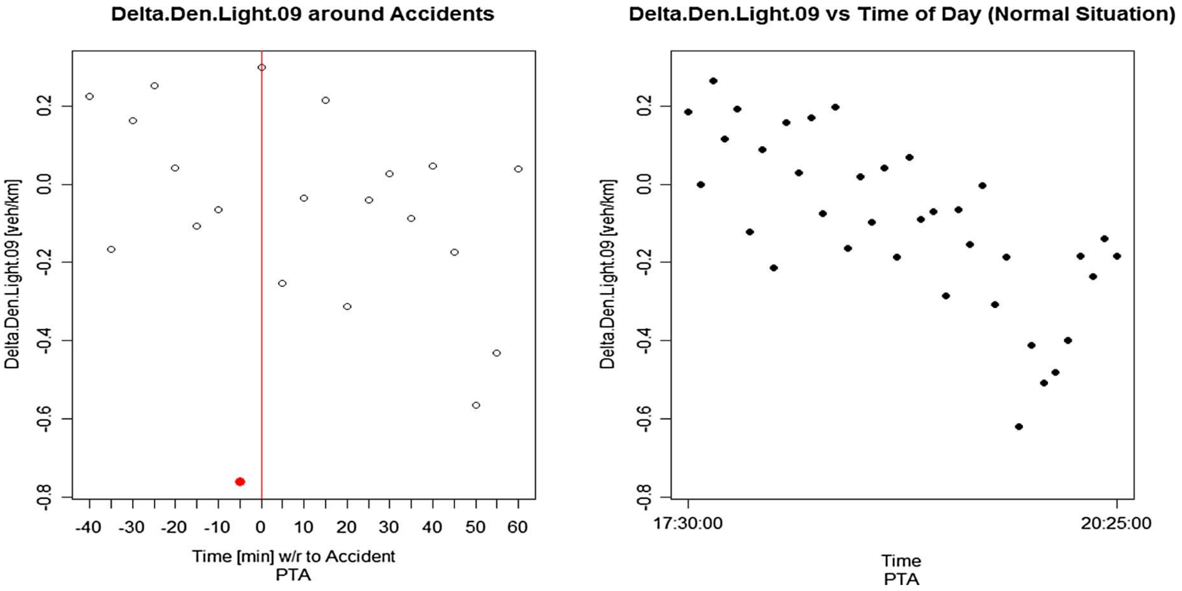


Fig. 8. Delta.Den.Light.09 behavior prior and after an accident.

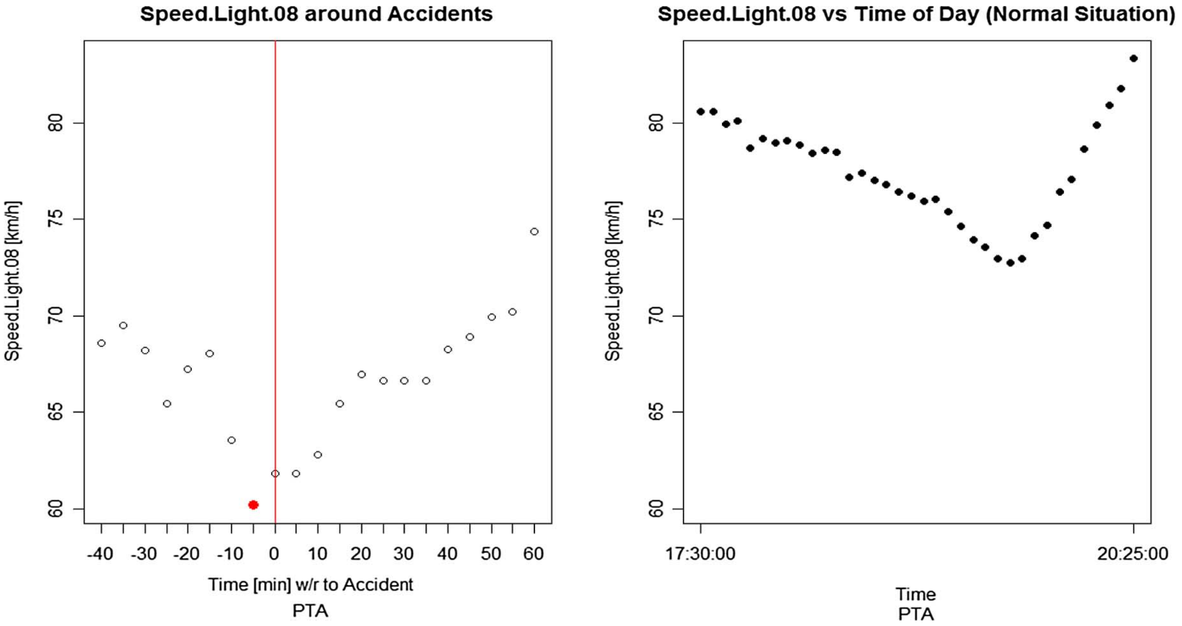


Fig. 9. Speed.Light.08 behavior prior and after an accident.

variables Delta.Den.Light.09 and Speed.Light.08 are attained just 5 min before the accident. As these two variables are in the top 5 of Gini's index, we conjectured that these variables would be highly relevant in the classification models and became our starting point when calibrating.

A final point to make is that the RF procedure help to assess the importance of one variable at the time. It does not help, however, to assess whether non-linear combinations of variables help to further separate 0 s from 1 s or not. Non-linear specifications of the classification models are tested below.

4. Classification method: support vector machines

Support Vector Machine (SVM) could be used to solve binary classification problem. We first provide some theoretical background on how SVM works, then use it on our data set. SVM seek to find a separator hyperplane $f(x) = wx + b$ between two classes in order to maximize the distance between the classes and the decision frontier. Given linear separable data

$(x_1,y_1),...,(x_n,y_n)y_i \in \{-1,1\},x_i \in \mathbb{R}^d,i = 1,...,n$, one and only one of the following statement holds:

$$\begin{aligned} x_i \cdot w + b &\geqslant 1 \quad \text{for } y_i = 1 \\ x_i \cdot w + b &\leqslant -1 \quad \text{for } y_i = -1 \end{aligned}$$

We can combine the above statements into only one as follows:

$$y_i(x_i \cdot w + b) \geqslant 1 \;\; \forall\, i$$

It is possible to prove (Cortes and Vapnik, 1995) that the w vector which maximizes the margin must be the solution of the following non-linear optimization problem:

$$\min \frac{1}{2}\|w\|^2 \text{ s. t. } \; y_i(x_i \cdot w + b) \geqslant 1,\; i = 1,...,n$$

If data is not linear separable, it is possible to add slack variables ξ_i to penalize misclassification. Cortes and Vapnik (1995) proposed the following SVM optimization problem:

$$\begin{aligned} &\min \tfrac{1}{2}\|w\|^2 + C\sum_i \xi_i \\ &\text{s.t } y_i(x_i \cdot w + b) \geqslant 1 - \xi_i,\; i = 1,...,n \\ &\quad \xi_i \geqslant 0, i = 1,...,n \end{aligned}$$

If we introduce the KKT multipliers, the SVM optimization problem[4] can be stated as follows:

$$\begin{aligned} &\max_\alpha \sum_i \alpha_i - \tfrac{1}{2}\sum_i \sum_j \alpha_i\alpha_j y_i y_j x_i \cdot x_j \\ &\quad \text{s. t}\sum_i \alpha_i y_i = 0 \\ &\quad 0 \leqslant \alpha_i \leqslant C \;\; \forall\, i \end{aligned}$$

If the decision function is nonlinear, it is possible to map the data to another Euclidean space H through a function Φ. Note that in the dual formulation, the data appears only as product $x_i \cdot x_j$. The mapping to the Euclidean space H could be done by computing the kernel function K which represents the dot product in H: $K(x_i,x_j) = \Phi(x_i)\cdot\Phi(x_j)$ (Hastie et al., 2001). In this paper we used the following classical kernels:

- Radial Kernel: $K(x_i,x_j) = \exp(-\gamma\|x_i - x_j\|^2)$
- Polinomial Kernel: $K(x_i,x_j) = (\gamma x_i \cdot x_j + 1)^q$,with $q = 3$
- Sigmoid Kernel: $K(x_i,x_j) = \tanh(\gamma x_i \cdot x_j + 1)$

As discussed above, and as can be seen from the actual data, accidents are rare events. This implies that SVM has to calculate the best separating hyperplane with a large number of observations in one class, and a very small number of observations on the other. This has proved to be troublesome for SVM as it ends up providing poor predictions (Akbani et al., 2004). To overcome this problem in the calibration phase, we use the Synthetic Minority Over-sampling Technique (SMOTE). This technique introduced by Chawla et al. (2002) sub-samples the majority class and over-samples the minority class (see Fig. 10). To do the latter, synthetic examples of the minority class are created. These examples are randomly introduced among the minority class and some of their closest k-neighbors. Some adjustment could be done in order to tune the proportions of both classes. It is very important to point out this artificially more balanced data is only used to calibrate the model; to test the predictive performance the full, original, unbalanced data is used. Previous articles have shown the benefits of using the SMOTE-SVM combined procedure which usually outperforms other classification methods for unbalanced data set (Drosou et al., 2014; Fergani, 2016).

Following to the previous section, we tried many different specifications using from the most relevant to less relevant variables, according to the RF analysis. As explained, in this section only the first 80% of the data set is used for calibration purposes, leaving the remaining 20% for validation. This ensures that we validate on data that was not used for training or calibrating the model, thus really placing stress on the model capabilities to predict. The 20% corresponds to an average of 7.8 accidents per validation sample. We think this is a number large enough to test the quality of our model. We also tried with 70% vs 30% cross validation ratios, obtaining very similar results.

The calibration data was adjusted through the SMOTE technique, varying some key parameters (the kernel type and the gamma parameter). Regarding variables, we ended up finding that the best specifications had Speed.Light.08 and Delta.Den.Light.09 as main variables. A specification including, additionally, Speed.Light.09 worked well also. In Table 4 the best models are shown, using the predictive performance on the full training data (i.e. not adjusted with SMOTE), as a model adjustment metric.

Arguably, the best model is the one that used a radial kernel for the SMOTE procedure. Calibrating without adjusting the data set with SMOTE proved to deliver much worse results. The poor performance of SVM for unbalanced datasets has been, as discussed before, documented in the literature (Akbani et al., 2004). Fig. 11 shows that the decision frontier is almost a straight line, something

[4] The Lagrangian dual of the previous optimization problem.

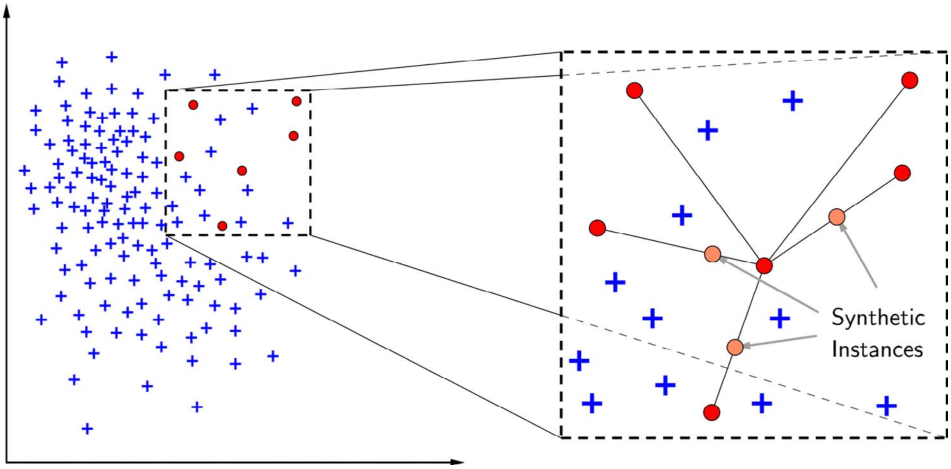


Fig. 10. Oversampling using SMOTE.

Table 4
Results of SVM models.

Kernel	Radial	Sigmoid	Polynomial (Grade 3)
Gamma	0.001	0.001	1
Cost	10	100	1
SMOTE.perc.over	500	500	500
SMOTE.perc.under	100	100	100
Variables	Speed.Light.08 Delta.Den.Light.09	Speed.Light.08 Delta.Den.Light.09	Speed.Light.08 Speed.Light.09 Delta.Den.Light.09
Sensitivity (%)	100	100	87.50
False Positives Rate (%)	20.17	28.56	20.86

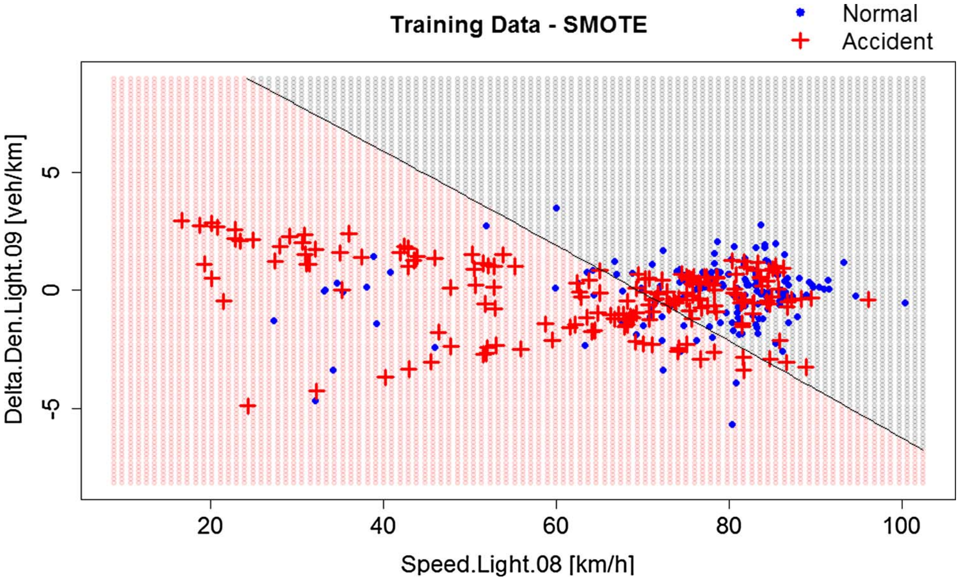


Fig. 11. Decision frontier for SVM with radial kernel for the training SMOTE data-set.

that also happens for the sigmoid and polynomial kernels.

In Fig. 12 we show the SVM decision frontier and the (unbalanced) validation data set: the last 20% of the data we were provided with. It clearly shows why the sensitivity rate was 100%: the frontier perfectly separates the accident events. Yet, as we have discussed before, the sensitivity rate may be ‘too good’: the value may heavily depend on the specific partition. To see this in Fig. 13 we show the SVM decision frontier and the full data set (100%). What this shows is that 8 accidents are above the frontier and would

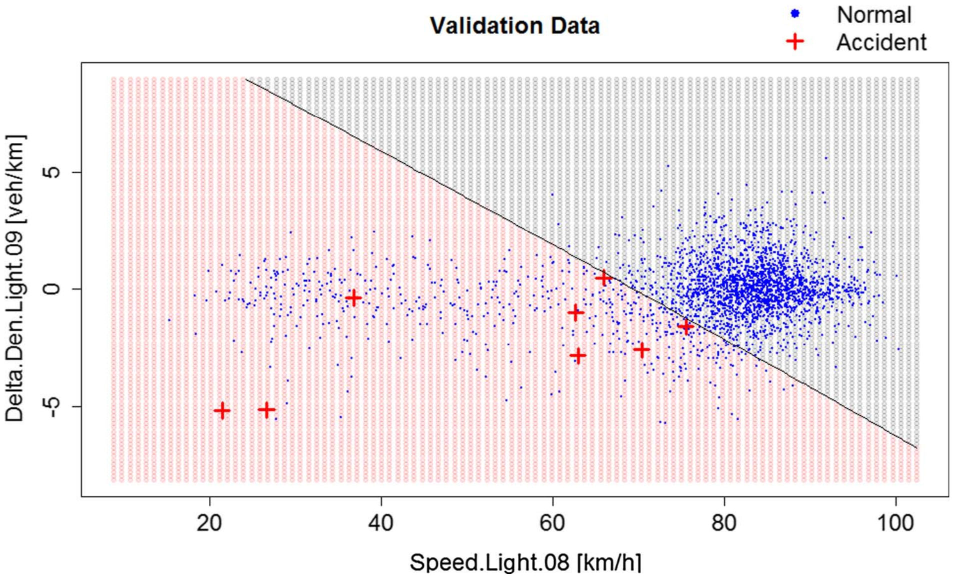


Fig. 12. Decision frontier for SVM with radial kernel for the full validation data-set.

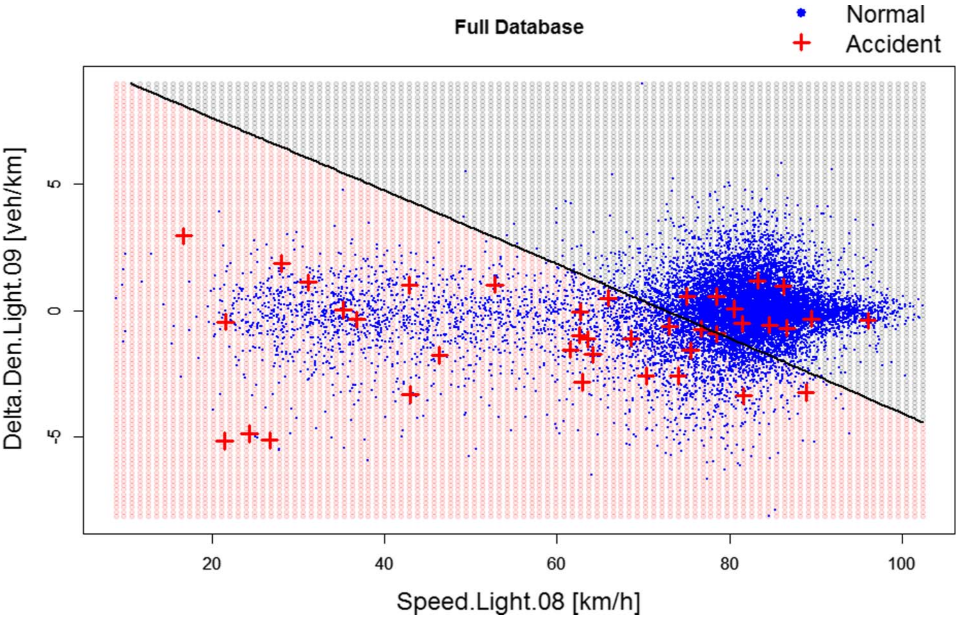


Fig. 13. Decision frontier for SVM with radial kernel over the full data-set (100%).

not be predicted. It just happened that none of those accident took place in the last 3 and half months of the data we received, thus not making it to the validation data set, and enabling a 100% sensitivity. The partition used was a good draw. It is also clear that if a different partition of the data was used the sensitivity would not be as high; in fact, in a really bad draw, the 10 points above the frontier would be all in the validation data set and the result would be very poor. This discussion shows the importance of repeating the calibration/validation process over many random partitions of the data, something we do in Section 6, and indicates that sensitivity values coming from calibration/validation on just one partition of the data should be taken with care.

5. Classification method: logistic regression

We now explore a second classification method: the logistic regression. The previous section showed that the SVM classifier frontier was quite similar to a line. Thus, the upside of using a logistic regression model is that one obtains parameters that may be easier to interpret opposed to SVM which is more of a black-box which requires multiple rules extraction (Martens et al., 2007).

The generalized linear models (of which the logistic regression is a particular case) aim to relax the restrictions given by the classical linear model

$$y = \beta_0 + \boldsymbol{\beta}^T\mathbf{x} + \varepsilon$$

which has to satisfy the Gauss-Markov assumptions in order to have the BLUE (best linear unbiased estimator) property for the ordinary least squares estimators. In particular, the error must be normally distributed with zero mean and has to satisfy the homoscedasticity property: $\mathrm{Var}(\varepsilon_i) = \sigma^2 < \infty$, i.e. constant variance.

As Hastie and Pregibon (1992) remarks, in some situations this is not appropriate. Generalized linear models deal with these problems by introducing a reparametrization to induce linearity and by allowing a non-constant variance (homoscedasticity violation) of the error. Specifically, GLM require:

- Link function, which describes how the mean depends on linear predictors $g(\mu) = \beta_0 + \boldsymbol{\beta}^T\mathbf{x}$.
- Variance function that captures how the variance of y depends upon the mean, $\mathrm{Var}(y) = \phi V(\mu)$, with ϕ constant.

In our case, we consider $y = 1$ if an accident occurs in the next five minutes, and $y = 0$ in other case, so $y \sim \mathrm{Bernoulli}$ (p). As link function between p and the independent variables $\mathbf{x}$, we use the logit link function g given by:

$$g(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \boldsymbol{\beta}^T\mathbf{x}$$

This implies:

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

The parameters $\boldsymbol{\beta}$ and β_0 are then estimated by maximum likelihood. As in the SVM case, we tried many different specifications using from the most relevant to less relevant variables, starting with linear specifications. As in SVM, the best results were obtained when the vector $\mathbf{x}$ contained Speed.Light.08 and Delta.Den.Light.09. The results of the logistic regression are shown in Table 5. The significance tests indicate that the three parameters are different from 0 at 90% confidence level.

To use this model in prediction phase we need to define a value p_0 such that, when the model delivers a value of $p > p_0$, those traffic conditions are classified as leading to an accident in the next five minutes. In order to define p_0 we calculate the values of p for all observations in the estimation data set and choose p_0 so that the false alarm rate stands at about 20%. This lead to $p_0 = 0.299\%$, a value that may seem low to the reader yet, it leads to satisfactory results in terms of sensitivity, while keeping the false alarm rate at 22% over the training data-set. The low value is explained by the extreme unbalanced characteristic of the data set.

The decision frontier –a straight line indeed–is shown in Fig. 14. Changing the threshold p_0 moves the decision frontier in parallel fashion. The tradeoff is as follows: If the straight line goes up, both the sensitivity and the false alarms increase. If the straight line goes down, both the sensitivity and the false alarms decrease. We attempted to estimate the Logistic regression model by adjusting the estimation data set using SMOTE, yet the results end up being worse, an interesting fact that shows that balancing data sets, through SMOTE or match-control may not always be the best course of action.

We now turn to the validation set, the last 20% of the data. The logistic model delivers, as in the SVM case a sensitivity of 100%, with a false alarm rate of 21.29%. The decision frontier and the validation data set are shown Fig. 15.

Many observations can be made: first, the decision frontier is similar, yet not identical to the SVM one. Second, the 100% sensitivity is explained by the same reasons as in SVM: the partition that uses the first 14.4 months for estimation and the remaining 3.6 months for validation is, by chance, a very good draw. Third, while SVM and the logistic model are quite similar, the latter has the advantage of providing us with an explicit function for the decision frontier, thus being easier to interpret.

In fact, we can now provide an interpretation of the actual process by which most accidents happen. As Figs. 8 and 9, the parameters of the Logistic model, and Figs. 14 and 15 show, accidents occur when, simultaneously, there is a dramatic, absolutely abnormal drop in density at AVI gate AC09 (upstream) while, downstream (AVI gate AC08), speeds are abnormally low. This means that, upstream, there were atypical congestion conditions that start to ease, leading to vehicles probably speeding more than usual in order to catch up. Those vehicles however, will ran into an abnormally low speed zone downstream, slower than what drivers are used. In a nutshell, the perfect storm occurs when vehicles that were trapped in heavier than usual congestion upstream, race to recover time lost but ran into an unexpected, atypically low speed zone downstream.

The interpretation of the process above hints us that not only the dramatic drop in density at the upstream AVI gate AC09 may be of interest but also the speed. Including that variable linearly, however did not improved the model so we attempted non-linear specifications. After several attempts we obtained better estimation using the processed variables Speed.Light.08^2 and

Table 5
Maximum likelihood parameters for the two variables logistic regression model.

Variable	Estimate	St. deviation	p-value
Intercept	−3.378	0.555	$1.15 \cdot 10^{-9}$
Speed.Light.08	−0.035	0.008	$1.24 \cdot 10^{-5}$
Delta.Den.Light.09	−0.214	0.119	0.073

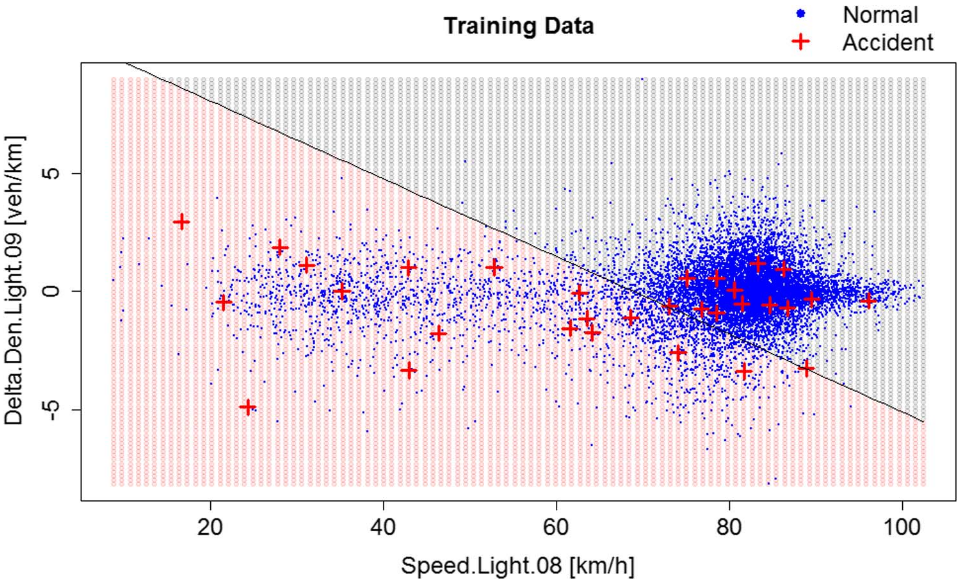


Fig. 14. Decision frontier for the two variables logistic regression for the training data-set.

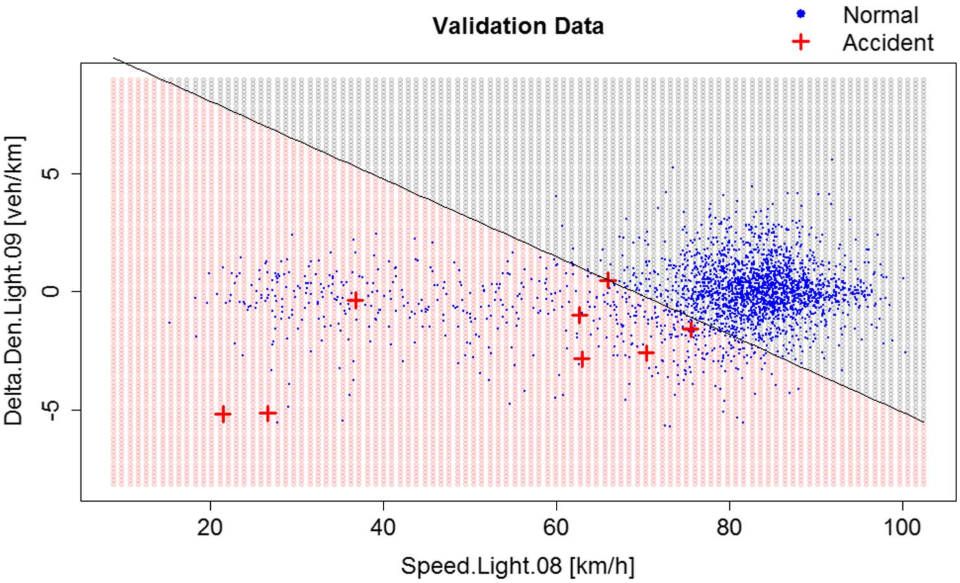


Fig. 15. Decision frontier for the two variables logistic regression over the validation data-set.

Delta.Den.Light.09 · Speed.Light.09^2. The estimated parameters for this model are presented in Table 6.

From this, it can be seen that both variables (which are a function of three original variables) are significant at the 95% confidence level. The signs of the variables are intuitively correct: high speeds at the downstream gate decrease the likelihood of accidents, and at a more than a linear rate. On the other hand, negative Delta.Den.Light.09, that is, drops in density increase the likelihood of an accident (as found before) but, now, the effect is amplified by the square of the speed. In summary, the situation that causes the highest probability of accidents is: (i) substantially density drops upstream with ensuing high speeds (ii) unusual low speeds at gate

Table 6
Maximum likelihood parameters for the three variables logistic regression model.

Variable	Estimate	St. deviation	p-value
Intercept	-4.287	0.393	$< 2 \cdot 10^{-16}$
Speed.Light.08^2	$-2.99 \cdot 10^{-4}$	$7.26 \cdot 10^{-5}$	$3.89 \cdot 10^{-5}$
Delta.Den.Light.09 · Speed.Light.09^2	$-5.58 \cdot 10^{-5}$	$2.60 \cdot 10^{-5}$	0.032

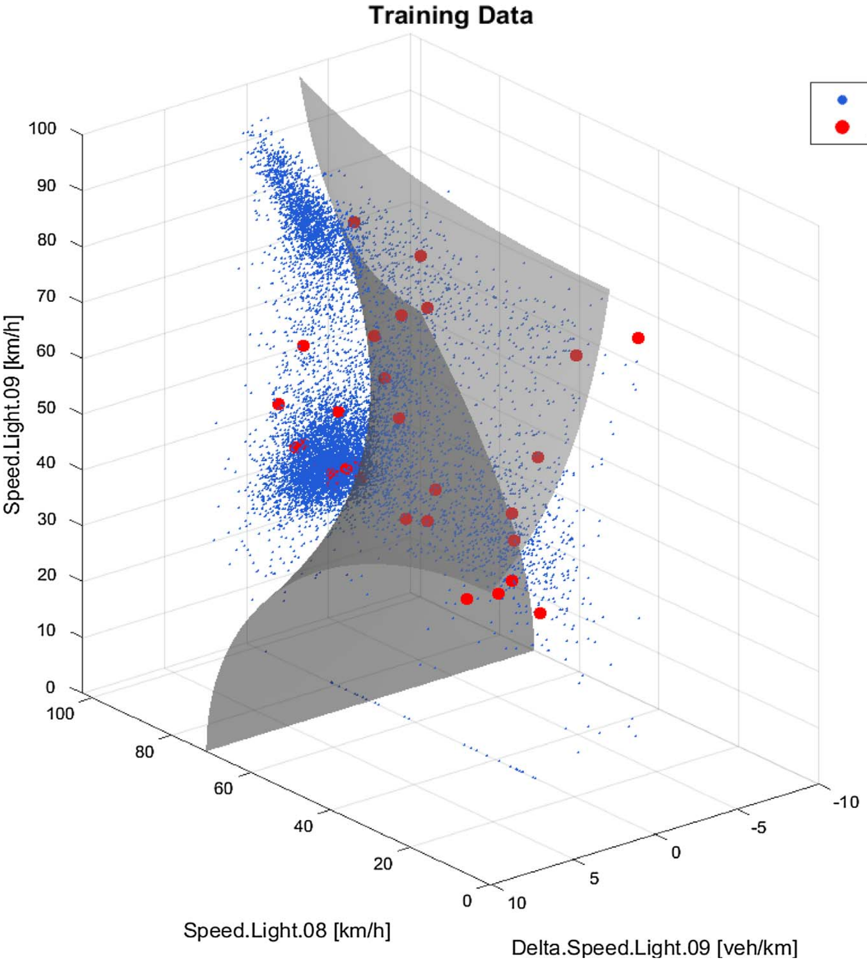


Fig. 16. Decision frontier for the three variables logistic regression for the training data-set.

AC08.

For this non-linear (in the variables) logistic model, the threshold probability is also set at $p_0 = 0.299\%$, corresponding to a false alarm rate of 21% over the estimation data set. Turning to validation/prediction, the model naturally achieved a sensitivity of 100% but decreased the false alarm rate to 20.17%., as shown in Fig. 16. That the non-linear model achieves better results than the linear model can be seen by comparing Figs. 14 and 16 below: the linear model had 10 accident events at the wrong side of the decision frontier, the non-linear model only six.

We also tried using a random-parameter logistic regression for both linear and non-linear logistic regression but the performance did not increase. Moreover, the likelihood ratio test suggests that the random-parameters logistic regression is not significantly better than the deterministic logistic regression. The estimated parameters of both models have the same sign and magnitude order. We also tried k-neighbors and CART but they ended up having considerably less predictive power than SVM and logistic regressions.

6. Robustness and model comparisons

We now turn to robustness. As clearly show, the sensitivity of the models may heavily depend on the partition of the data used so, what we did is to repeat 300 hundred times the following: we randomly select 80% of the data base, calibrate both SVM and logistic models and then validate using the remaining 20%.[5] We then calculated 300 values for sensitivity and false positive rates and then calculated the averages, maximum, minimum and standard deviations. The number of repetitions used was based in our experimental results, which showed that the mean of the sensitivity and false alarm rate stabilizes when the number of repetitions was around 200, therefore 300 hundred repetitions were chosen to be on the safe side. Also, is important to remark that with this number of

[5] In each logistic regression model construction, the threshold probability is chosen in order to reach a false positive rate near to 20% when making a prediction on the training base. This probability (percent) always ranges between the interval [0.29%, 0.30%], which is consistent with the accident proportion in the base considered (0.299% of the data).

Table 7
Prediction power for adjusted models.

Indicators	SVM radial kernel	SVM sigmoid kernel	SVM polinomial kernel	Logistic regression linear	Logistic regression nonlinear
Mean sensitivity (%)	69.06	68.50	77.13	62.26	67.89
Maximum sensitivity (%)	100	100	87.50	100	100
Minimum sensitivity (%)	53.50	54.64	52.06	45.63	59.17
Mean false alarm rate (%)	28.44	27.72	59.78	20.94	20.94
Standard deviation (%)	1.62	0.88	7.92	0.10	0.07

repetitions, the probability that two sub-samples are equal is indistinguishable from zero. The main results are shown in Table 7 and Fig. 17.

Table 7 shows that SVM models have high sensitivity percentages, particularly that of a degree 3 polynomial kernel, which reaches a mean prediction of almost 80%, however they provide high false positive rates as they overestimate the zones where accidents should occur, probably due to the base balancing when using SMOTE. Yet if SMOTE is not used for training, SVM does not deliver good sensitivities. On the other hand, logistic regression models show false positive rates near to 20% adjusted on the estimation base; an expected behavior when using cross validation with random selection. In this last category, the non-linear logistic regression model shows the best results, with a mean sensitivity of 67.89% (similar to that obtained through SVM models with radial and sigmoid kernel), and a mean false positive rate of 20.94% (much lower than the same SVM models).

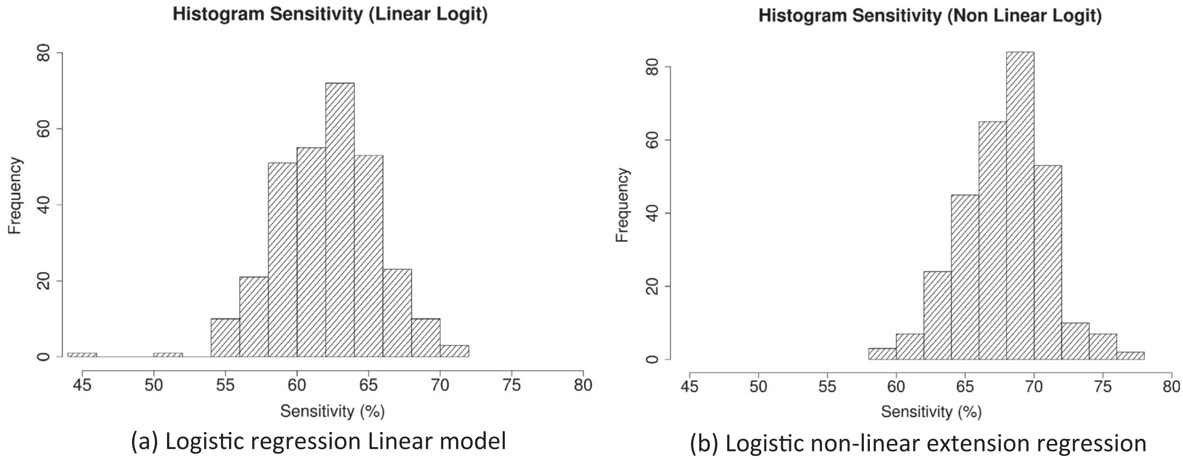
These results are quite promising when compared to the literature, as shown in Table 8 –which draws from Lin et al. (2015). First, because we achieve high sensitivity values with low false alarm rates. More importantly, because our results are averages over 300 randomly selected calibration/estimation data sets and, therefore, we are positive that they are not conditioned by a particular partition of the data set. Also, because the 300 validations were done over non-balanced data sets.

Keeping in mind these differences, note that studies that reach higher sensitivity rates than those we obtain here have a false positive rate much higher than the 20% we achieve. On the other hand, when Ahmed et al. (2012) projected a false alarm rate near to 20%, their sensitivity dropped to about 60% according to the ROC curve of this article. On the other hand, the high sensitivity percentage found by Sun and Sun (2015) has the methodological disadvantage of coming from an artificially balanced base, where for each accident record, only 5 normal situation records where selected, to later use this base for both training and validation purposes; therefore, the percentages shown does not necessarily reflect the predictive power it would actually have in real world situations.

A final robustness check worth of study, is whether the calibrated model would perform well on data collected after the period used for calibration. This would serve two purposes: first, it is what comes closer to learn what would have been the result should a real-time model been at work. Second, it enables a look to how far or close in the future are the models able to predict. Hence, we used the non-linear logistic regression model on traffic and crash data that was collected by Autopista Central on a period of time later than the one we had at hand (June and July 2016). The results are encouraging: the sensitivity was actually better than the mean, reaching a sensitivity of 75.03%, while keeping the false alarm rate at 22.47%.

7. Concluding remarks

Road accidents imply congestion, delays and sometimes loss of human life. That is why in the last two decades researchers have tried to stablish relations between crashes, flow states and environmental variables. Even though loop detectors (electromagnetic dispositive flow data collectors) exist since early 60s, predictive models for crash prediction appeared only in the early 2000s. These models showed that predicting road accidents is possible, but the lack of online data and the high failure rate of loop detectors have



(a) Logistic regression Linear model

(b) Logistic non-linear extension regression

Fig. 17. Sensitivity histogram (300 repetitions with 5-fold CV) in logistic regression.

Table 8
Prediction power for previous research.

Authors	Variable selection method	Classification method	Sensitivity (%)	False alarm rate (%)
Abdel-Aty et al. (2004)	N/A	Logistic regression	69	N/A
Pande and Abdel-Aty (2006)	Classification tree	Neural Network	57.14	28.83
Abdel-Aty et al. (2008)	Random forest	Neural Network	61	21
Hossain and Muromachi (2012)	Random multinomial logit	Bayesian Network	66	20
Ahmed and Abdel-Aty (2012)	Random forest	Matched case-control method	68	46
Lin et al. (2015)	Frequent Pattern tree	Bayesian Network	61.11	38.16
Sun and Sun (2015)	N/A	Dynamic Bayesian network	76.4	23.7

truncated the construction of computational tools. Recently, Automatic Vehicle Identification gates have been introduced in some urban expressways, such as Autopista Central in Chile. AVIs have almost no failures (less than 1%) and they are sometimes able to distinguish among vehicles classes such as Cars and SUVs, Buses and Trucks, and Motorcycles. In this study, techniques based on machine learning and logistic regression models to classify and forecast accidents on a stretch of the Autopista Central in Santiago are introduced. To the best of authors' knowledge, this paper is the first in making predictions based on disaggregated variables per type of vehicle using AVIs information. This allows isolating the contribution of each class to the increase the probability of accidents. Moreover, this paper is also the first to use non-artificially balanced data to validate the predictive models, which is quite important in order to think in a real-time application tool, and the first, as far as we know, to use repetitions to randomly select the calibration/estimation data set, in order to ensure robustness. The procedure described in this paper is as follows. First, we defined a stretch in Autopista Central to collect the flow data. The election was done in order to maximize the rate of accident per km per month. Second, we built a Random Forest model to classify the importance of the available variables. We complement it with visual inspection which permitted to identify the main explanatory variables of accident occurrence. Using these, SVM and logistic regression models were adjusted using the first 80% of the available data. Then the models were validated using the last 20%. The best models (radial SVM and non-linear extension logistic regression) predicted the 100% of the accidents with a relatively low false alarm rate of 20% approximately. To prove the robustness of our approach, we made 300 additional repetitions, randomly selecting the calibration/estimation data set and keeping the remaining for validation. We trained and estimated our models in each instance, thus obtaining robust average sensitivity and failure rates.

The main conclusions of our paper are:

1. In the studied stretch, the selected modeling variables are related only to vehicles in the “Car and pickup truck” category, which is directly related to the central location of this stretch, and its intrinsically urban nature. This is reflected on the traffic composition: 93.1% of the vehicles registered in the period studied belong to the “Car and pickup trucks” category (light). This means that, for this stretch, vehicle composition variables were not relevant, contrary to what one may have conjectured, a finding on itself. Yet, in extensions of the study to more rural stretches, we have find that variables related to the rest of the types of vehicles (particulary regarding the interaction of motor bikes and trucks, and their speed differences) are indeed of first order importance, showing the advantages of using disaggregate data. Moreover, our preliminary work in this stretch, which also has more distant traffic counters, has shown results as positive as the ones described in the paper.
2. SVM models reach a high percent of sensitivity, but tend to overestimate the “accident” prediction zone, prompting high rates of false positives, much higher than the 20% sought a priori. This can be caused by the Synthetic Minority Over-sampling Technique (SMOTE) we used to balance the data. Yet, without SMOTE, sensitivity itself drops.
3. The non-linear logistic model reaches, at validation, a mean sensitivity of 67.89% with just 20.94% of false positives. This sensitivity is comparable to the best results obtained in contemporary literature although their failure rates are usually higher. The comparison though is not really fair to our model, as we did not use a specific partition of data but used 300 random ones, and we validated on actual data and not artificially balanced data.
4. From this same model (non-linear logistic), the situation where accidents are most likely to occur is identified: In summary, the situation that causes the highest probability of accidents is: (i) substantially density drops upstream with ensuing high speeds (ii) unusual low speeds downstream. This concurs with empirical intuition and experience: a sudden, unusual traffic congestion prompts, once it starts to dissipate, a more aggressive behavior of drivers who try to recover their lost time by speeding yet, a couple of kilometers down they face unusual low speeds, causing braking maneuvers that can lead to crashes.

We believe that our results are promising and that studying the rest of the expressway stretches and in other hours is warranted. We expect that due to differences in traffic geometry, length, and vehicle composition on the different stretches and periods, models will change, both in terms of variables as in terms of specifications. From a methodological point of view, we would like to stress the importance of validating using the original data set and using more than one data partition to ensure robustness.

We conclude by highlighting what we consider is an important avenue of future research: the matter of which preemptive actions could be taken when an accident is predicted. This is indeed crucial, yet escapes the scope of this paper. Possibly, providing a reasonable response to this question will require the work of a multidisciplinary team composed by psychologist, occupational safety, health experts and engineers to decide the best measures to prevent the accidents predicted.

Acknowledgements

We thank Autopista Central, and particularly Cindy Carmona, Elena Hurtado and Christian Barrientos for providing us with data and collaborating enthusiastically with the project. We also thank Mauro Huenupi and Gabriel Montero for very insightful comments, and Erasmo Hidalgo for competent research assistance. We gratefully acknowledge financial support from the Complex Engineering Systems Institute, ISCI (grant CONICYT FB0816) and a CONICYT-Chile doctoral grant (Franco Basso).

References

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., 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.

Abdel-Aty, M.A., Hassan, H.M., Ahmed, M., Al-Ghamdi, A.S., 2012. Real-time prediction of visibility related crashes. Transp. Res. Part C: Emerg. Technol. 24, 288–298.

Ahmed, M., Abdel-Aty, M., Yu, R., 2012. Assessment of interaction of crash occurrence, mountainous freeway geometry, real-time weather, and traffic data. Transp. Res. Rec.: J. Transp. Res. Board (2280), 51–59.

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.

Akbani, R., Kwek, S., Japkowicz, N., 2004. Applying support vector machines to imbalanced datasets. Mach. Learn.: ECML 2004, 39–50.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Tree. Wadsworth International Group, Belmont, CA.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J. Artificial Intell. Res. 16, 321–357.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297.

Drosou, K., Georgiou, S., Koukouvinos, C., Stylianou, S., 2014. Support vector machines classification on class imbalanced data: a case study with real medical data. J. Data Sci. 12 (4), 143–155.

Fergani, B., 2016. Comparing HMM, LDA, SVM and smote-SVM algorithms in classifying human activities. In: Proceedings of the Mediterranean Conference on Information & Communication Technologies 2015. Springer, pp. 639–644.

Golob, T.F., Recker, W.W., 2004. A method for relating type of crash to traffic flow characteristics on urban freeways. Transp. Res. Part A: Policy Practice 38 (1), 53–80.

Hastie, T.J., Pregibon, D., 1992. Statistical models in S, chapter generalized linear models. Wadsworth & Brooks/Cole, 51.

Hastie, T., Friedman, J., Tibshirani, R., 2001. Model Assessment and Selection. The Elements of Statistical Learning. Springer, pp. 193–224.

Hossain, M., Muromachi, Y., 2012. A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways. Accid. Anal. Prev. 45, 373–381.

Kwak, H.-C., Kho, S., 2016. Predicting crash risk and identifying crash precursors on Korean expressways using loop detector data. Accid. Anal. Prev. 88, 9–19.

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.

Lv, Y., Tang, S., Zhao, H., Li, S., 2009. Real-time highway accident prediction based on support vector machines. In: Control and Decision Conference, 2009. CCDC'09. Chinese. IEEE, pp. 4403–4407.

Martens, D., Baesens, B., Van Gestel, T., Vanthienen, J., 2007. Comprehensible credit scoring models using rule extraction from support vector machines. Eur. J. Oper. Res. 183 (3), 1466–1476.

Pande, A., Abdel-Aty, M., 2006. Assessment of freeway traffic parameters leading to lane-change related collisions. Accid. Anal. Prevent. 38 (5), 936–948.

Rizzi, L.I., de Dios Ortúzar, J., 2003. Stated preference in the valuation of interurban road safety. Accid. Anal. Prev. 35 (1), 9–22.

Sawalha, Z., Sayed, T., 2006. Traffic accident modeling: some statistical issues. Can. J. Civ. Eng. 33 (9), 1115–1124.

Shi, Q., Abdel-Aty, M., 2015. Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transp. Res. Part C: Emerg. Technol. 58, 380–394.

Shi, Q., Abdel-Aty, M., Yu, R., 2016. Multi-level Bayesian safety analysis with unprocessed Automatic Vehicle Identification data for an urban expressway. Accid. Anal. Prev. 88, 68–76.

Sun, J., Sun, J., 2015. A dynamic Bayesian network model for real-time crash prediction using traffic speed conditions data. Transp. Res. Part C: Emerg. Technol. 54, 176–186.

Theofilatos, A., Yannis, G., 2014. A review of the effect of traffic and weather characteristics on road safety. Accid. Anal. Prev. 72, 244–256.

Theofilatos, A., Yannis, G., Kopelias, P., Papadimitriou, F., 2016. Predicting road accidents: a rare-events modeling approach. Transp. Res. Proc. 14, 3399–3405.

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.

Yu, R., Abdel-Aty, M., 2013. Utilizing support vector machine in real-time crash risk evaluation. Accid. Anal. Prev. 51, 252–259.

Yu, R., Abdel-Aty, M.A., Ahmed, M.M., Wang, X., 2014. Utilizing microscopic traffic and weather data to analyze real-time crash patterns in the context of active traffic management. IEEE Trans. Intell. Transp. Syst. 15 (1), 205–213.