Real-time crash prediction in an urban expressway using disaggregated data [1]

Franco Basso, Leonardo J. Basso, Francisco Bravo and Raúl Pezoa

Escuela de Ingeniería Industrial, Pontificia Universidad Católica de Valparaíso
Escuela de Ingeniería Industrial,Universidad Diego Portales
Departamento de Ingeniería Civil, Universidad de Chile

PhD Program Presentation, PUCV, Valparaíso, June 3rd, 2020.

[1] Basso, F., Basso, L. J., Bravo, F., & Pezoa, R. (2018). Real-time crash prediction in an urban expressway using disaggregated data. Transportation research part C: emerging technologies, 86, 202-219.

My research

Methodologies

Operations Research and Analytics.

Application areas

Logistics and Transportation.

Industries

Highways, Public Transportation, Wine, Mining.

My research

Publications [1/3]

(1) Varas, M., Basso, F., Maturana, S., Osorio, D., & Pezoa R. (2020). A multi-objective approach for supporting wine grape harvest operations. Computers & Industrial Engineering, 145: 106497.

(2) Contreras, J. P., Bosch, P., Varas, M., & Basso, F. (2020). A new genetic algorithm encoding for coalition structure problems. Mathematical Problems in engineering, 2020: 1203248.

(3) Basso, F., Basso, L. J., & Pezoa, R. (2020). The importance of flow composition in real-time crash prediction. Accident Analysis and Prevention, 137: 105436.

(4) Basso, F., Guajardo M., & Varas, M. (2020). Collaborative job scheduling in the wine bottling process. Omega, 91: 102021.

My research

Publications [2/3]

(5) Varas, M., Basso, F., Lüer-Villagra, A., Mac Cawley, A., & Maturana, S. (2019). Managing premium wines using an (s-1,s) inventory policy: a heuristic solution approach. Annals of Operations Research, 280: 351-376.

(6) Basso, F., Epstein L.D., Pezoa, R. & Varas, M. (2019). An optimization approach and a heuristic procedure to schedule battery charging processes for stackers of palletized cargo. Computers & Industrial Engineering, 133: 9-18.

(7) Basso, F., D'Amours, S., Rönnqvist, M., & Weintraub, A. (2019). A survey on obstacles and difficulties of practical implementation of horizontal collaboration in logistics. International Transactions in Operational Research, 26(3): 775-793.

(8) Frez, J., Baloian, N., Pino, J. A., Zurita, G., & Basso, F. (2019). Planning of Urban Public Transportation Networks in a Smart City. Journal of Universal Computer Science, 25(8), 946-966.

My research

Publications [3/3]

(9) Basso, F., Basso, L. J., Bravo, F., & Pezoa, R. (2018). Real-time crash prediction in an urban expressway using disaggregated data. Transportation Research Part C: Emerging Technologies, 86: 202-219.

(10) Varas, M., Maturana, S., Cholette, S., Mac Cawley, A., & Basso, F. (2018). Assessing the benefits of labelling postponement in an export-focused winery. International Journal of Production Research, 56(12): 4132-4151.

(11) Basso, F., Varas, M. (2017). A MIP formulation and a heuristic solution approach for the bottling scheduling problem in the wine industry. Computers & Industrial Engineering, 105: 136-145.

(12) Silva, J. D., Amaya, J. G., & Basso, F. (2017). Development of a predictive model of fragmentation using drilling and blasting data in open pit mining. Journal of the Southern African Institute of Mining and Metallurgy, 117(11): 1089-1094.

Outline

1. Introduction
2. Data set and preparation
3. Variable Selection
4. Classification Methods
5. Robustness and model comparisons
6. Concluding remarks
7. Pilot software
8. Appendix

Introduction

- Car accidents in cities are an important externality caused by traffic. Accidents imply congestion, delays and sometimes fatalities.
- For example, there were 94,879 road accidents in 2017 in Chile, the highest number ever. In that same year, 1,483 people died in road accidents (CONASET, 2018).
- Rizzi and 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.

Introduction

Purpose of the research

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.

Introduction

- There has been previous work on this area (Abdel-Aty et al., 2004; Lv et al., 2009; Hossain and Muromachi, 2012; Yu and Abdel-Aty, 2013; Sun and Sun, 2015; Yang et al., 2018; Wang et al., 2019).
- However, there are three main general differences with previous efforts.

What is new?

(1) We work with highly accurate and disaggregated data obtained from Automatic Vehicle Identification (AVI) system and provided by a major tolled urban highway in Santiago, Chile, Autopista Central.

(2) Our models are validated on the original unbalanced data set (where accidents are quite rare events), rather than on artificially balanced data.

(3) We do not use only one partition of the data set for calibration and validation but conduct 300 repetitions of randomly selected partitions.

Data set and preparation

- Autopista Central is an expressway in Santiago, Chile, which is 60.5 kms long and has a north-south orientation.
- The raw traffic data set they provided us with has traffic information from November 1st 2014 to April 30th 2016. We consider the afternoon rush hour, that is, Monday to Friday from 5:30p to 8.30p.

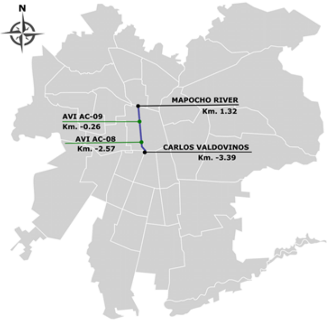
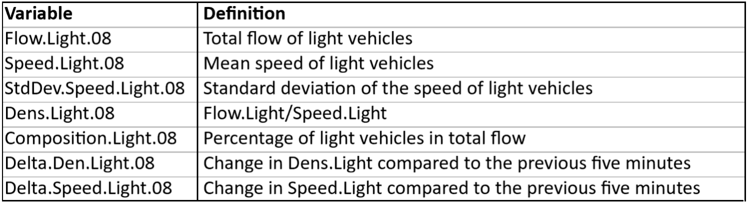


Figure 1: Section of the expressway studied.

Data set and preparation

- 10,745,766 vehicles crossed the two AVI gates during the studied period (5,298,683 correspond to the AVI gate AC-09 and 5,447,083 to the AVI gate AC-08).
- 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.
- 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.



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	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

Figure 2: Variables used for light vehicles crossing AVI gate AC-08.

Data set and preparation

- The final data set has 13,029 observations (5 minutes periods) of which only 39, i.e. 0.30% had an accident, confirming the rare event feature

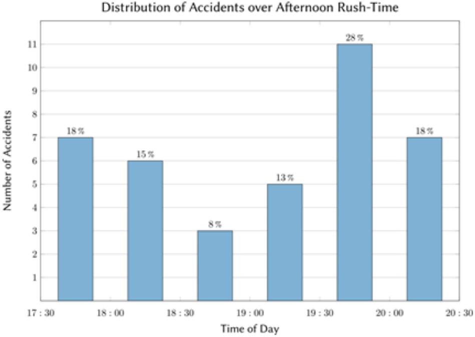
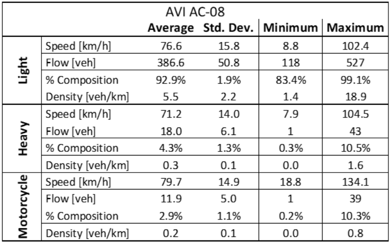


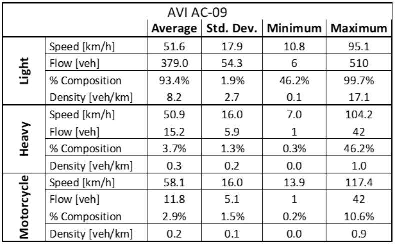
Figure 3: Distribution of the accidents for 30 minutes intervals

Data set and preparation



	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

Figure 4: Descriptive statistics of AVI gate AC-08.



	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

Figure 5: Descriptive statistics of AVI gate AC-09.

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.
- In order to do this, Pearson correlations, Random Forest (RF) techniques and graphical analysis are used.

What is Random Forest?

- 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).
- 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 (Breiman, 1984).

Variable Selection

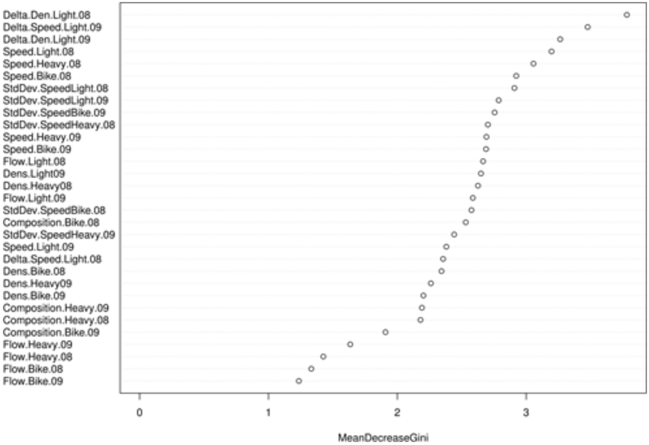


Figure 6: Change in Gini impurity index to determine variable importance

Variable Selection

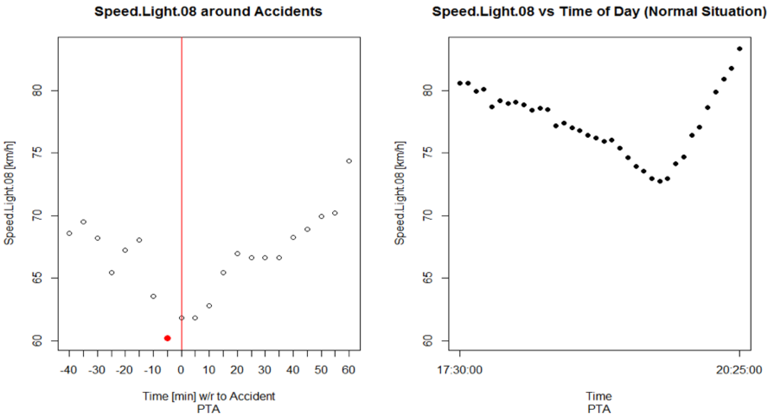


Figure 7: Speed.Light.08 behavior prior and after an accident.

Variable Selection

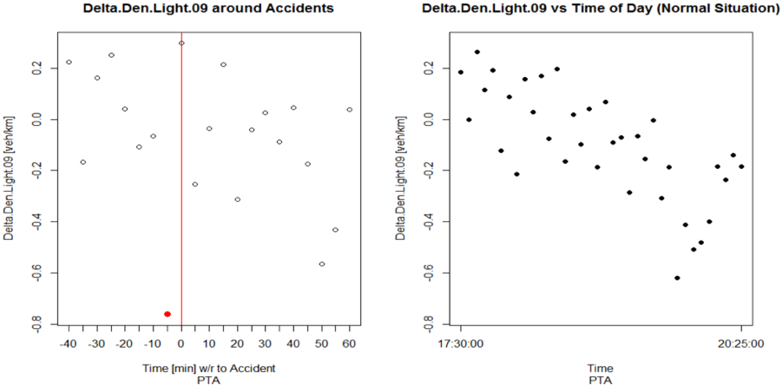


Figure 8: Delta.Den.Light.09 behavior prior and after an accident.

Classification Methods

- In this research, we used Logistic Regression and Support Vector Machine (SVM) as classification methods.

What is SVM?

- SVM was first introduced by Cortes and Vapnik (1995) and seek to find a separator hyperplane between two classes in order to maximize the distance between the classes and the decision frontier.
- If data is not linear separable, it is possible to add slack variables to penalize misclassification.
- If the decision function is nonlinear, it is possible to map the data to another Euclidean space through a kernel function.

Classification Methods

- Because accidents are rare events, 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.

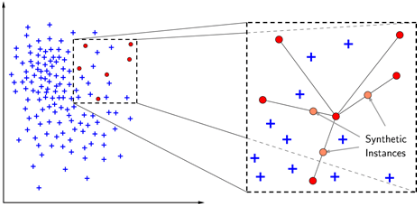


Figure 9: Oversampling using SMOTE.

Classification Methods

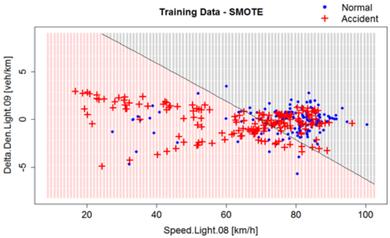


Figure 10: Decision frontier for SVM with radial kernel for the training SMOTE data-set.

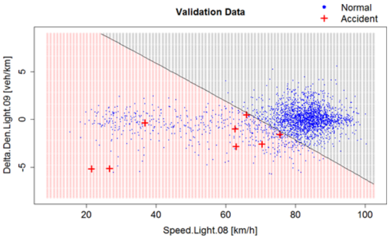


Figure 11: Decision frontier for SVM with radial kernel for the full validation data-set.

Classification Methods

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

Figure 12: Maximum likelihood parameters for the three variables logistic regression model.

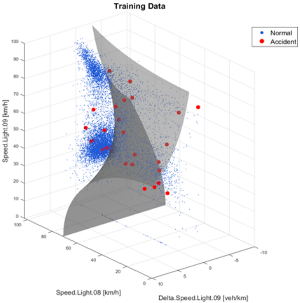
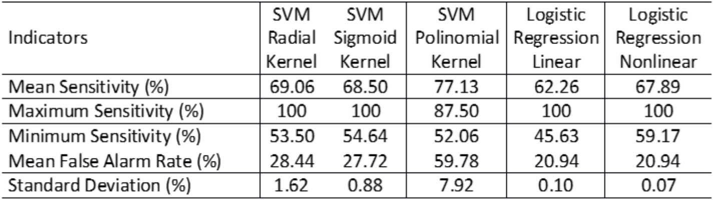


Figure 13: Decision frontier for the three variables logistic regression for the training data-set.

Robustness and model comparisons

- 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%.
- We then calculated 300 values for sensitivity and false positive rates and then calculated the averages, maximum, minimum and standard deviations.



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

Figure 14: Prediction power for adjusted models.

Concluding remarks

- 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.
- 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.
- 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.
- 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.

Concluding remarks

Future research

- Assess the importance of flow composition in real-time crash prediction (Basso et al., 2020)
- Does vehicle message signs help to prevent crashes? (submitted paper)
- Include trajectory variables (working paper)
- Use machine learning models to predict drivers' behaviors (lane change, speed reduction, etc.)
- Determine drivers' risk profiles using passive data.
- Extend the models to local roads.

Support Vector Machines

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:

$$\begin{array}{l}\min \frac{1}{2}||w||^2 \\ \text{s.t.} \\ y_i(x_i \cdot w + b) \geq 1 \qquad i = 1, ..., n\end{array}$$

If data is not linear separable, it is possible to add slack variables ξ_i to penalize misclassification.

$$\begin{array}{l}\min \frac{1}{2}||w||^2 + C\sum_i \xi_i \\ \text{s.t.} \\ y_i(x_i \cdot w + b) \geq 1 - \xi_i \qquad i = 1, ..., n \\ \xi_i \geq 0 \qquad i = 1, ..., n\end{array}$$

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

$$\begin{array}{l}\max \sum_i a_i - \frac{1}{2}\sum_i \sum_j a_i a_j y_i y_j x_i \cdot x_j \\ \text{s.t.} \\ \sum_i a_i y_i = 0 \\ 0 \leq a_i \leq C \qquad \forall i\end{array}$$

Support Vector Machines

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 $H : K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$. We have 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: $K(x_i, x_j) = tanh(\gamma x_i \cdot x_j + 1)$

Random Forest

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, 1984) is defined for node as:

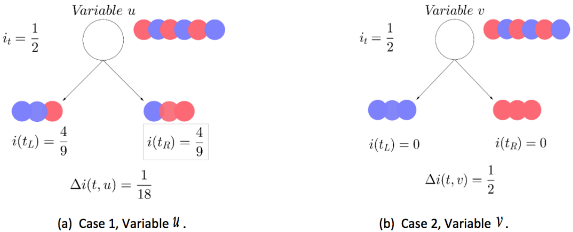
$$i(t) = 2p(1/t)p(2/t)$$

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:

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

Random Forest

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.



(a) Case 1, Variable u .

(b) Case 2, Variable v .

Figure 15: Example of two splittings, with case 2 (variable V) preferred.

References I

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, F., and Hsia, L. (2004). Predicting freeway crashes from loop detector data by matched case-control logistic regression. *Transportation Research Record: Journal of the Transportation Research Board*, (1897):88–95.

Akbani, R., Kwek, S., and Japkowicz, N. (2004). Applying support vector machines to imbalanced datasets. In *European conference on machine learning*, pages 39–50. Springer.

Basso, F., Basso, L. J., and Pezoa, R. (2020). The importance of flow composition in real-time crash prediction. *Accident Analysis & Prevention*, 137:105436.

Breiman, L. (1984). *Classification and regression trees*. Routledge.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

References II

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Hossain, M. and Muromachi, Y. (2012). A bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways. *Accident Analysis & Prevention*, 45:373–381.

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

Rizzi, L. I. and Ortúzar, J. d. D. (2003). Stated preference in the valuation of interurban road safety. *Accident Analysis & Prevention*, 35(1):9–22.

Sun, J. and Sun, J. (2015). A dynamic bayesian network model for real-time crash prediction using traffic speed conditions data. *Transportation Research Part C: Emerging Technologies*, 54:176–186.

Wang, L., Abdel-Aty, M., Ma, W., Hu, J., and Zhong, H. (2019). Quasi-vehicle-trajectory-based real-time safety analysis for expressways. *Transportation Research Part C: Emerging Technologies*, 103:30–38.

References III

Yang, K., Wang, X., and Yu, R. (2018). A bayesian dynamic updating approach for urban expressway real-time crash risk evaluation. *Transportation research part C: emerging technologies*, 96:192–207.

Yu, R. and Abdel-Aty, M. (2013). Utilizing support vector machine in real-time crash risk evaluation. *Accident Analysis & Prevention*, 51:252–259.