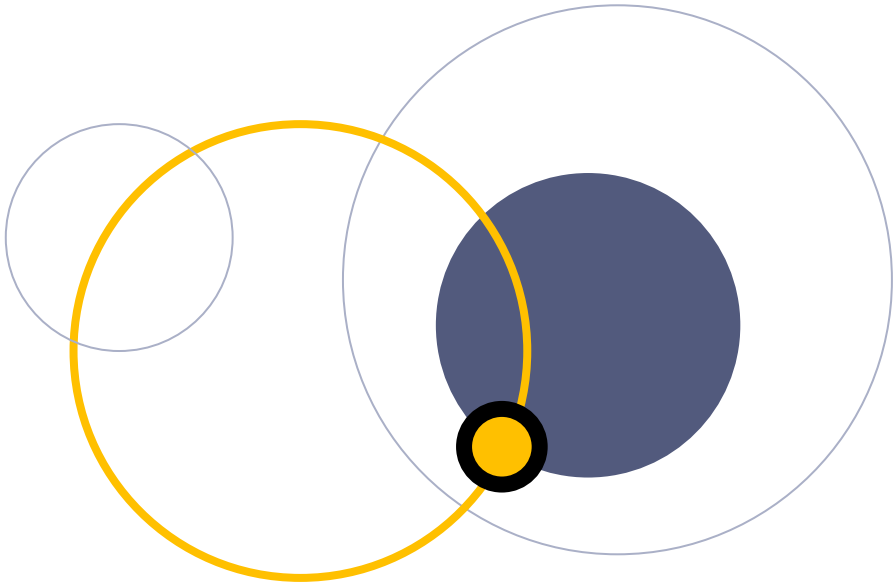
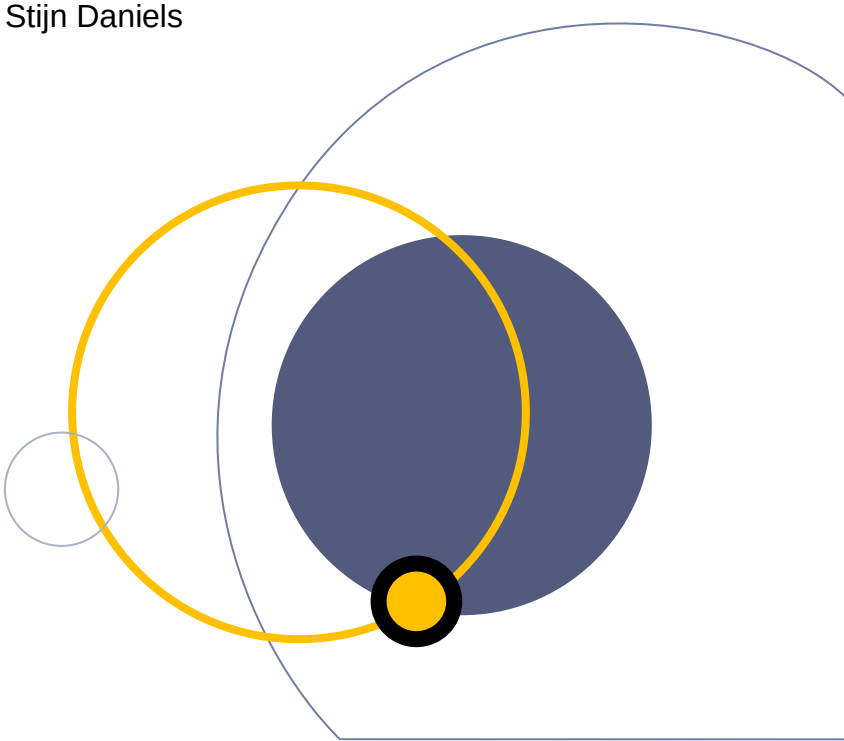


Steunpunt beleidsrelevant onderzoek 3de generatie Verkeersveiligheid

Real-time crash risk prediction models using loop detector data for dynamic safety management system applications

Ali Pirdavani, Maarten Magis, Ellen De Pauw, Stijn Daniels

RA-2014-003



17/7//2014

Wetenschapspark 5 bus 6 | 3590 Diepenbeek

Consortium UHasselt, KU Leuven en VITO

Dit rapport kwam tot stand met de steun van de Vlaamse Overheid, programma ‘Steunpunten voor Beleidsrelevant Onderzoek’. In deze tekst komen onderzoeksresultaten van de auteur(s) naar voor en niet die van de Vlaamse Overheid. Het Vlaams Gewest kan niet aansprakelijk gesteld worden voor het gebruik dat kan worden gemaakt van de meegedeelde gegevens.

Het Steunpunt Verkeersveiligheid 2012-2015 voert in opdracht van de Vlaamse overheid beleidsondersteunend Wetenschappelijk onderzoek uit over verkeersveiligheid. Het Steunpunt Verkeersveiligheid is een samenwerkingsverband tussen de Universiteit Hasselt, de KU Leuven en VITO, de Vlaamse Instelling voor Technologisch Onderzoek.

Table of content

List of Figures

List of Tables

Summary

Real-time crash risk prediction models using loop detector data for dynamic safety management system applications

There is a growing trend in development and application of real-time crash risk prediction models within dynamic safety management systems. These real-time crash risk prediction models are constructed by associating crash data with the real-time traffic surveillance data, collected by double loop detectors. The objective of this paper is to develop a real-time prediction model that will potentially be utilized within safety management systems. This model aims to predict the traffic safety condition of a motorway where potential prediction variables are confined to traffic related characteristics. Given that the dependent variable (i.e. traffic safety condition) is considered dichotomous (i.e. “no-crash” or “crash”), the binary logistic regression technique is selected for model development. The crash and traffic data used in this study were collected between June 2009 and December 2011 on a part of the European route E313 in Belgium between Geel-East and Antwerp-East exits, on the direction towards Antwerp. The results of analysis show that several traffic flow characteristics such as standard deviation of speed and occupancy at the upstream loop detector, and the difference in average speed on upstream and downstream loop detectors are significantly contributing to the crash occurrence prediction. The final chosen model is able to predict more than 60% of crash occasions while it predicts more than 90% of no-crash instances correctly. The findings of this study can be used to predict the likelihood of crashes on motorways within dynamic safety management systems.

Samenvatting

Real-time risicomodellen op basis van inductieve lusdetectordata voor toepassingen in dynamische verkeersmanagementsystemen

Dynamisch verkeersmanagement wordt reeds wereldwijd toegepast met als doel het verbeteren van mobiliteit en de verkeersveiligheid. In de aansturing van deze systemen wordt meer en meer gebruik gemaakt van risicomodellen. Deze risicomodellen trachten door middel van geobserveerde verkeersdata, vaak verzameld door middel van dubbele inductieve lussen in het wegdek, de verkeers(on)veiligheid op een bepaalde locatie te voorspellen. Het doel van dit onderzoek is de ontwikkeling van een risicomodel dat kan gebruikt worden binnen het Vlaamse verkeersveiligheidsbeheer op autosnelwegen. Dit model tracht op basis van geobserveerde data op een bepaalde autosnelweglocatie te voorspellen of de situatie al of niet veilig is. Gegeven het feit dat de afhankelijke variabele (de verkeersveiligheidssituatie) dichotoom is, wordt de binaire logistische regressie techniek gehanteerd voor de ontwikkeling van het model. De onderzoekslocatie betrof de E313 tussen Geel-Oost en Antwerpen-Oost, in de richting van Antwerpen. Hier werden alle data, ongevallen enerzijds en verkeersdata anderzijds, verzameld tussen juni 2009 en december 2011. Het model resulteerde in drie significante variabelen die de (on)veiligheid op een bepaalde locatie konden voorspellen, namelijk (1) de standaard deviatie van snelheid stroomopwaarts, (2) de bezetting van de lussen stroomopwaarts en (3) het verschil in de gemiddelde snelheid tussen de lussen stroomafwaarts en stroomopwaarts. Het finale model kan meer dan 60% van de gevaarlijke situaties en 90% van de veilige situaties correct voorspellen. De resultaten van deze studie kunnen gebruikt worden om door middel van dynamisch verkeersmanagement systemen gevaarlijke situaties op autosnelwegen te voorspellen en bijgevolg de nodige maatregelen te treffen om ongevallen te voorkomen.

1 Introduction

In the recent years, proactive traffic management systems have increasingly attracted researchers and policy makers' attention. These systems, which are mainly implemented on motorways, are meant to improve traffic safety by smoothening the traffic flow. In such dynamic safety management systems, real-time crash risk prediction models are major elements. These models estimate the likelihood of crash occurrence by using real-time traffic flow characteristics that are collected by traffic surveillance systems such as loop detectors. These models can dynamically evaluate the traffic safety condition of motorways and identify crash conditions that would potentially lead to crash occurrence. When such crash condition is identified, proactive safety countermeasures can be implemented to alleviate crash occurrence risk. Among others, variable speed limits (Lee et al., 2004, 2006c; Abdel-Aty et al., 2006; Jo et al., 2012), ramp metering (Abdel-Aty et al., n.d.; Lee et al., 2006b) and intelligent speed adaptation (Chen et al., 2002; Carsten and Tate, 2005; Servin et al., 2008; Lai et al., 2012) are effective measures that are known to improve the traffic safety. These measures are intended to smoothen the traffic flow, increase average time headways, reduce speed variation and subsequently improve the traffic safety. For instance, safety benefits will be gained by simultaneously lowering down the speed upstream and increasing the speed downstream of the location where a crash condition is identified by the real-time crash risk prediction models.

In recent years, several studies were conducted where real-time crash risk prediction models were developed by associating real-time traffic flow data with crash data (Lee et al., 2002, 2003; Chang and Chen, 2005; Oh et al., 2005; Abdel-Aty et al., 2006; Lee et al., 2006c; Oh et al., 2006; Pande and Abdel-Aty, 2006a; Abdel-Aty et al., 2007; Zheng et al., 2010; Pande et al., 2011; Xu et al., 2012; Abdel-Aty et al., 2012; Xu et al., 2013; Ahmed and Abdel-Aty, 2013).

The most commonly used modeling technique in developing real-time crash prediction model is logistic regression. Lee et al. (Lee et al., 2002, 2003) investigated a number of traffic flow characteristics that would be linked with crash occurrence on the Gardiner Expressway in Toronto. They developed an aggregate log-linear model by associating the crash precursor variables with crash data observed over a period of 13 months. The results of their analysis revealed that the variation of speed, speed difference between upstream and downstream loop detectors and traffic density are significant predictors of crash occurrence. In a study conducted by Abdel-Aty et al. (Abdel-Aty et al., 2004) matched case-control logistic regression models were developed to associate real-time traffic flow characteristics with crash likelihood. The results of this study showed that the likelihood of crash occurrence would be predicted by speed variation on downstream loop detector station and the average occupancy on the upstream loop detector station. Abdel-Aty et al. (Abdel-Aty et al., 2005) further extended their study by developing crash prediction models under low speed and high speed traffic regimes. Zheng et al. (Zheng et al., 2010) also used a matched case-control logistic regression model to assess the impacts of speed variance on crash occurrence, whereas Ahmed and Abdel-Aty (Ahmed and Abdel-Aty, 2012) used this technique to develop a real-time crash risk prediction model. Hourdos et al. (Hourdos et al., 2006) developed a binary-response logistic model where crash-prone conditions were identified by using real-time traffic characteristics. The results of analysis showed that the crash likelihood increases when the difference in speed variability increases. In another study conducted by Lee et al. (Lee et al., 2006a) the real-time traffic factors associated with sideswipe crashes were investigated by means of logistic regression models. The results revealed that the variation in speed and the variation in flow were among the most significant predictors of sideswipe crashes. Golob et al. (Golob et al., 2008) investigated the level of safety on freeways, based on data collected from single loop detectors and by means of multinomial logistic regression models. The results showed significant relationships between traffic flow data and crash likelihood. Recently, Xu et al. (Xu et al., 2013) utilized binary logistic models to predict crash likelihood and severity. In their study, several traffic characteristics such as the upstream occupancy, the upstream speed variance, the downstream speed variance and etc. were found as significant predictors of crash likelihood.

Besides application of logistic regression technique, other approaches were followed to associate crash likelihood and real-time traffic flow characteristics. Oh et al. (Oh et al., 2005) developed a Bayesian model in which the standard deviation of speed was found to be an appropriate predictor of hazardous traffic conditions. Abdel-Aty and Pande (Abdel-Aty and Pande, 2005) applied the probabilistic neural network model to predict crash occurrences on freeways. In another study conducted by Chang et al. (Chang and Chen, 2005) classification and regression tree technique was employed to analyze freeway accident frequency. The results of this study revealed that the average

daily traffic and precipitation were the two major determinants of crash occurrence. Pande et al. (Pande and Abdel-Aty, 2006b) also developed a crash risk prediction model based on the classification tree and neural network while random forests technique was used by Pande et al. (Pande et al., 2011). In the context of real-time crash prediction modeling, there are also other employed techniques such as Bayesian semi-parametric Cox (Ahmed et al., 2012) and Stochastic Gradient Boosting (Ahmed and Abdel-Aty, 2013).

As can be seen from the literature, there exist different traffic flow characteristics that are associated with crash occurrence, such as speed, speed variation, traffic density, occupancy and etc. The primary objective of this research is set to investigate the possibility of developing real-time crash prediction models based on the traffic flow characteristics that are collected by double loop detector stations in Flanders, Belgium. To this end, the binary logistic regression technique will be employed for model development. This is considered as the first step in realization of a proactive highway safety management system. When the developed models predict the crash condition appropriately, the transportation authorities are enabled to implement crash prediction countermeasures within available dynamic traffic management systems in order to improve the traffic safety conditions of motorways.

2 Data preparation

2.1 Study area

The study area in this research is a part of the European route E313 in Belgium between Geel-East and Antwerp-East exits, on the direction towards Antwerp. The total length of the studied road segment is about 42.5 km. However, because of problems with the loop detectors, a 10 km length segment was excluded from the study (see Figure 1). From the starting point of the study area up to 10 km before the ending point, the E313 has two lanes (+ hard shoulder) and the speed limit is 120 km/h. In the last 10 km, the motorway has three lanes (+ hard shoulder/bus lane) and the speed limit is 100 km/h.

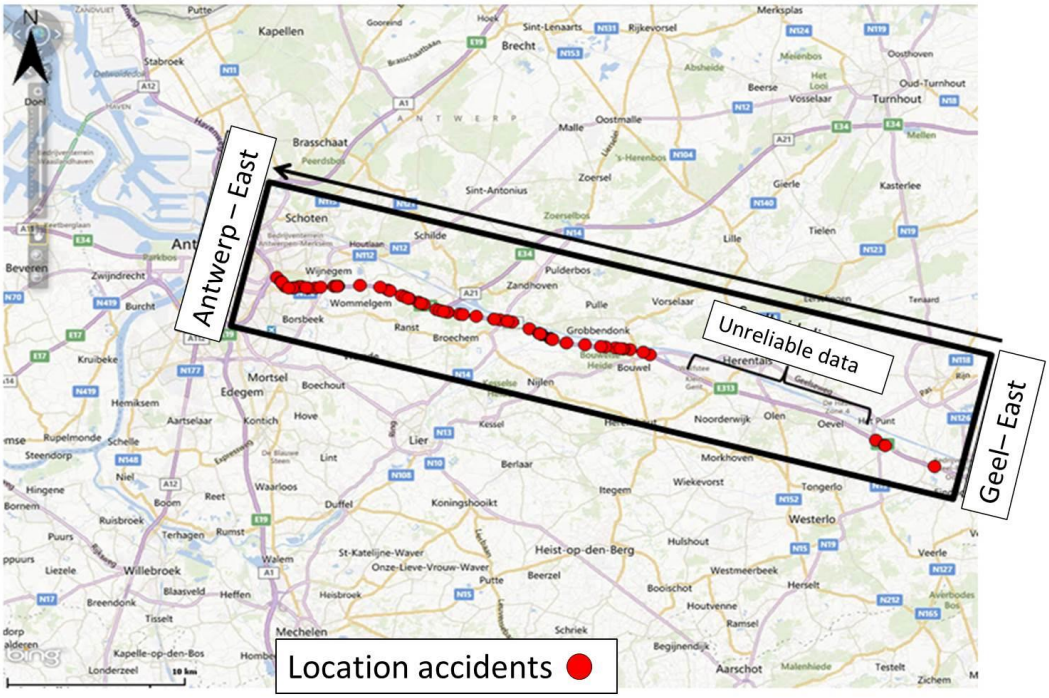


Figure 1: Study area and crash locations.

The primary crash dataset includes all crashes that occurred in the study area between June 2009 and December 2011. Among others a total number of 78 crash records are selected to be considered in model development. Due to the necessity of having precise crash occurrence time (in the order of 1 minute) and since the crash data (gathered by the police) were obtained from a different authority than the one which provides traffic flow data (Ministry of Mobility and Public Works, Flemish Traffic Center), the accuracy of crash occurrence time was double checked by matching these two datasets. To this end, for each crash record the traffic flow data derived from adjacent loop detectors were collected for a period of one hour around the crash occurrence time. Subsequently all crash records were verified whether their corresponding traffic flow data show any speed-drop event or not. This was carried out to ensure that each crash record perfectly matches with its linked traffic flow data. If a speed-drop event is observed for a crash record, then this record is considered as a valid record and, therefore, is added to the final dataset. An example of this matching task is shown in Figures 2, 3 and 4 where changes in traffic flow characteristics such as average speed, traffic volume and occupancy (i.e. percentage of time that a loop detector is occupied by a vehicle) are observed for a crash, which occurred at 17:11. These figures are the outputs of the program “Mindat” which are provided by the Flemish Traffic Center. This program enables users to derive different traffic flow characteristics for any specific place and time. Traffic flow data were collected from four consecutive loop detectors spaced at approximately 750 meters; one loop detector downstream (DS1) and three loop detectors upstream (US1-3). The first upstream loop detector station is named US1; the subsequent stations in the upstream direction were labeled US2 and US3.

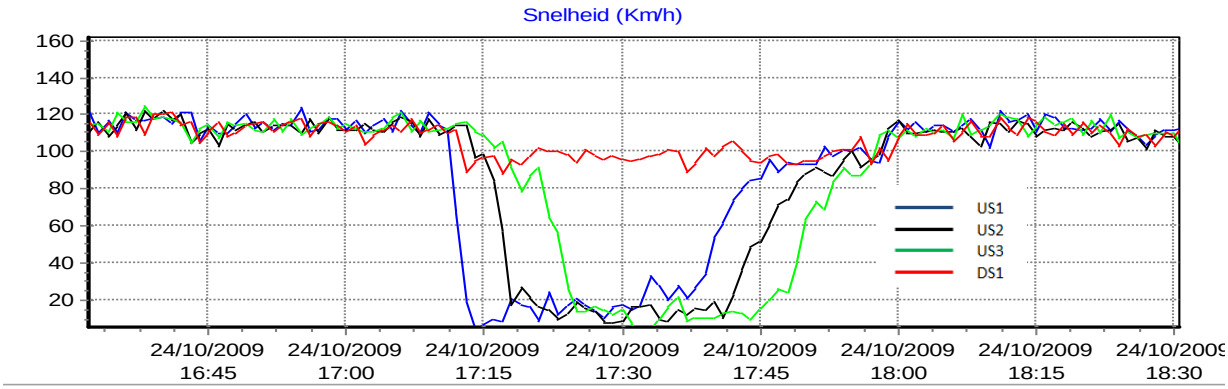


Figure 2: Mindat” program output: Speed (km/h)µ

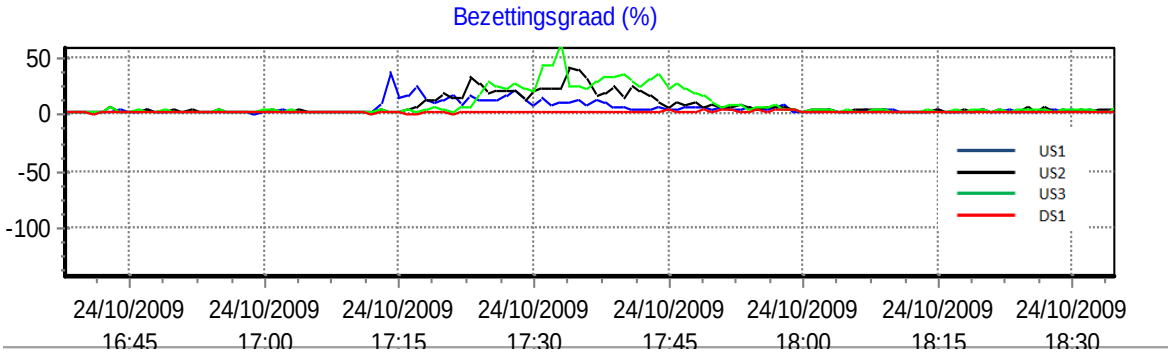


Figure 3: “Mindat” program output: Occupancy (percentage)

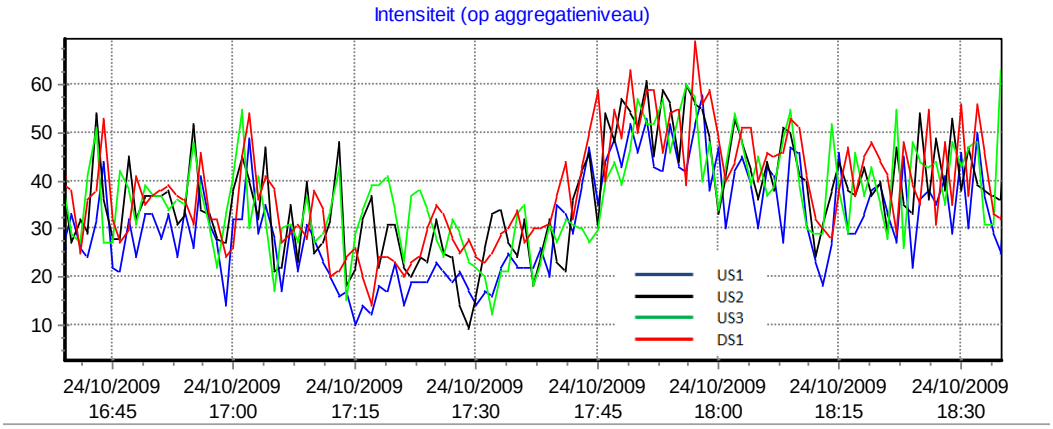


Figure 4: "Mindat" program output: Traffic volume (veh/min)

2.2 Aggregation levels

The next step in data preparation is the data aggregation. The 1-minute raw data seemed to have random noise and, therefore, the primary raw data should be combined into 5-min level (Abdel-Aty et al., 2012). The extracted raw data were then aggregated to three different aggregation levels, namely 5-minute, 10-minute and 15-minute intervals prior to crash occurrence time. All these three aggregation levels will be investigated to identify the best level that will result in better crash prediction.

In the next step and for preparing the complete dataset that will be used for the modeling task, four non-crash cases were also taken from the same location, the same day of the week and the same time given the condition that no crash had occurred within one hour time period around the targeted time. To eliminate the seasonal effects and to avoid possible bias resulting from dissimilar traffic patterns on different days of the week, non-crash cases were extracted from exactly one and two weeks before and after the crash occurrence time. All non-crash cases matched the condition that no crash had occurred within a one hour time period around the targeted time. This results in utilizing traffic flow data for the following records and for each location:

- Exactly two weeks before crash occurrence
- Exactly one week before crash occurrence
- Crash occurrence
- Exactly one week after crash occurrence
- Exactly two weeks after crash occurrence

To summarize, the final dataset consists of the traffic flow data corresponding to each crash record and four matched non-crash records. This dataset includes 390 observations (i.e. 78 crashes and 312 non-crash records).

2.3 Traffic flow characteristics

Several traffic flow variables are collected by double loop detectors and might be relevant to this study. However, in order to save time and effort in the model development stage, a pre-analysis is performed to minimize the number of potential explanatory variables. To this end, firstly the non-parametric Spearman's correlation test was performed to investigate which variables do have a significant correlation with the dependent variable (i.e. safety condition). After removing all uncorrelated variables (e.g. occupancy and average speed on the first downstream loop detector station), due to the existence of inter-relationship among remained variables, the variance inflation factor test needs to be performed to ensure that the collinearity issue do not exist among explanatory variables (Kutner et al., 2004). Final variables that will be considered for model development are listed in Tables 1-3 together with their descriptive statistics. These variables are prepared for different time intervals before crash occurrence time (i.e. 5-minute, 10-minute and 15-minute interval).

Here is a short definition of each variable:

Variable	Definition
TV_US1:	Traffic volume on first upstream loop detector station
STDEV_SP_US1:	Standard deviation of speed on first upstream loop detector station
OC_US1:	Occupancy (% time that a loop detector is occupied) on first upstream loop detector station
Diff_STDEV_SP_US1-DS1:	Difference between standard deviation of speed on first upstream and downstream loop detector stations
Diff_SP_US1-DS1:	Difference between average speed on first upstream and downstream loop detector stations
TV_DS1:	Traffic volume on first downstream loop detector station
STDEV_SP_DS1:	Standard deviation of speed on first downstream loop detector station

	TV_US1	STDEV_SP_US1	OC_US1	Diff_STDEV_SP_US1-DS1	Diff_SP_US1-DS1	TV_DS1	STDEV_SP_DS1
Traffic safety condition: Crash (dependent variable = 1)							
Min	0.1	1	0.3	0.1	0.2	0	0
Max	34.1	46.5	46.8	40.4	101	24.2	53.7
Mean	13.65	17.71	20.13	12.71	35.97	11.99	9.73
St. Deviation	6.66	11.97	11.47	10.99	28.21	6.087	9.31
Traffic safety condition: No-crash (dependent variable = 0)							
Min	1	1	0.3	0	0	0.8	0.9
Max	33.9	57.5	48.3	52.3	120.5	33.9	36.2
Mean	18.27	7.72	10.45	4.65	17.82	16.53	8.06
St. Deviation	6.35	7.38	7.83	6.53	27.42	6.99	6.85

Table 1: Descriptive statistics of final variables for 5-minute interval

	TV_US1	STDEV_SP_US1	OC_US1	Diff_STDEV_SP_US1-DS1	Diff_SP_US1-DS1	TV_DS1	STDEV_SP_DS1
Traffic safety condition: Crash (dependent variable = 1)							
Min	0.1	0.8	0.2	0.2	0	1.2	3
Max	34.5	46.3	37.1	42.4	83.2	27.3	49.6
Mean	15.76	22.74	16.43	11.85	27.99	13.91	15.51
St. Deviation	6.72	11.03	8.88	9.97	24.43	6.44	10.71
Traffic safety condition: No-crash (dependent variable = 0)							
Min	1.2	1.2	0.4	0	0	1.2	1.5
Max	33.6	38.7	43.4	30.4	119.9	33.2	38.8
Mean	18.27	9.19	10.26	3.91	16.39	16.52	9.64
St. Deviation	6.15	7.55	7.08	5.21	26.87	6.81	7.58

Table 2: Descriptive statistics of final variables for 10-minute interval

	TV_US1	STDEV_SP_US1	OC_US1	Diff_STDEV_SP_US1-DS1	Diff_SP_US1-DS1	TV_DS1	STDEV_SP_DS1
Traffic safety condition: Crash (dependent variable = 1)							
Min	1.2	8.9	0.6	0.1	0	0	0
Max	33.7	52.5	30.5	42.2	88.7	29	47.7
Mean	16.73	26.04	14.85	11.92	21.46	14.88	17.05
St. Deviation	6.01	10.02	7.25	10.23	20.08	6.62	10.57
Traffic safety condition: No-crash (dependent variable = 0)							
Min	0	0	0	0	0	1.3	1.7
Max	33	35.6	36	90.7	118.4	31.4	90.7
Mean	18.15	10.16	10.19	4.61	16.45	16.58	11.48
St. Deviation	5.94	7.96	6.74	10.52	28.06	6.68	11.70

Table 3: Descriptive statistics of final variables for 15-minute interval

3 Model development

3.1 Model structure

This study aims to predict the traffic safety condition of motorways by associating crash data with traffic flow characteristics that are collected by traffic loop detectors. Due to the dichotomous nature of the dependent variable Y (i.e. dependent variable can only take two values; Y=1 for crash condition and Y=0 for no-crash condition), application of binary logistic regression technique is appropriate. This type of model facilitates the probability estimate of being involved in a crash or no-crash condition based on the independent variables incorporated into the regression model. These independent variables can be categorical or continuous. The binary logistic regression model assumes a binomial distribution for the dependent variable. The probability of experiencing a crash or no-crash condition is modeled as the following:

$$\pi(x) = \frac{e^{g(x)}}{1+e^{g(x)}} \quad (1)$$

The Logit transformation of the $\pi(x)$ logistic function is shown in Eq. 2:

$$g(x) = ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (2)$$

Where $\pi(x)$ describes the probability of experiencing crash or no-crash condition. This probability falls between 0 and 1 (i.e. $0 \leq \pi(x) \leq 1$); values close to 1 signify crash conditions while values close to 0 denote no-crash conditions. x_i's are the independent variables and β_i's are the regression coefficients for each variable. These coefficient estimates determine the odds ratio of crash occurrence. The odds of an event are defined as the probability of the outcome event occurring divided by the probability of the event not occurring (Hosmer. et al., 2013). The odds ratio that is equal to the exponential of the β_i indicates the relative amount by which the odds of the outcome increase (ratios greater than 1.0) or decrease (ratios less than 1.0) when the value of the independent variable increases by 1.0 unit.

3.2 Model validation technique

In this study the n-fold cross-validation technique is employed to validate the accuracy and robustness of the prediction models (Olson and Delen, 2008). The n-fold cross-validation technique minimizes the possible bias caused by the random sampling of the training and testing datasets. In the n-fold cross-validation technique, the complete dataset is equally divided into n subsets. In each step of the model development one subset is considered as for validation dataset while n-1 subsets are used as for training dataset. The cross-validation process is then repeated n times, when each of the subsamples

will be used only once as the validation data. Subsequently the n results from the n developed models can be averaged or combined to deliver one single estimation. In this study, a 10-fold cross-validation approach is followed.

3.3 Model development

For developing the real-time prediction models, the final variables (see Tables 1, 2 and 3) were considered separately for each time interval. In other words, three binary logistic regression models were developed using explanatory variables of each time interval (i.e. 5-minute, 10-minute and 15-minute intervals). The results of analysis showed that the model developed based on 5-minute data outperforms the other two models and, therefore, is selected as the final prediction model. All models' classification results are reported in Tables 4, 5 and 6.

4 Model performance evaluation

The classification of results are shown in Tables 4, 5 and 6; commonly referred to as contingency table or confusion matrix. In a binary prediction problem, the outcomes are labeled either as positive or negative. In the context of this study and since the ultimate objective is to predict crash conditions, a positive outcome is set to be a crash condition while predicting a no-crash condition is considered as a negative outcome. Hence, there will be four possible outcomes by which the prediction accuracy of the model can be evaluated. If the outcome of a prediction is positive and the observed value is also positive, then this condition is considered as true positive (TP) while if the observed value is negative then it is stated to be a false positive (FP). Similarly, a true negative (TN) will occur when both the prediction outcome and the observed values are negative, and false negative (FN) is when the prediction outcome is negative while the observed value is positive. Model outcomes that are labeled with this convention are shown in Tables 4, 5 and 6.

Contingency Table[a]				
Observed		Predicted		ROC = 0.853
		Traffic safety condition		Percentage Correct
		Crash	No-crash	
Traffic safety condition	Crash	47 (TP)	31 (FN)	60.26
	No-crash	29 (FP)	283 (TN)	90.71
Overall Percentage				84.62
a. The classification threshold value is 0.4				

Table 4: contingency table of the 5-minute prediction model

Contingency Table[a]				
Observed		Predicted		ROC = 0.806
		Traffic safety condition		Percentage Correct
		Crash	No-crash	
Traffic safety condition	Crash	38 (TP)	40 (FN)	48.72
	No-crash	22 (FP)	290 (TN)	92.95
Overall Percentage				84.1
a. The classification threshold value is 0.4				

Table 5: The contingency table of the 10-minute prediction model

Contingency Table[a]				
Observed		Predicted		ROC = 0.819
		Traffic safety condition		Percentage Correct
		Crash	No-crash	
Traffic safety condition	Crash	41 (TP)	37 (FN)	52.56
	No-crash	24 (FP)	288 (TN)	92.31
Overall Percentage				84.36
a. The classification threshold value is 0.4				

Table 6: The contingency table of the 15-minute prediction model

The prediction performance of a binary model of outcome can be illustrated by means of a graphical plot which is called the receiver operating characteristic (ROC) curve (Olson and Delen, 2008). This graph is constructed by plotting the true positive rate (i.e. TP divided by total observed positives) against the false positive rate (i.e. FP divided by total observed negatives). Figure 5 illustrates the ROC curve for the final prediction model. Each instance or prediction outcome of the contingency table represents one point in the ROC space. The best possible prediction would yield a point in coordinate (0,1) of the ROC space (i.e. the upper left corner) which implies no FN and no FP. Hence, the larger the area under the ROC curve, the better the prediction accuracy. The area under the ROC curve for the 5-minute model is 0.853, greater in comparison with other model results, indicating an appropriate predictive performance of this model. Moreover, it is important to distinguish between different costs imposed by different false or true predictions. In many cases optimizing the classification rate without considering the cost of the errors often leads to misleading results. A very well-known example of this would be applicable in loan decisions. As it is the case for the current study, the cost of lending to a defaulter is far greater than the lost-business cost of refusing a loan to a non-defaulter (Witten et al., 2011). The same rule is applicable in the context of this study where the cost of having more false negatives is greater than false positives. The results reported in Tables 4-6 showed that the 5-minute model outperform other models by predicting less false negatives.

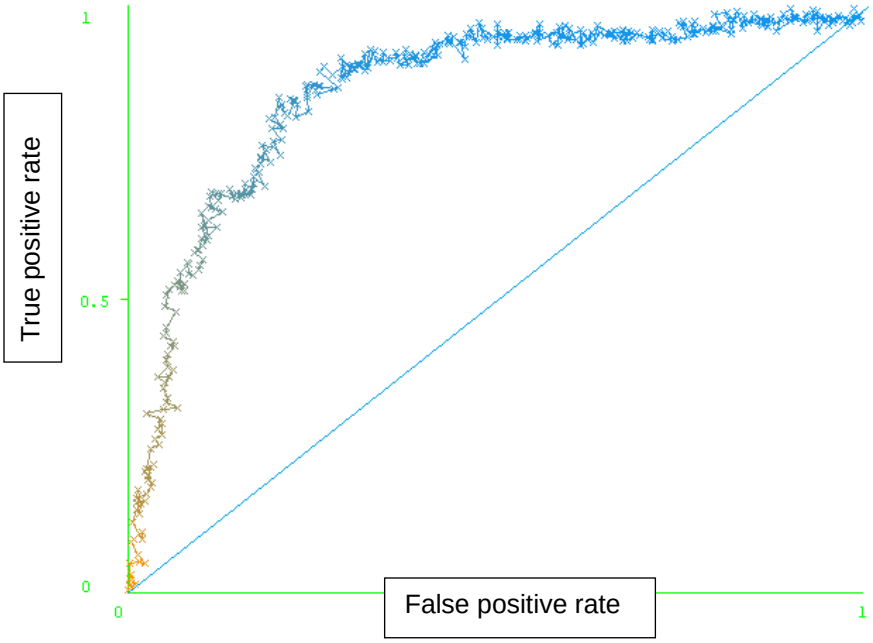


Figure 5: The ROC curve of the 5-minute prediction model

Expectedly the prediction accuracy of the model increases if the acceptable FP rate (also referred to as false alarm rate) also increases. However, the trade-off between the prediction accuracy of the model and the false alarm rate needs to be considered and an appropriate threshold should be set by traffic authorities. To this end, the classification threshold (indicated in Tables 4, 5 and 6) can be adjusted in order to deliver appropriate classification accuracy for both traffic safety conditions (i.e. no-crash and crash conditions). Depending on the application and the definition of dependent variable, this classification threshold can be increased or decreased to provide less false positive and false negatives. In the context of this research, it would be beneficial to decrease this threshold and conservatively predict more crash occasions (even if they are not observed as crash occasions) in order to stay on the safe side. In practice and in the case of predicting a crash condition, different countermeasures can be implemented (e.g. intelligent speed adaptation or variable speed limits) to avoid a potential crash occurrence. Although this crash condition might not be correctly predicted (i.e. it might not eventually lead to crash occurrence), implementation of safety countermeasures would be anyhow beneficial since it reduces traffic flow disturbance. With the classification threshold of 0.4, the false alarm rate of the 5-minute model is less than 10% which is significantly lower in comparison to the results of previous studies reported by Xu et al. (Xu et al., 2013).

5 Model results

Based on the discussion of the previous section, the 5-minute model is selected as the final model. As can be seen from the model output in Table 7, three predictor variables have significant association with traffic safety condition.

	Coefficient estimate	p-value	Odds ratio
Intercept	-4.3078	0.000	-
STDEV_SP_US1	0.1044	0.000	1.1101
OC_US1	0.0832	0.000	1.0868
Diff_SP_US1-DS1	0.0149	0.000	1.0151

Table 7: Coefficient estimates and odds ratios for 5-minute model

The final chosen model's formulation is shown in Eq.3.

$$ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = -4.3078 + 0.1044 \times \text{STDEV\_SP\_US1} + 0.0832 \times \text{OC\_US1} + 0.0149 \times \text{Diff\_SP\_US1} - \text{DS1} \quad (3)$$

Where $\pi(x)$ denotes the probability of experiencing crash or no-crash condition. All estimate signs are also in line with intuitive expectations. As can be seen from the results, standard deviation of speed upstream the crash location has a positive association with crash occurrence. This implies that higher standard deviation of speed at a location will potentially increase the risk of crash occurrence at a downstream location. Occupancy at the upstream loop detector has also a positive sign. This indicates that higher occupancy will also increase the likelihood of crash occurrence. Another significant variable is the difference in average speed of upstream and downstream loop detectors. This interesting result reveals the importance of speed and its derivative variables in crash likelihood prediction. This signifies that if the difference in average travel speed of two consecutive locations is becoming greater, there will be a higher probability of crash occurrence in between those two locations. Figures 6, 7 and 8 illustrate the distribution of traffic safety condition in association with the final chosen explanatory variables.

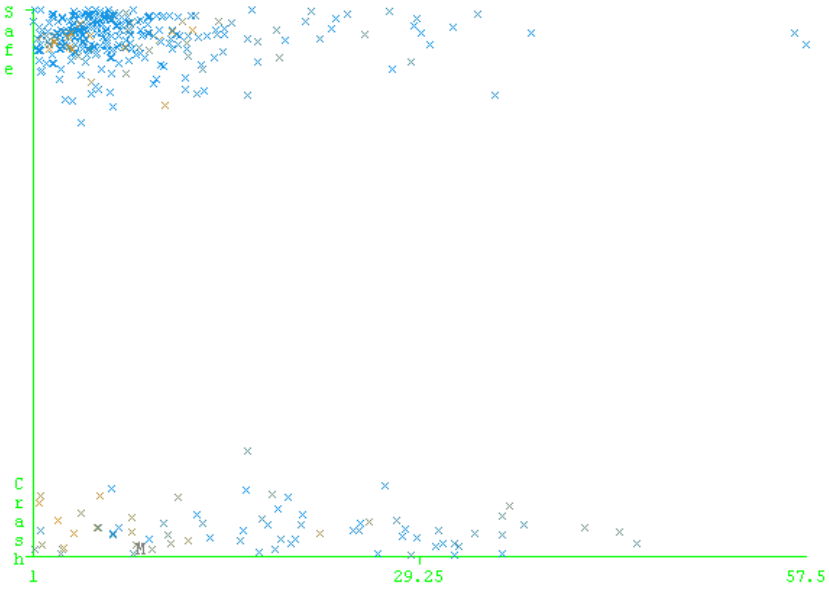


Figure 6: The distribution of traffic safety condition in association with variable STDEV_SP_US1

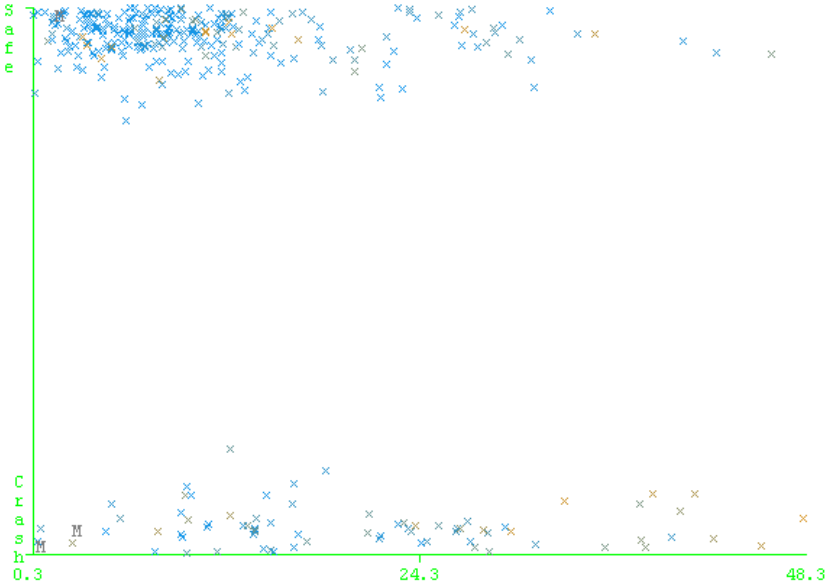


Figure 7: The distribution of traffic safety condition in association with variable OC_US1

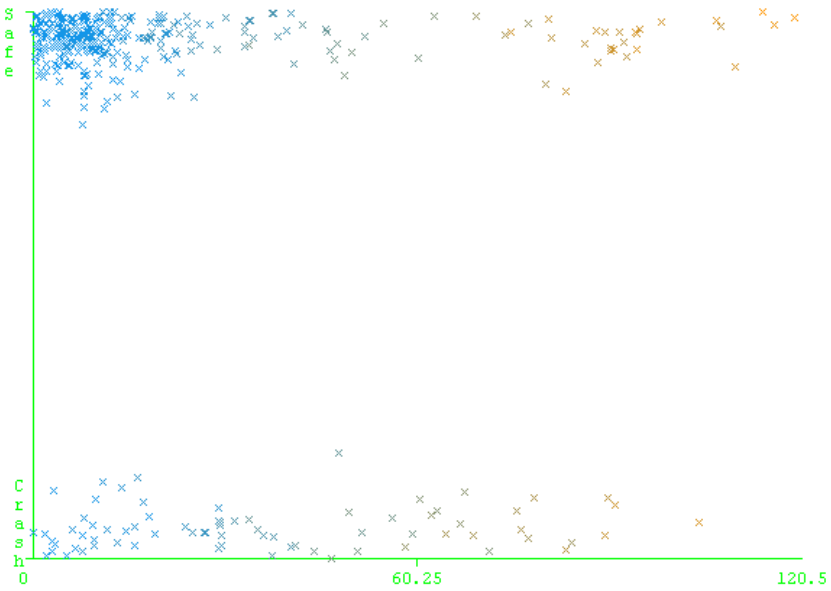


Figure 8: The distribution of traffic safety condition in association with variable Diff_SP_US1-DS1

6 Conclusions and Discussion

The main objective of this study was to explore the possibility of predicting traffic safety conditions on motorways by means of traffic flow characteristics collected by double loop detector stations. Various variables such as traffic volume, occupancy, average speed, standard deviation of speed, difference between average speeds on two consecutive loop detector stations were among the potential predictor variables that were considered for model development. The raw data were at 1-minute level of aggregation, which would potentially bias the results due to their random noise. To avoid this problem, the primary data were aggregated into three different levels, namely 5-minute, 10-minute and 15-minute intervals prior to crash occurrence time. This also enables us to identify the best level of aggregation that will result in better crash prediction accuracy. All of these three aggregation level data were used to develop individual prediction models by means of the binary logistic regression technique.

The results of analysis showed that the 5-minute model outperforms the other two models by means of more correctly predicted crash traffic conditions. The results showed that the 5-minute model was able to correctly predict more than 60% and 90% of crash and no-crash instances respectively. The false alarm rate (i.e. false positive in this study) resulted from the 5-minute model is less than 10%, significantly lower than false alarm rates reported in the literature. This low percentage of false alarm rate allows system users to decrease the classification threshold (i.e. currently set to be 0.4). By doing so, the number of true positive predictions (i.e. number of crash instances that are predicted correctly) will increase, meaning that the predictability of crash conditions will be improved. In return, decreasing the classification threshold yields to more no-crash occasions being incorrectly predicted as crash occasions. Depending on the application of safety management systems, this increase in false alarm rate does not impose any critical problem on road users. In application of some dynamic safety management systems (e.g. ramp metering or variable speed limits), road users are not aware of the reason behind a change in ramp metering rate or a temporary reduced speed limit. Therefore, a controlled decrease of the classification threshold will improve crash condition prediction accuracy. This trade-off between the prediction accuracy and the false alarm rate must be determined cautiously by traffic authorities based on their own specific preferences.

The first requirement in realization of a proactive highway safety management system is having accurate real-time prediction models. The performance of the developed prediction model in this study (i.e. the 5-minute model) appropriately fulfills this condition by predicting an acceptable rate of crash

conditions. Having said that, there is always room for improving the accuracy level of developed model by enriching the crash and traffic data (i.e. data collected by traffic surveillance cameras) and by employing other modeling techniques. This would improve model accuracy and robustness and subsequently would increase the acceptability of the prediction model by traffic authorities who are going to utilize this model in their dynamic safety management system. Another extension for future research would be the transferability check of the developed prediction model. To this end, the model should be validated against crash and traffic data collected from various motorways. This should be carried out to ensure that the final model is able to correctly predict traffic safety conditions on any motorway under the same jurisdiction and with the same infrastructural basis (e.g. speed limit, traffic volume order or geometric conditions).

7 Acknowledgements

The authors thank the Flemish Traffic Centre for their support during the data gathering and for their collaborated efforts in conducting the survey. The content of this paper is the sole responsibility of the authors.

References

Abdel-Aty, M., Dilmore, J., Dhindsa, A., 2006. Evaluation of variable speed limits for real-time freeway safety improvement. Accid. Anal. Prev. 38, 335–345. doi:10.1016/j.aap.2005.10.010

Abdel-Aty, M., Haleem, K., Cunningham, R., Gayah, V., n.d. Application of variable speed limits and ramp metering to improve safety and efficiency of freeways. Department of Civil, Env. & Construction Engineering University of Central Florida.

Abdel-Aty, M., Pande, A., 2005. Identifying crash propensity using specific traffic speed conditions. J. Safety Res. 36, 97–108. doi:10.1016/j.jsr.2004.11.002

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. 11, 107–120. doi:10.1080/15472450701410395

Abdel-Aty, M., Uddin, N., Pande, A., 2005. Split Models for Predicting Multivehicle Crashes During High-Speed and Low-Speed Operating Conditions on Freeways. Transp. Res. Rec. J. Transp. Res. Board 1908, 51–58. doi:10.3141/1908-07

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. doi:10.3141/1897-12

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. doi:10.1016/j.trc.2012.04.001

Ahmed, M., Abdel-Aty, M., 2012. The viability of using automatic vehicle identification data for real-time crash prediction. IEEE Trans. Intell. Transp. Syst. 13, 459–468.

Ahmed, M., Abdel-Aty, M., 2013. A data fusion framework for real-time risk assessment on freeways. Transp. Res. Part C Emerg. Technol. 26, 203–213. doi:10.1016/j.trc.2012.09.002

Ahmed, M., Abdel-Aty, M., yu, 2012. A Bayesian updating approach for real-time safety evaluation using AVI data, in: The 91st Annual Meeting of the Transportation Research Boar. Washington, DC., USA.

Carsten, O.M.J., Tate, F.N., 2005. Intelligent speed adaptation: accident savings and cost–benefit analysis. Accid. Anal. Prev. 37, 407–416. doi:10.1016/j.aap.2004.02.007

Chang, L.-Y., Chen, W.-C., 2005. Data mining of tree-based models to analyze freeway accident frequency. J. Safety Res. 36, 365–375. doi:10.1016/j.jsr.2005.06.013

Chen, G., Meckle, W., Wilson, J., 2002. Speed and safety effect of photo radar enforcement on a highway corridor in British Columbia. Accid. Anal. Prev. 34, 129–138. doi:10.1016/S0001-4575(01)00006-9

Golob, T.F., Recker, W., Pavlis, Y., 2008. Probabilistic models of freeway safety performance using traffic flow data as predictors. Saf. Sci. 46, 1306–1333. doi:10.1016/j.ssci.2007.08.007

Hosmer, Jr., D.W., Lemeshow, S., Sturdivant, R.X., 2013. Applied Logistic Regression. John Wiley & Sons.

Hourdos, J., Garg, V., Michalopoulos, P., Davis, G., 2006. Real-Time Detection of Crash-Prone Conditions at Freeway High-Crash Locations. Transp. Res. Rec. J. Transp. Res. Board 1968, 83–91. doi:10.3141/1968-10

Jo, Y., Yoon, K., Jung, I., 2012. Variable Speed Limit to Improve Safety near Traffic Congestion on Urban Freeways. Int. J. Fuzzy Syst. 14, 278–288.

Kutner, M.H., Nachtsheim, C.J., Neter, J., 2004. Applied Linear Regression Models, 4th ed. McGraw-Hill, Boston; New York.

Lai, F., Carsten, O., Tate, F., 2012. How much benefit does Intelligent Speed Adaptation deliver: An analysis of its potential contribution to safety and environment. Accid. Anal. Prev. 48, 63–72. doi:10.1016/j.aap.2011.04.011

Lee, C., Abdel-Aty, M., Hsia, L., 2006a. Potential Real-Time Indicators of Sideswipe Crashes on Freeways. Transp. Res. Rec. J. Transp. Res. Board 1953, 41–49. doi:10.3141/1953-05

Lee, C., Hellinga, B., Ozbay, K., 2006b. Quantifying effects of ramp metering on freeway safety. Accid. Anal. Prev. 38, 279–288. doi:10.1016/j.aap.2005.09.011

Lee, C., Hellinga, B., Saccomanno, F., 2003. Real-Time Crash Prediction Model for Application to Crash Prevention in Freeway Traffic. Transp. Res. Rec. J. Transp. Res. Board 1840, 67–77. doi:10.3141/1840-08

Lee, C., Hellinga, B., Saccomanno, F., 2004. Assessing Safety Benefits of Variable Speed Limits. Transp. Res. Rec. J. Transp. Res. Board 1897, 183–190. doi:10.3141/1897-24

Lee, C., Hellinga, B., Saccomanno, F., 2006c. Evaluation of variable speed limits to improve traffic safety. Transp. Res. Part C Emerg. Technol. 14, 213–228. doi:10.1016/j.trc.2006.06.002

Lee, C., Saccomanno, F., Hellinga, B., 2002. Analysis of Crash Precursors on Instrumented Freeways. Transp. Res. Rec. J. Transp. Res. Board 1784, 1–8. doi:10.3141/1784-01

Oh, C., Park, S., Ritchie, S.G., 2006. A method for identifying rear-end collision risks using inductive loop detectors. Accid. Anal. Prev. 38, 295–301. doi:10.1016/j.aap.2005.09.009

Oh, J.-S., Oh, C., Ritchie, S.G., Chang, M., 2005. Real-Time Estimation of Accident Likelihood for Safety Enhancement. J. Transp. Eng. 131, 358–363. doi:10.1061/(ASCE)0733-947X(2005)131:5(358)

Olson, D.L., Delen, D., 2008. Advanced data mining techniques. Springer, Berlin.

Pande, A., Abdel-Aty, M., 2006a. Comprehensive Analysis of the Relationship Between Real-Time Traffic Surveillance Data and Rear-End Crashes on Freeways. Transp. Res. Rec. J. Transp. Res. Board 1953, 31–40. doi:10.3141/1953-04

Pande, A., Abdel-Aty, M., 2006b. Assessment of freeway traffic parameters leading to lane-change related collisions. Accid. Anal. Prev. 38, 936–948. doi:10.1016/j.aap.2006.03.004

Pande, A., Das, A., Abdel-Aty, M., Hassan, H., 2011. Estimation of Real-Time Crash Risk. Transp. Res. Rec. J. Transp. Res. Board 2237, 60–66. doi:10.3141/2237-07

Servin, O., Boriboonsomsin, K., Barth, M.J., 2008. Preliminary Design of Speed Control Strategies in Dynamic Intelligent Speed Adaptation System for Freeways. Presented at the Transportation Research Board 87th Annual Meeting.

Witten, I.H., Frank, E., Hall, M.A., ITPro, 2011. Data mining practical machine learning tools and techniques, third edition. Morgan Kaufmann Publishers, Burlington, Mass.

Xu, C., Liu, P., Wang, W., Li, Z., 2012. Evaluation of the impacts of traffic states on crash risks on freeways. Accid. Anal. Prev. 47, 162–171. doi:10.1016/j.aap.2012.01.020

Xu, C., Tarko, A.P., Wang, W., Liu, P., 2013. Predicting crash likelihood and severity on freeways with real-time loop detector data. Accid. Anal. Prev. 57, 30–39. doi:10.1016/j.aap.2013.03.035

Zheng, Z., Ahn, S., Monsere, C.M., 2010. Impact of traffic oscillations on freeway crash occurrences. Accid. Anal. Prev. 42, 626–636. doi:10.1016/j.aap.2009.10.009

Het Steunpunt Verkeersveiligheid 2012-2015 is een samenwerkingsverband tussen de volgende partners: