Using Data Mining Techniques to Understand Collision Processes

Nicolas Saunier, Assistant Professor,
Department of Civil, Geological and Mining Engineering, École Polytechnique de Montréal
Nadia Mourji, Graduate Student
Department of Mathematics and Industrial Engineering, École Polytechnique de Montréal
Bruno Agard, Associate Professor
Department of Mathematics and Industrial Engineering, École Polytechnique de Montréal

ABSTRACT

In order to improve road safety, it is necessary to better understand collision processes, i.e. the chains of events that lead to collisions. Among the most important benefits, more efficient countermeasures can be found to target causes and factors known to lead to collisions. This would also help develop more reliable surrogate safety measures based on traffic events without a collision that have stronger links to collisions. This paper reports on the first phase of a project relying on microscopic data extracted from video sensors and on data mining techniques to identify patterns in a dataset of traffic events with and without a collision. This approach is demonstrated on a dataset collected in Kentucky of 295 traffic events, constituted of 213 conflicts and 82 collisions. Using the k-medoid algorithm, its clustering yields three groups with distinct characteristics, especially related to speed and the type, or lack, of evasive action. The most important attributes that determine the traffic event outcome (collision or not) are identified through a logit model.

RESUME

Dans le but d'améliorer la sécurité routière, il est nécessaire de mieux comprendre les processus de collision, i.e. les chaînes d'événements qui mènent à la collision. Parmi les bénéfices, de meilleures mesures pourraient être prises pour cibler les causes et les facteurs connus pour mener à des collisions. Cela aiderait aussi à développer des mesures substitutives de sécurité basées sur des événements sans collision dont les liens à la collision sont connus. Cet article décrit la première phase d'un projet reposant sur des données microscopiques extraites de capteurs vidéo et sur des méthodes de fouille de données afin d'identifier des régularités dans un ensemble d'événements de la circulation avec et sans collision. Cette approche est appliquée à un ensemble de données collectées

au Kentucky comprenant 295 événements, dont 213 conflits et 82 collisions. Leur classification non-supervisée à l'aide de l'algorithme des k-médoïdes produit 3 groupes avec des caractéristiques distinctes reliées à la vitesse et au type, ou à l'absence, de manœuvre d'évitement. Les attributs les plus importants pour déterminer le résultat d'un événement (collision ou pas) sont identifiés à l'aide d'un modèle logit.

INTRODUCTION

It is difficult to overstate the terrible cost of road collisions all over the world, and in particular in developing countries where the toll is expected to continue rising for the coming decades (1). While many general factors are known to decrease safety, i.e. increase the probability of a road collision (e.g. drunk driving) and the severity of collision outcomes (e.g. speed, lack of use of vehicle safety features such as safety belts, children seats), the actual processes that lead to collisions are not well known in details. The main reason is that collisions and the chains of events that lead to them (often called pre-crash events) are rarely observed.

Recent advanced data collection techniques may help in that regard. Video analysis in particular has generated a lot of interest for transportation applications: rich data, from the macroscopic variables flow, density and speeds to complete road users' trajectories, can thus be automatically extracted over large areas. Such a tool was developed at the University of British Columbia (2) and used to compute surrogate measures of safety such a more robust time to collision (3) and applied to various studies, including pedestrian-vehicle interactions and the before and after study of a pedestrian scramble phase (4). However, very large amounts of data may be generated from video data, in particular all road users' trajectories. Data mining provides the tools to efficiently explore, interpret and extract knowledge from large amounts of data.

This work relies on the assumption of the existence of a safety hierarchy (5), i.e. a framework that places all traffic events on a continuum with collisions at the top, undisturbed passages or "safe traffic events'" at the bottom and traffic conflicts in between. The position of a traffic event in the safety hierarchy measures its proximity to a potential collision. Significant effort has been invested to develop techniques to collect and link to collisions the specific class of the most severe traffic conflicts. It is believed that the observation of all traffic events can provide a complementary safety diagnosis, more complete than can be done using collision data alone. It is in particular a way to gain more knowledge about the factors and processes that lead to collisions.

The objective of this work is to better understand collision processes using microscopic road user data (trajectories). The success of this research would yield many benefits:

- more efficient countermeasures could be found to target causes and factors known to lead to collisions;
- more reliable surrogate safety measures could be developed based on traffic events without a collision that have stronger links to collisions.

The second point is critical as (6) suggests, on a small set of traffic events, that the evasive actions undertaken by road users involved in conflicts may be of a different nature than the ones attempted in collisions.

This paper presents part of the work done in a larger research project that aims to better understand collision processes and the relationship of interactions with and without a collision. This work relies on a large set of traffic events composed of conflicts and collisions and various statistical and data mining techniques to identify patterns in the dataset and classify interactions. To the authors' knowledge, this work is unique in the size of the analyzed dataset and the actual observation of safety-related events. The organization of the rest of the paper is as follows: related work, proposed approach, dataset description, results and conclusion.

RELATED WORK

There has been a considerable amount of research to estimate safety models as a function of explanatory variables describing the transportation system: the road, the vehicle and the driver. These safety models, also called a safety performance functions (SPF), typically take the form of an equation linking the expected number of collisions to a set of variables and rely on historical collision data. These models are at the core of the recently published Highway Safety Manual (HSM).

Historical collision data obtained from insurance and police reports is ill-suited for the analysis of collision processes (4). Other methods are required: in-depth accident analysis and naturalistic driving studies may help to better understand collision factors and processes. In-depth accident analysis relies on detailed reconstitutions to investigate collision factors (7) and as such may provide some information on the chain of events that led to the collision. However, they share many shortcomings with methods based on historical collision data: they provide only limited amounts of data, at a higher cost, they rely on reconstitutions in which the collision processes may be only guessed at and they still require to wait for collisions to occur. Naturalistic driving studies rely on the continuous collection of data on a road user, his driving behaviour, the vehicle and the environment, over extended periods of time (8). Very large projects, e.g. in the Strategic Highway Research Program 2 Safety research area (9), are in the making and should provide unprecedented information. An advantage will be the observation of all traffic events, not only collisions. Nevertheless, naturalistic driving studies also have limitations: they typically provide detailed information only on one of the road users involved in a safety-related event; vehicle instrumentation is costly and requires access to the vehicle, while fixed video cameras provide external non-intrusive monitoring of all traffic events and their context at a lower cost.

There has been a strong renewed interest in proactive methods for road safety analysis (10), the most famous being the traffic conflict technique (5) (11). Although mixed validation results, issues of cost and reliability have hindered their development, they have been integrated into traditional approaches, including the HSM, providing complementary information and alternative methods. The framework of the safety hierarchy, developed in the context of traffic conflict studies, is the basis for more recent approaches that take into account all road users' interactions, not only the most severe traffic conflicts, for more complete and robust diagnoses (3) (5). To understand collision processes, it is necessary to

expand the approach and use automated data collection techniques to provide sufficiently large amounts and objective microscopic data.

Studying collision processes can be helped by good classifications, as they are based on some form of similarity measure between collisions and the processes that lead to them. One of the best-known classification studies was published in 1993 with the goal of helping the evaluation of collision avoidance strategies (12). In a more recent work (13), the authors recommend to “use relatively homogenous class of accidents, all involving the same manoeuvre or ‘accident mechanism'” to study behavioural factors in road collisions.

Machine learning models, like artificial neural networks (ANN) and support vector machines (SVM) have been widely applied to estimate SPFs. This project however requires extracting patterns from data and can be achieved through data mining techniques (14). These include classification, using for example decision trees that can be interpreted, as opposed to the “black box'” nature of ANNs and SVMs, and clustering, i.e. finding groups through some similarity measure, using for example the k-means algorithm. Data mining has been used for the analysis of databases containing only collisions, without any microscopic data, and the readers are referred to (15) for a review. Despite significant use of data mining techniques to analyze collision data, it is apparent that the lack of microscopic data describing traffic events with and without a collision limits the scope of the collision factors that can be identified and the analysis of the similarities of traffic events of different severities.

PROPOSED APPROACH

Preliminary work has already been done to cluster interactions with and without a collision (15). Interactions between pairs of vehicles are described by indicators based on the two vehicles' speeds and speed differentials computed from the road users' trajectories, the type of evasive action and other vehicle and contextual information (see Table 1 and Table 2). The interaction dataset described below was mined for patterns using two well-known data mining techniques: decision trees (namely the C4.5 algorithm) and the k-means algorithm (14). Association rules were tried, but did not yield any strong result.

A decision tree was learnt to predict the outcome of the interaction (collision or not) from the other interaction attributes. The result highlighted the importance of evasive actions in the interaction outcome (more than 90 % of interactions where no evasive action was attempted resulted in a collision), as well as speed differential for interactions where one of the road users braked.

The number of clusters in the dataset was obtained using sequentially the k-means algorithm and a hierarchical agglomerative clustering (HAC) method. The “best” number of cluster is given from the HAC dendogram where the biggest step is observed. Three clusters were found to better segment the data. The dataset of the interactions was then partitioned in three clusters using only the speed information as a rough proxy for the road users’ relative interacting movements, with the k-mean method. Two types of clusters were obtained:

- two mixed clusters of similar interactions with and without a collision,
- one pure cluster with very few collisions in the cluster.

The underlying assumption is that interactions without a collision in mixed clusters can be used as surrogates for the collisions in the same cluster, while interactions without a collision

in the pure cluster cannot be used as surrogates to any type of collision (the last possibility, a pure cluster with no or very few interactions without a collision, would indicate that no interaction should be used as surrogates to the collisions in the cluster). In the analyzed dataset, the conflicts with the lowest speeds do not seem similar enough to many collisions and therefore should not be used for surrogate safety analysis. The two mixed clusters have distinct characteristics, in particular regarding the interaction categories (while this information was not used for clustering).

The present study includes more detailled analysis about interactions (see Figure 1). The work described in this paper relies on the use of all interaction attributes, not only speed information, to build a classification of all interactions and identify factors that may favour collisions. The distance used to compare interactions is also enhanced to better take into account various levels of similarities of the interaction categories (e.g. two same direction interactions, say rear-end and lane change, are different but not as dissimilar as a same direction and a side interaction, and will have an intermediate distance). Also the k-medoid clustering algorithm is now employed: it is a variation on the k-means algorithm where the cluster centroids are constrained to actual elements in the dataset. Following on the previous work, and after running similar tests to determine the number of clusters, the same number (3) is used. The three clusters produced by the k-medoid algorithm are finally analyzed and described. A more traditional statistical technique, a logit model, is also used to identify the most important attributes that determine the interaction outcome.

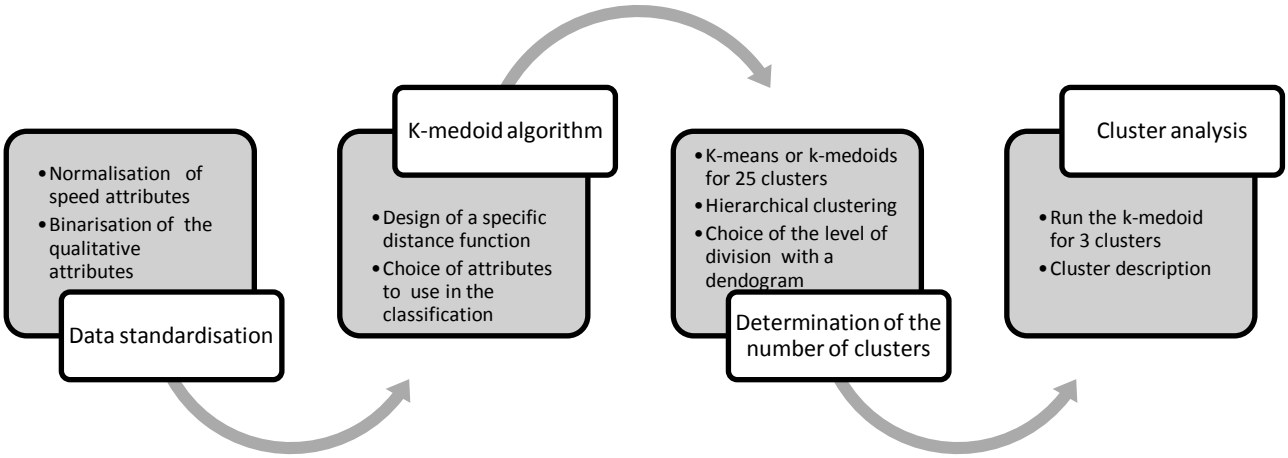

Figure 1 Methodology

DATASET DESCRIPTION

This work relies on the same dataset of interactions with and without a collision that was used in (3) and (15). The dataset was collected at a single signalized intersection in Kentucky and consists of 311 traffic events, conflicts and collisions. Video recordings were kept for a few seconds before and after the sound-based automatic detection of an interaction of interest. Unfortunately, some information is missing about this dataset, in particular the exact operational definition of conflicts. The existence of an interaction or its severity is not always obvious in some videos. Also regular traffic conditions (without conflicts or collisions) are not present. After some clean-up, 295 traffic events remain, split between 213 conflicts and 82 collisions. The interactions recorded in this dataset involve only motorized vehicles. Finally, the quality of the extracted trajectory information is limited by the

quality of the video data: limited resolution, compression artefacts, weather and lighting conditions. Despite these challenges, the microscopic data of the road users involved in the traffic events could be extracted, as well as other vehicle and contextual information: the attributes describing the interactions are described in Table 1 and Table 2, as well as the interaction categories in Figure 2. Most of the information was manually obtained, except for the day and microscopic data.

Categorical Attributes	**Values**
Type of day	weekday, week end
Lighting condition	daytime, twilight, night-time
Weather condition	normal, rain, snow
Interaction category	same direction (turning left and right, rear-end, lane change), opposite direction (turning left and right, head-on), side (turning left and right, straight) (see Figure 2)
Interaction outcome	conflict, collision

Table 1 Categorical interaction attributes

Numerical Attributes	**Units**
Road user type passenger car - van, 4x4, SUV - bus - truck (all sizes) - motorcycle - bike - pedestrian	number of road users per type
Type of evasive action No evasive action - Braking - Swerving - Acceleration	number of evasive actions per evasive action
Road user origin 4 street origins	number or road users per origin
3 attributes from the speed differential Δv (minimum, maximum and mean, denoted respectively Δv_{min}, Δv_{max} and Δv)	km/h
6 values from the road users' speeds (minimum, maximum and mean for each, ordered, denoted respectively s_{min1}, s_{max1}, s_1, s_{min2}, s_{max2}, and s_2)	km/h

Table 2 Numerical interaction attributes

Video Number	**Day**	**Month**	**Year**	**Hour**	**Lighting Condition**	**Type of Day**	**Weather Condition**
107031	7	1	2003	13h11	daytime	weekday	normal
112050	12	1	2005	00h39	night-time	weekday	normal
114021	14	1	2002	08h55	daytime	weekday	normal
114022	14	1	2002	21h55	night-time	weekday	normal
117051	17	1	2005	11h07	daytime	weekday	normal

Table 3 Dataset excerpt showing a few interactions and attributes

Table 3 shows an excerpt of the interaction dataset that interaction number 114021 occurred on 14th January 2002, at 08h55, during daytime, a weekday, with normal conditions.

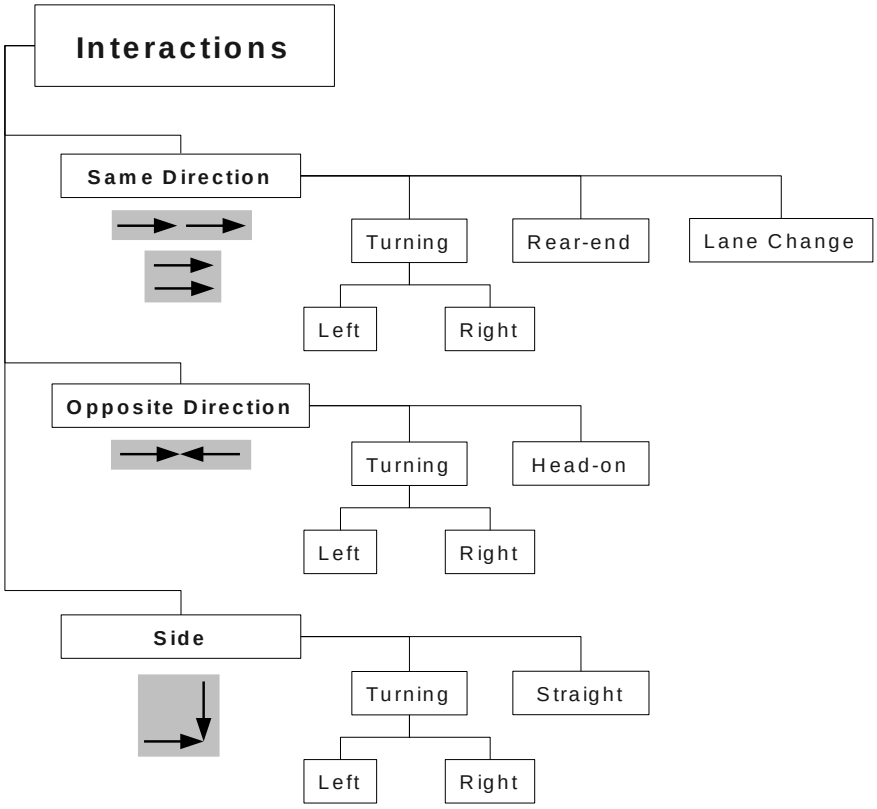

Figure 2 Hierarchy of interaction categories

RESULTS

Classification

Table 4 summarizes the characteristics of the three clusters obtained with the k-medoid algorithm, based on the analysis of the attribute importance and systematic examination of their distribution per cluster and in the whole dataset (see Figure 3 to Figure 8). The road user origin attribute is not used as it is not a generic characteristic of interactions: it could yield insight on that particular intersection, but not about collision processes.

While the whole dataset consists of 72.2 % of conflicts and 27.8 % of collisions, the proportions vary in the clusters (see Figure 3). Cluster 2 contains a larger proportion of collisions, which is 48.5 %. The conflicts are preponderant in cluster 3 where they constitute 84 % of the observations whereas the collisions represent only 16 %: cluster 3 could be considered pure. Cluster 1 consists of a mixture of collisions and conflicts in proportions which are appreciably similar to the whole dataset.

	CLUSTER 1	CLUSTER 2	CLUSTER 3
NUMBER OF INTERACTIONS	168	33	94
SPEED DIFFERENTIALS	Lowest speed differentials	Highest speed differentials	Medium speed differentials
SPEEDS	Lowest to medium speeds alternating with cluster 3	Highest speeds	Lowest to medium speeds alternating with cluster 1
INTERACTION OUTCOME	30.4 % of collisions 79.6 % of conflicts	48.5 % of collisions 51.5 % of conflicts	16.0 % of collisions 84.0 % of conflicts
INTERACTION CATEGORY	45.8 % Same direction turning left 44.6 % Same direction turning right	51.5 % Side straight 33.3 % Same direction turning right	60.6 % Side Straight 18.0 % Same direction turning left 17.0 % Same direction turning right
TYPE OF ROAD USERS	59.7 % Passenger car 30.9 % 4X4, VAN, VUS 8.6 % Truck	55.4 % Passenger car 44.6 % 4X4, VAN, VUS	53.4 % Passenger car 41.1 % 4X4, VAN, VUS 5.5 % Truck
TYPE OF EVASIVE ACTIONS	41.0 % No evasive action 45.4 % Braking	59.6 % No evasive action 36.2 % Braking	22.9 % No evasive action 61.1 % Braking 15.3 % Swerving
TYPE OF DAY	59.5 % Weekday 40.5 % Week-end	30.3 % Weekday 69.7 % Week-end	78.7 % Weekday 21.3 % Week-end

Table 4 Cluster characteristics (the analysis proved inconclusive for all attributes that are not included in the table)

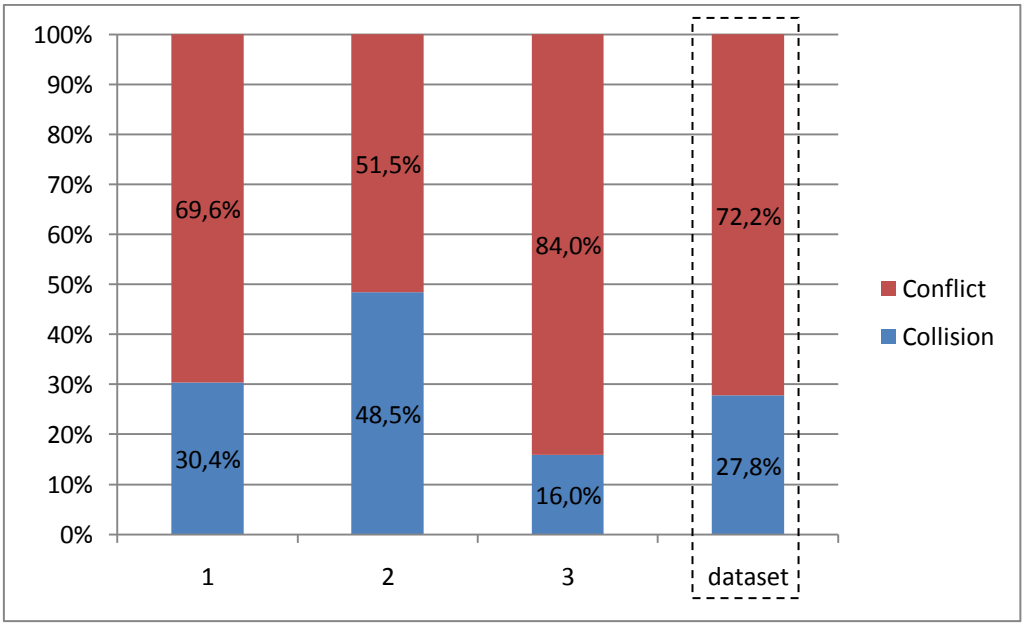

Figure 3 Comparison of the interaction outcome in the clusters (numbers 1 to 3) and the whole dataset

All speed attributes (average, minimum and maximum speeds for the two road users denoted, as well as speed differentials) are the largest overall in cluster 2 while cluster 1 has the smallest speed differential values, cluster 3 has the medium speed differential values and these two clusters share the lowest and medium speed values (see Figure 4 and Figure 5).

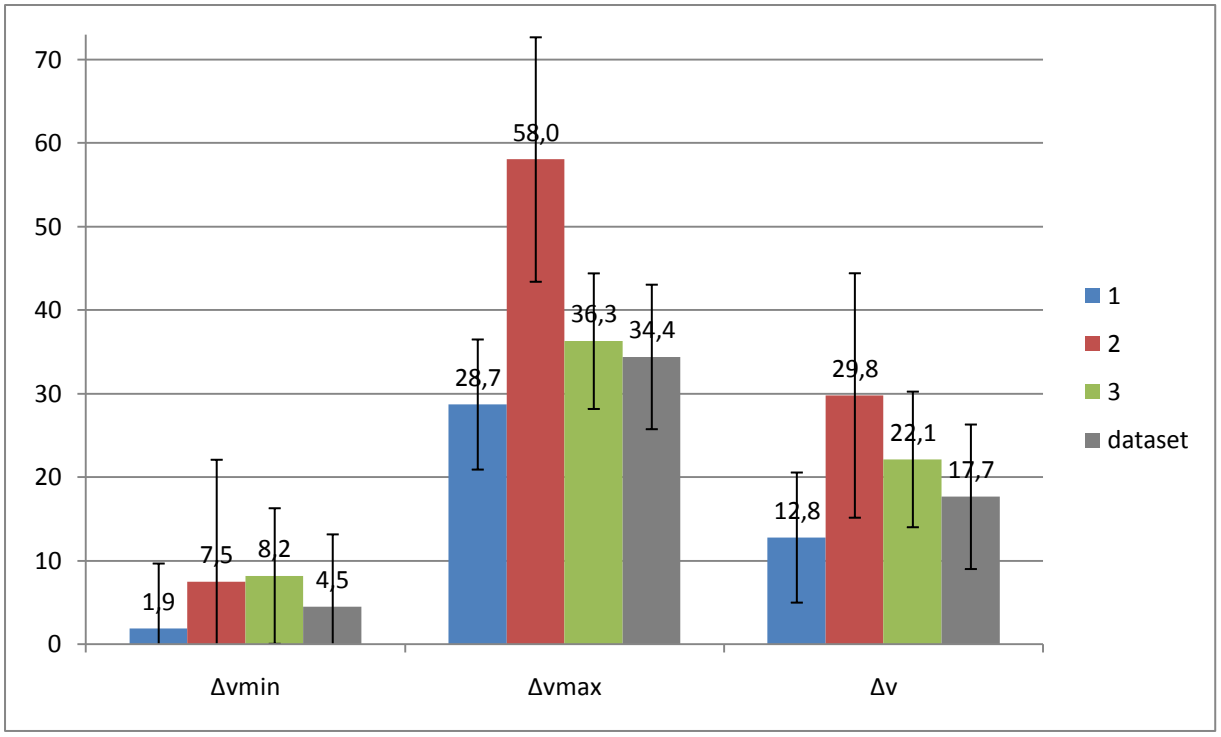

Figure 4 Comparison of the speed differential attributes in the clusters (numbers 1 to 3) and the whole dataset

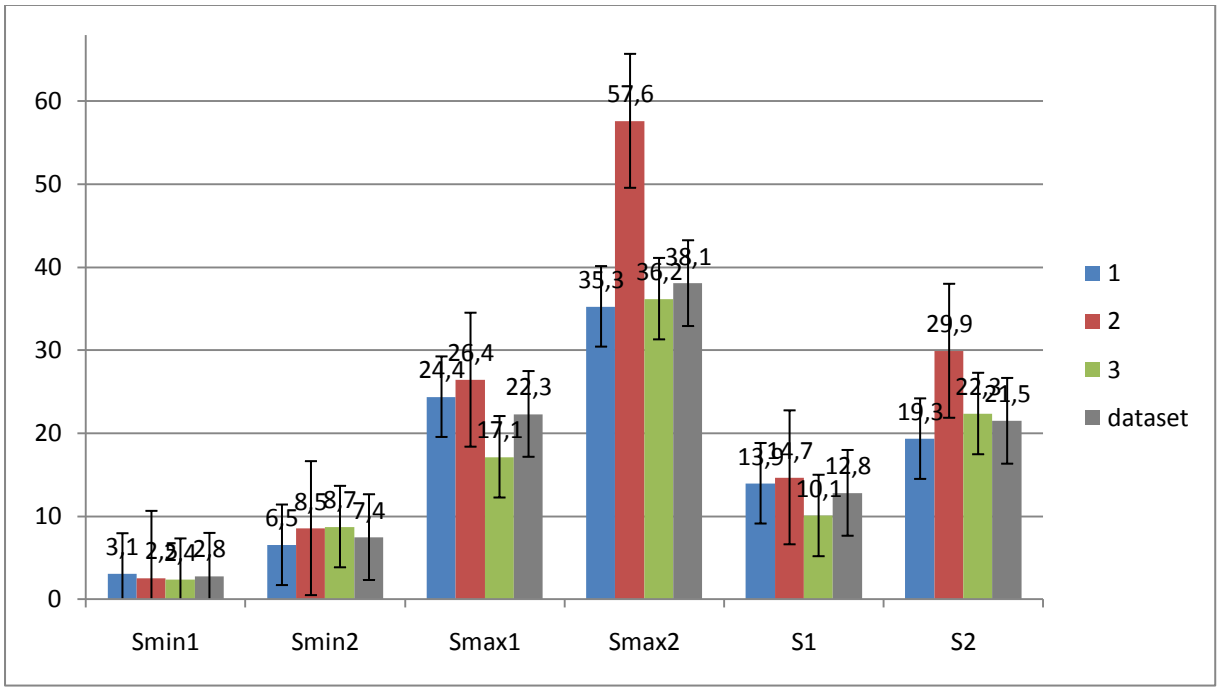

Figure 5 Comparison of the speed attributes in the clusters (numbers 1 to 3) and the whole dataset

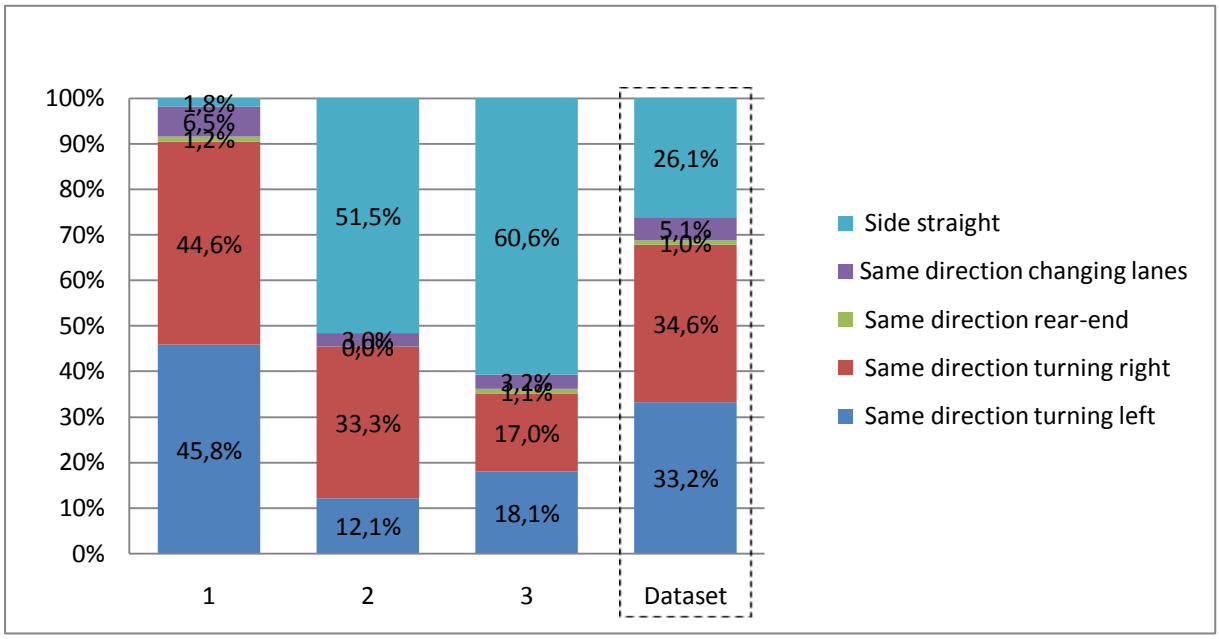

Figure 6 Comparison of the interaction category in the clusters (numbers 1 to 3) and the whole dataset

Regarding the interaction categories (see Figure 6), for which the distance used in the clustering algorithm was specifically designed, the proportions are more clear-cut. In particular the side straight interactions are almost all in clusters 2 and 3, which correspond to the most unbalanced cluster with respect to interaction outcome, with respective majorities of collisions and conflicts. The majority of the overall same direction category is concentrated in cluster 1, which is again favoured by the special distance. It would mean however that the same direction interactions that are not in cluster 1 are more similar to the side straight interactions based on their other characteristics, e.g. speeds.

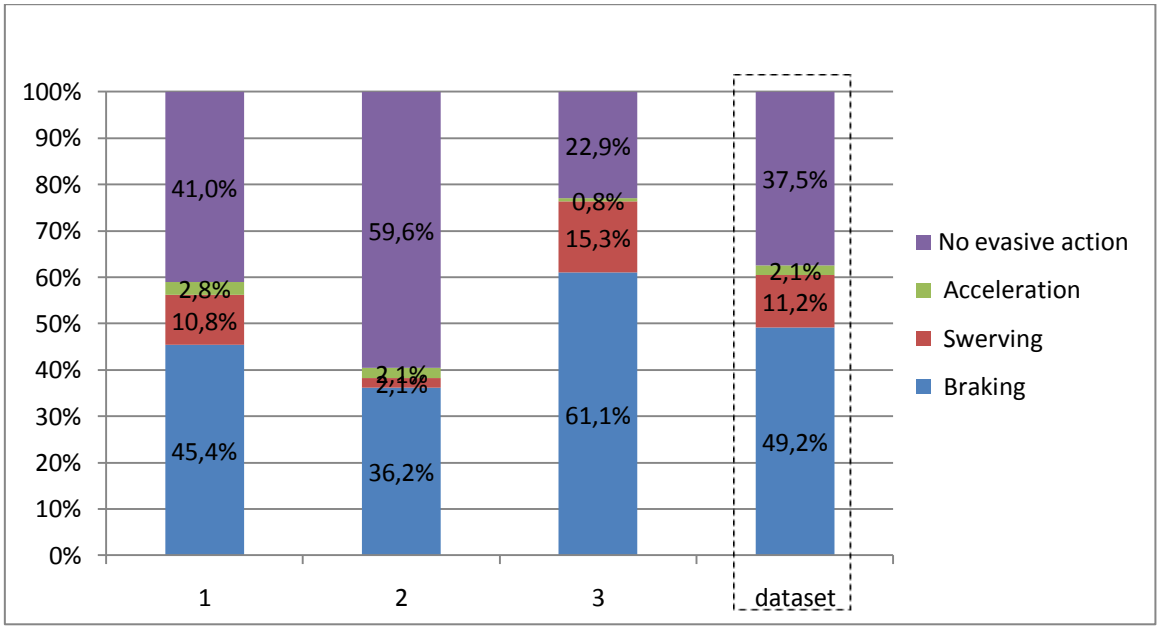

Figure 7 Comparison of the type of evasive action in the clusters (numbers 1 to 3) and the whole dataset

The types of evasive actions in the clusters are also investigated (see Figure 7): no evasive action was undertaken in 59.6 % of the observations in cluster 2 while braking was the most common evasive action (36.2 %). The proportion of interactions where no evasive action was undertaken in cluster 2 is not surprising given that it has a higher than average share of the collisions. As expected, the situation is the opposite in cluster 3. The proportion of braking is more important in clusters 1 and 3, with 45.4 % and 61.1 % respectively.

The last interesting attribute that appears to be related to the interaction outcome is the type of day (see Figure 8). Cluster 2 and 3 contain respectively 69.7 % and 21.3 % of weekends compared to 37.6 % in the whole dataset. The weekend may therefore have a link with the occurrence of collisions on this intersection.

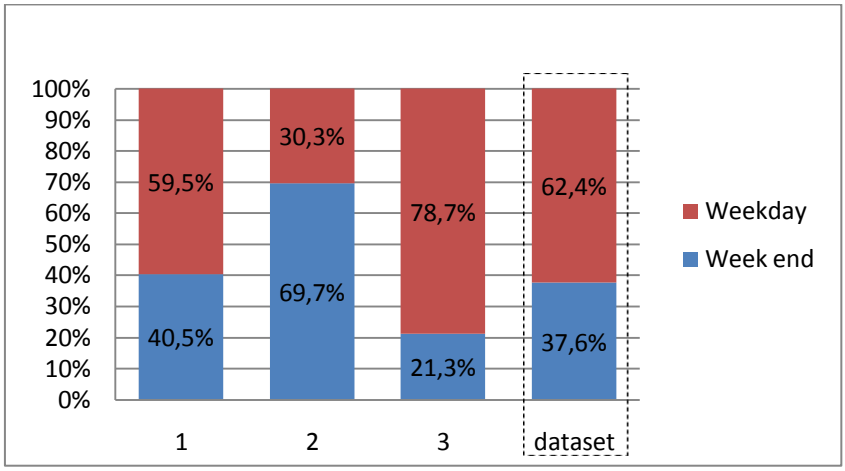

Figure 8 Comparison of the type of day in the clusters (numbers 1 to 3) and the whole dataset

The proposed classification provides knowledge about the factors that may favour collisions, and the characteristics that differentiate interactions with and without a collision. The results may not appear as clear-cut as in the previous work (15), but they are not expect to be the same since different attributes are used for classification, and a similar conclusion may be drawn about the use of interaction without a collision as surrogates for collisions: interactions with and without a collision share many similarities, and even more within sub-groups, while they are markedly different in-between groups (see some of the speed attributes, the interaction categories and the types of evasive actions). It follows that different prediction models could be calibrated for different interaction subgroups: for example, in a simple linear prediction model where the expected number of collisions is equal to the number of observed conflicts multiplied by a coefficient α ( ), a first approach would be to estimate different α for different groups.

Logit Model of Interaction Outcome

After converting all nominal attributes to numerical ones (creating *n-1* attributes for each nominal attribute that may take *n* values), a logit model that predicts the interaction outcome was computed via maximum likelihood using the open source econometrics software gretl (16). The linear model for interaction outcome is presented in Table 5, after removing iteratively all variables that are not significant with a p-value less than 0.1. The model

correctly predicts the outcome of 90.2 % of all interactions and its McFadden coefficient of determination R^2 is 0.5462.

	Coefficient	Std. Error	z-stat	Slope
const	-1.72947	1.28607	-1.3448	
Same direction turning left	2.78372	1.04016	2.6763	0.439349
Same direction turning right	1.72514	1.0261	1.6813	0.244256
Side straight	4.44196	1.34845	3.2941	0.757887
Braking	-4.1418	0.571796	-7.2435	-0.701337
Swerving	-2.67496	0.767919	-3.4834	-0.17601
No evasive	1.41745	0.546812	2.5922	0.160854
Δv	-0.180444	0.0553516	-3.2600	-0.0208568
s_2	0.138837	0.0504446	2.7523	0.0160476

Table 5 Logit model of interaction outcome (collision is 1, conflict 0) (the slope is computed at the mean)

The model is consistent with the decision tree built in (15): the most important attributes determined by the decision tree, the type of evasive action, the average speed differential and the maximum average road user speed s_2, all appear in the model, with the same, intuitively logical, sign:

- Evasive actions are negatively correlated with collisions, while it is obviously the opposite for the absence of evasive action: braking has the strongest influence on the collision outcome (both by coefficient absolute value and by p-value).
- The speed differential is also negatively correlated to collision, which may be related to the success of an evasive action, in particular braking as shown by the previous result of the decision tree.
- The maximum average speed s_2 is positively correlated, which may indicate the risks of collision carried by high speed.
- New significant attributes are related to the interaction category. There is some variability in the results depending on the order of inclusion of the various categories, and the results are especially not stable, including the sign, for all the same direction categories. However, the side-straight category is consistently positively associated with collisions, which relates to the particular danger of this interaction category.

CONCLUSION

This work is the first step in a larger project that studies collision processes based on microscopic data. It demonstrates that useful relationships between interactions with and without a collision may be automatically analyzed using statistical and data mining techniques. This paper proposes a new classification of interactions with and without a collision that highlights the main similarities and differences, related mainly to speeds and speed differentials, interaction categories, and evasive actions. Such a classification paves the way for surrogate models of safety calibrated for the various types of interactions. The

logit model of interaction outcome confirms previous results and allows quantifying the contribution of interaction characteristics and collision factors.

It is difficult to draw strong conclusions based on this classification, given the dataset limitations. This work however demonstrates new techniques for the analysis of large datasets of traffic events that are bound to become more and more common as more such data becomes available, in particular with the ongoing naturalistic driving data collections (8).

ACKNOWLEDGEMENT

The authors wish to acknowledge the financial support of the Fondation Polytechnique and the Research and Innovation Directorate of École Polytechnique of Montréal, and to thank Zu Kim of California PATH and Ann Stansel of the Kentucky Transportation Cabinet for providing the video dataset. The authors also thank the reviewer for the helpful comments.

REFERENCES

1. **Organization, World Health.** *Global status report on road safety: time for action.* 2009.

2. *A feature-based tracking algorithm for vehicles in intersections.* **Saunier, N. and Sayed, T.** s.l. : IEEE, 2006. Canadian Conference on Computer and Robot Vision.

3. *Large Scale Automated Analysis of Vehicle Interactions and Collisions.* **Saunier, N., Sayed, T. and Ismail, K.** 2010, Transportation Research Record: Journal of the Transportation Research Board, Vol. 2147, pp. 42-50.

4. *Automated Analysis Of Pedestrian-vehicle Conflicts: A Context For Before-and-after Studies.* **Ismail, K., Sayed, T. and Saunier, N.** 2010, Transportation Research Record: Journal of the Transportation Research Board. In press.

5. *Estimating the severity of safety related behaviour.* **Svensson, A. and Hydén, C.** 2, 2006, Accident Analysis & Prevention, Vol. 38, pp. 379-385.

6. *Outline of Causal Theory of Traffic Conflicts and Collisions.* **Davis, G. A, Hourdos, J. and Xiong, H.** 2008. Transportation Research Board Annual Meeting. 08-2431.

7. *Accident prototypical scenarios, a tool for road safety research and diagnostic studies.* **Fleury, D. and Brenac, T.** 33, 2001, Accident Analysis & Prevention, Vol. 2, pp. 267-276.

8. *An overview of the 100 car naturalistic study and findings.* **Neale, V.L., et al.** 2005.

9. **Implementation, Committee for the Strategic Highway Research Program 2:.** *Implementing the Results of the Second Strategic Highway Research Program: Saving Lives, Reducing Congestion, Improving Quality of Life.* Transportation Research Board. 2009. Special Report.

10. **Tarko, A., et al.** *Surrogate Measures of Safety.* ANB20(3) Subcommittee on Surrogate Measures of Safety. 2009. White Paper.

11. *Traffic conflict standards for intersections.* **Sayed, T. and Zein, S.** 1999, Transportation Planning and Technology, Vol. 22, pp. 309-323.

12. *Development of a collision typology for evaluation of collision avoidance strategies.* **Massie, Dawn L., Campbell, Kenneth L. and Blower, Daniel F.** 3, 1993, Accident Analysis & Prevention, Vol. 25, pp. 241-257.

13. *A statistical profile of road accidents during cross-flow turns.* **Clarke, David D., Forsyth, Richard and Wright, Richard.** 4, 2005, Accident Analysis & Prevention, Vol. 37, pp. 721-730.

14. **Duda, R. O. and Hart, P. E.** *Pattern Classification.* s.l. : Wiley-Interscience, 2000.

15. *Investigating Collision Factors by Mining Microscopic Data of Vehicle Conflicts and Collisions.* **Saunier, N., Mourji, N. and Agard, B.** 2011. Transportation Research Board Annual Meeting. 11-0117.

16. *Econometrics with gretl, Proceedings of the gretl Conference 2009.* **Diaz-Emparanza, I., Mariel, P. and Esteban, M.V., [ed.].** 2009.