INFLUENTIAL EVALUATION OF DATA SAMPLING TECHNIQUES ON ACCURACY OF MOTORWAY CRASH RISK ASSESSMENT MODELS

Minh-Hai Pham
PhD student, School of Architecture, Civil and Environmental Engineering,
Swiss Federal Institute of Technology in Lausanne,
EPFL-LAVOC, Station 18, 1015 Lausanne, Switzerland, e-mail: minhhai.pham@epfl.ch

André-Gilles Dumont
Professor, School of Architecture, Civil and Environmental Engineering,
Swiss Federal Institute of Technology in Lausanne,
EPFL-LAVOC, Station 18, 1015 Lausanne, Switzerland, e-mail: andre-gilles.dumont@epfl.ch

Submitted to the 3rd International Conference on Road Safety and Simulation, September 14-16, 2011, Indianapolis, USA

ABSTRACT

Recently, many studies have been focusing on real-time detection of rear-end and sideswipe crash risks on motorways thanks to the availability of traffic data provided by traffic detectors and crash record databases. In these studies, traffic evolution leading to individual crashes called *pre-crash cases* is considered and differentiated with traffic conditions where there is no crash recorded, called *non-crash cases*. This trend of studies reflects the need of identifying traffic crash risk in real-time in order that appropriate countermeasures could be implemented to prevent the risk from further developing and ending up with a crash. These studies are called *disaggregate studies* as the units of analysis are crashes themselves, according to (Golob et al., 2004).

However, one of issues for these studies is the imbalance between pre-crash and non-crash cases because crashes are rare events on motorways and there should be certain traffic conditions for rear-end and sideswipe crashes to occur. Therefore, it is important to choose appropriate non-crash cases to compare with pre-crash cases, which is usually neglected in previous disaggregate studies. In the present paper, four different techniques for sampling non-crash cases is reviewed and compared using individual vehicle traffic data and crash databases altogether available from 2003 to 2007 on Swiss motorways A1.

Keywords: traffic safety, safety indicators, real-time, traffic individual data, accident data, non-crash data sampling

BACKGROUND & STUDY SITE

Background

Recently, more studies such as the ones presented in (Abdel-Aty et al., 2008; Hossain and Muromachi, 2010; Hourdakis et al., 2006; Lee et al., 2003; Oh et al., 2001; Pande and Abdel-Aty, 2007; Pham et al., 2011) focus on detecting in real-time the risk of rear-end and sideswipe crashes on motorways thanks to the availability of traffic data recorded by traffic detectors and crash record databases. In these studies, traffic evolution leading to individual crashes, called *pre-crash cases* is considered and differentiated with traffic conditions where there is no crash, called *non-crash cases*. The models developed to differentiate between pre-crash and non-crash cases are called *risk assessment models* as the models aim to classify whether a traffic case that has occurred is pre-crash or non-crash. This trend of studies reflects the need of identifying traffic crash risk in real-time in order that appropriate countermeasures could be implemented to prevent the risk from further developing and ending up with a crash. These studies are also different from the incident detection studies that attempt to detect traffic incidents right after they occur to provide necessary healthcare, urgent services and to avoid secondary incidents.

Pre-crash cases are compared with some of non-crash cases which are selected according to certain criteria. Several methodologies were proposed to develop models aiming to differentiate between pre-crash and non-crash cases. On the one hand, the outcome of model development is to understand the causality of pre-crash cases and thereafter, the causality of crashes. On the other hand, the developed models can be used to evaluate the risk status of new traffic conditions. In real-time, new traffic conditions are the traffic evolution during the last time interval. In model development process, new traffic conditions are represented by validation data sets. In most of the previous studies, the reported performance of developed models for the validation data sets is relatively high. However, the test to those models cannot be replicated within the present study due to the difference of data sets and approaches.

Crashes are rare events on motorways and there should be certain traffic conditions for rear-end and sideswipe crashes to occur. The selection of non-crash cases to be compared with pre-crash cases might have impact on the performance of risk assessment models. Therefore, the objective of current study is to verify how non-crash data sampling can influence the accuracy of such models. Here, non-crash data sampling techniques from literature are reviewed and compared. There are mainly five techniques for sampling non-crash cases applied in the previous studies which are:

- Technique 1: non-crash data are taken for 5 minutes at 30 minutes before crashes. Pre-crash data are taken at 5 minutes right before crashes. For each pre-crash case, there is one corresponding non-crash case. This technique is applied by Oh et al., (2001).
- Technique 2: use of matched-case control. For each pre-crash case, take 5 non-crash cases at the same time of the day, day of the week, under the same weather conditions. This technique is applied by Lee et al., (2003) and Pande and Abdel-Aty, (2007).
- Technique 3: random selection. For each pre-crash case, select 5 non-crash cases at random. This technique is applied by Abdel-Aty and Pande, (2005).
- Technique 4: Develop models using all available non-crash cases. This technique is applied by Hossain and Muromachi, (2010).

- Technique 5: cluster all non-crash cases into clusters then classify pre-crash cases into obtained clusters such that non-crash cases are matched with pre-crash cases. This technique is applied by Pham et al., (2011).

The five non-crash data sampling techniques are tested and evaluated using traffic data, meteorological data and crash databases altogether available from 2002 to 2007 on Swiss motorways A1.

Study Site

A study site for the present study is selected on Swiss motorway A1 between two cities Bern and Zurich. The main criteria for study site selection is the simultaneous availability of individual vehicle data from double loop traffic detectors, crash records around the location of traffic detectors and meteorological data. The selected study site is presented in Figure 1. At the study site, there are two lanes per traffic direction.

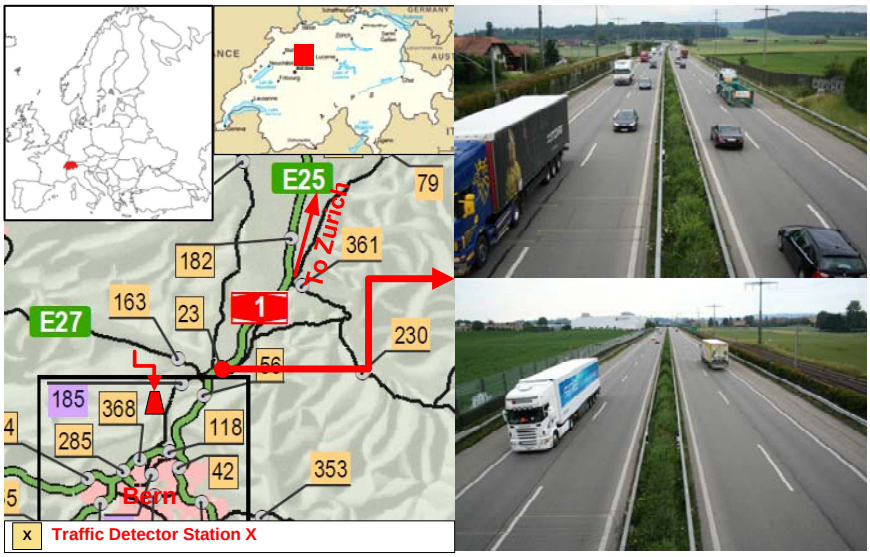

Figure 1: Study site

The double loop detectors at the study site provide individual vehicle data as presented in Table 1. Important data fields include time gap (column Gap), speed (Speed), the length of vehicles (Length), and the class of vehicles (V. Class).

Table 1: Individual vehicle data sample

Index	Date	HHMM	Sec	ms	Reserved	Lane	Dir	Hw	Gap	Speed	Length	V. Class
023198	150303	0001	21	30	000000	1	1	43.6	43.5	120	467	2
023199	150303	0001	38	42	000000	1	1	17.1	16.9	125	989	3
023200	150303	0001	47	12	000000	4	1	46.9	46.8	113	428	2
023201	150303	0001	50	58	000000	4	1	3.4	3.3	119	423	2

METHODOLOGY

Comparison setting

Here, 5-minute intervals are used for aggregating data. In most of the previous studies, 5-minute aggregation intervals are also used. To define pre-crash and non-crash cases, variables listed in Table 2 are used. The variables specify 5 different views of a traffic case, including the moment where the traffic case occurs, the state of traffic on each lane (there are two lanes each direction), the difference between two lanes, the traffic evolution from the last traffic case, and the meteorological information.

Table 2: List of variables

Variable	Alias	Explanation	Specification
X1	TDay	Time of the Day	Instantaneity
X2	WDay	Day of the Week	
X3	**L**Flow	Right Lane's Flow	Status of Right Lane (Prefix **L**)
X4	**L**ASpd	Right Lane's Average Speed	
X5	**L**AHw	Right Lane's Average Headway	
X6	**L**Occ	Right Lane's Occupancy	
X7	**L**VHw	Right Lane's Headway Variation	
X8	**L**VSpd	Right Lane's Speed Variation	
X9	**L**% HV	Right Lane's Percentage of Heavy Vehicles	
X10	**H**Flow	Left Lane's Flow	Status of Left Lane (Prefix **H**)
X11	**H**ASpd	Left Lane's Average Speed	
X12	**H**AHw	Left Lane's Average Headway	
X13	**H**Occ	Left Lane's Occupancy	
X14	**H**VHw	Left Lane's Headway Variation	
X15	**H**VSpd	Left Lane's Speed Variation	
X16	**H**%HV	Left Lane's Percentage of Heavy Vehicles	
X17	Spd#	Speed Difference between two lanes (HASpd-LASpd)	Difference between 2 lanes
X18	**L**FCg	Flow change on Right Lane (compared to the previous TS)	Traffic evolution
X19	**L**SCg	Speed change on Right Lane (compared to the previous TS)	
X20	**H**FCg	Flow change on Left Lane (compared to the previous TS)	
X21	**H**SCg	Speed change on Left Lane (compared to the previous TS)	
X22	Prec	Precipitation	Meteorological Information

Based on crash data, pre-crash and non-crash cases are defined as illustrated in Figure 2. For a crash, there is a *crash period* containing traffic evolution before and after the crash. The crash period is divided into three smaller parts, namely, *pre-crash buffer period*, *pre-crash period*, and *post-crash period*. Pre-crash cases are traffic cases occurring within pre-crash period. Non-crash cases are traffic cases occurring outside of crash periods for all crashes. Traffic cases occurring within post-crash periods and pre-crash buffer periods are not considered in this study. It's worth noting that pre-crash cases in the present study are extracted from pre-crash periods of sideswipe and rear-end crashes.

Here the duration of pre-crash period and pre-crash buffer period is 30 minutes, i.e. there are six pre-crash cases before each crash. The duration for post-crash period is 210 minutes as the results of an analysis on the influence of crashes on traffic conditions.

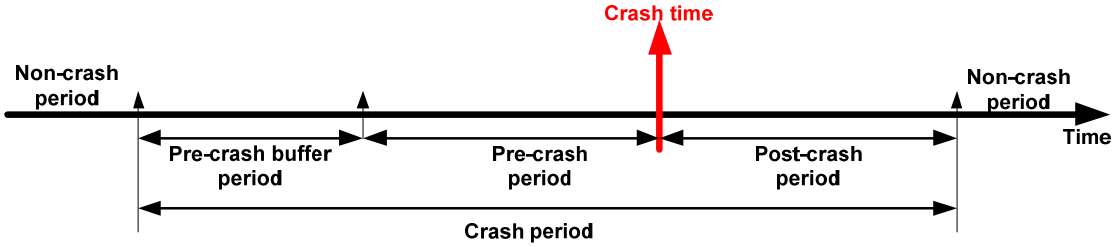

Figure 2: Definition of pre-crash and non-crash cases

Framework

The framework for comparing pre-crash and non-crash cases is illustrated in Figure 3. Firstly, non-crash cases are sampled using the sampling techniques presented in the background section. Thereafter, there pre-crash cases together with the sampled non-crash cases are used for developing risk identification models that aim at differentiating pre-crash and non-crash cases. The models are developed using Random Forest regression (Breiman, 2001). Finally, the performance of developed models in identifying pre-crash and non-crash cases in validation data sets is compared.

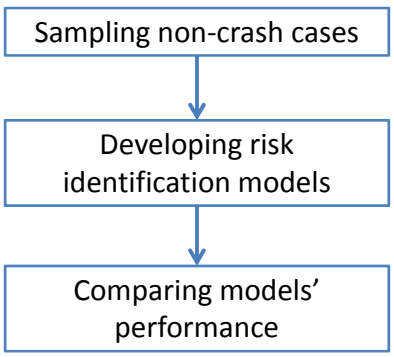

Figure 3: Comparison framework

In the framework illustrated in Figure 3, the only difference between developed models is the techniques for sampling non-crash data. Therefore, the final performance of these models should reflect the influence of sampling techniques.

Development of Risk Assessment Models

Risk assessment models are developed using Random Forest Regression - RFR (Breiman, 2001) with RFR output of 1 (or 0) representing pre-crash (or non-crash). Once the regression model is developed, it outputs the probability for an input, which is a traffic case, to be pre-crash (ranging from 0.0 to 1.0). After sampling non-crash cases, the working data set contains all pre-crash cases and sampled non-crash cases. The working data set is then divided into three smaller data sets: training; calibration; and validation data sets with the ratio of 6:2:2, respectively. It is worth mentioning that the distribution of pre-crash and non-crash cases into three smaller data sets is the same (i.e. the ratio of 6:2:2, respectively). The training data set is used to form the regression relationship between pre-crash and non-crash cases. After that, it is necessary to determine a probability threshold to classify regression outputs into pre-crash or non-crash. The calibration data set is then used to determine that probability threshold. With developed regression model and the calibrated probability threshold, validation data set is tested to verify the performance of the model on new data. As data in validation data set are not used for training regression models or calibrating probability thresholds, the models' performance against the validation data set reflects the capacity of the models in predicting new traffic cases.

Three rules called *Threshold Rules* are applied for determining the probability threshold:
1) At least 70% of pre-crash cases are correctly identified by regression models. This rule conforms to study's objective: assessing traffic risks. This also indicates the maximum possible missing rate of 30%.
2) At least 70% of non-crash cases are correctly identified by regression models. This is to guarantee that the model is not trivial.
3) Among the thresholds satisfying two rules above, choose the threshold maximizing the sum of percentages of non-crash and pre-crash cases correctly identified.

In the Threshold Rules, the threshold of 70% is decided based on the fact that in most of previous studies, the reported accuracy of risk assessment models is less than 70%.

By validating developed models, validation data set including both non-crash and pre-crash cases is input into models. Besides, for several non-crash sampling techniques, there are non-crash data unused. These data are also input into developed models.

RESULTS AND ANALYSIS

Sampling Non-Crash Cases

With a total of 120 crashes during six years from 2002 to 2007 considered in this study, there are totally 720 pre-crash cases whereas; the number of non-crash cases is 1'160'834. Therefore, the imbalance between non-crash and pre-crash cases is high (1'612.27 vs. 1, respectively). Table 3 summarizes the population of non-crash cases corresponding to the application of five different non-crash sampling techniques. Among fives techniques, the ratio between non-crash and pre-crash cases resulted by technique 1 is lowest (1:1) whereas; the ratio resulted by technique 4 is

highest. Techniques 2 and 3 result in the same ratio of 5:1 yet the search space of non-crash cases used by technique 2 is reduced due to the matched case control.

Table 3: Results of sampling techniques

Sampling techniques	Notes	Pre-crash cases	Non-crash cases	Pre-crash / Non-crash ratio
Technique 1	One non-crash, one pre-crash	720	720	1:1
Technique 2	Matched case control	720	2600	1:5
Technique 3	Random selection	720	2600	1:5
Technique 4	All non-crash cases are used	720	1'160'834	1:1612.27
Technique 5	Clustering-Classification scheme	720	Varies	Between 1:150 and 1:800

In Table 3, the ratio between non-crash and pre-crash cases given by Technique 5 varies because non-crash cases sampled by Technique 5 are clustered into traffic regimes (i.e. groups of non-crash cases) before pre-crash cases are classified into the traffic regimes. Therefore, under each traffic regime, there is a set of non-crash cases and a set of pre-crash cases. The ratio between pre-crash and non-crash cases is then traffic regime – specific. Traffic regimes where no pre-crash case is classified into are not considered here as in the future, if a new traffic case is classified into those traffic regimes, the traffic case is automatically declared as non-crash case.

Comparison of Models' Performance

After sampling non-crash data, the working data set is divided into three data sets for training, calibration, and validation. The performance of the developed models with three data sets is introduced in Table 4. Although non-crash cases have been sampled, the number of pre-crash cases is still smaller compared to the number of sampled non-crash cases (except for Technique 1). Therefore, the performance of developed models is based on non-crash and pre-crash cases identified within each data set. As there is a proportion of non-crash data unused for developing models for non-crash sampling techniques 1, 2, and 3, the unused non-crash data are also tested with the models developed using techniques 1, 2, and 3.

Table 4: Performance of models developed based on different non-crash sampling techniques (percentage of cases correctly identified)

Non-crash sampling technique	Training data set		Calibration data set		Validation data set		Unused non-crash data
	Non-crash	Pre-crash	Non-crash	Pre-crash	Non-crash	Pre-crash	
1	81.94	23.61	45.83	45.83	34.67	58.33	54.12
2	73.15	68.14	70.14	71.35	56.25	61.15	57.35
3	67.36	63.27	61.81	51.73	53.47	43.65	56.25
4	92.81	03.70	91.62	03.47	91.67	03.42	-
5	95.67	100.00	91.93	90.77	89.83	83.62	-

Results presented in Table 4 show that the model developed using Technique 5 performs better with validation data set than models developed using other non-crash sampling techniques. Model 1 that is developed using Technique 1 is vulnerable when new data are tested and the

performance of the model is low. Firstly, the equal population of non-crash and pre-crash cases used by Technique 1 does not reflect well the fact that crashes are rare events. That is why when Model 1 is tested with unused non-crash data, the performance of Model 1 is low as it can easily classify non-crash cases into pre-crash. Secondly, the fact that non-crash data are extracted at 30 minutes before crashes might mislead the model as those non-crash cases can be pre-crash cases in reality.

Similarly to Model 1, Models 2 and 3 are developed by using Techniques 2 and 3, respectively, and do not take into account all traffic data. As results, the performance of Models 2 and 3 is also low with unused non-crash data. The applicability of Models 1, 2, and 3 in real-time is questionable because when a new traffic case is tested by these models, the chance for that traffic case to be non-crash or pre-crash is likely to be equal.

Models 4 and 5 developed using Techniques 4 and 5, respectively, make use of all possible traffic cases. Therefore, there is no unused traffic case left to be tested. However, the significant difference between Model 4 and Model 5 is their performance: Model 4 performs badly with pre-crash data whereas; the performance of Model 5 is good for both pre-crash and non-crash cases. Model 4 reflects very well the imbalance between pre-crash and non-crash cases and when a new traffic case is input into Model 4, the chance for it to be identified as non-crash is high. As presented in Table 4, most of pre-crash cases are wrongly identified as non-crash.

Model 5 might be the most appropriate for this problem as all traffic cases are considered and there is a clustering-classification procedure used to match non-crash with pre-crash cases. Non-crash cases are clustered into traffic regimes and pre-crash cases are classified into the obtained traffic regimes. Under each traffic regime, there is a model for differentiating pre-crash cases with non-crash cases. There are also traffic regimes, called *low risk traffic regimes*, where there is no pre-crash case classified into. This is because traffic conditions under those traffic regimes are not favorable for the crash types under consideration which consist of rear-end and sideswipe crashes. New traffic cases, before being classified by traffic regime – based models, are classified into one of existing traffic regimes. If new traffic cases are classified into low risk traffic regimes, the traffic cases can be declared non-crash. Otherwise, the new traffic cases are classified by traffic regime – based models.

CONCLUSIONS

This paper addresses on an important methodological step of disaggregate traffic safety studies using traffic flow data: non-crash data sampling. The motivation for this discussion comes from the fact that crashes are rare events and non-crash cases need to be selected to be compared with pre-crash cases to improve the overall performance of real-time risk assessment models.

Most of existing studies do not sufficiently pay attention to the selection of non-crash cases. This usually leads to two issues:

1) The developed models propose low accuracy in identifying non-crash and/or pre-crash cases in validation data sets. Therefore, the performance of the models is low when applied in real-time with new traffic cases. This happens with Models 1, 2, 3, and 4.

2) The applicability of the developed models in real-time is questionable as not all non-crash cases are used to develop the models. Given a new traffic case, it is unknown whether the new traffic case can be input to the models. This happens with Models 1, 2, and 3.

Among five models considered in this paper, Model 5 with clustering-classification procedure for sampling non-crash cases seems to address well these two issues.

Therefore, for an aggregate traffic safety study, non-crash sampling is an important step and is not ignorable in developing risk assessment models.

ACKNOWLEDGEMENTS

In this study, traffic data are provided by Federal Roads Office, crash data are provided by Federal Statistics Office, and meteorological data are provided by Federal Office of Meteorology and Climatology.

This study is a part of project “Fusion of Safety INdicators” (FUSAIN), funded by Swiss Federal Roads Office (FEDRO).

REFERENCES

Abdel-Aty, M., Pande, A., 2005. Identifying crash propensity using specific traffic speed conditions. *Journal of Safety Research* 36, 97-108.

Abdel-Aty, M., Pande, A., Das, A., Knibbe, W., 2008. Assessing Safety on Dutch Freeways with Data from Infrastructure-Based Intelligent Transportation Systems. *Transportation Research Record: Journal of the Transportation Research Board* 2083, 153-161.

Breiman, L., 2001. Random Forests. *Machine Learning* 45, 5-32.

Golob, T.F., Recker, W.W., Alvarez, V.M., 2004. Freeway safety as a function of traffic flow. *Accident Analysis & Prevention* 36, 933-946.

Hossain, M., Muromachi, Y., 2010. Evaluating Location of Placement and Spacing of Detectors for Real-Time Crash Prediction on Urban Expressways, *89th TRB Annual meeting*, Washington DC.

Hourdakis, J., Garg, V., Michalopoulos, P., Davis, G., 2006. Real-Time Detection of Crash-Prone Conditions at Freeway High-Crash Locations. *Transportation Research Record: Journal of the Transportation Research Board* 1968, 83-91.

Lee, C., Hellinga, B., Saccomanno, F., 2003. *Real-time crash prediction model for application to crash prevention in freeway traffic*. National Research Council, Washington, DC, United States.

Oh, C., Oh, J.-S., Ritchie, S.G., Chang, M., 2001. Real-time Estimation of Freeway Accident Likelihood, *80th Annual Meeting of the Transportation Research Board, Washington, D.C., 2001*.

Pande, A., Abdel-Aty, M., 2007. Multiple-Model Framework for Assessment of Real-Time Crash Risk. *Transportation Research Record: Journal of the Transportation Research Board* 2019, 99-107.

Pham, M.-H., Bhaskar, A., Chung, E., Dumont, A.-G., 2011. Methodology for Developing Real-time Motorway Traffic Risk Identification Models Using Individual Vehicle Data, *90th Transportation Research Board annual meeting*, Washington DC.