Data Mining for driver behavior in normal driving

L.Guyonvarch, LAB, laurette.guyonvarch@lab-france.com,
M.Lutz, OCTO Technology, mlutz@octo.com
C.Chauvel, LAB, cyril.chauvel @lab-france.com
F.Josseaume, LAB, francoise.josseaume @lab-france.com,
A.Guillaume, LAB, anne.guillaume@lab-france.com
LAB - Laboratory of Accidentology, Biomechanics and driver behaviour

Note – for your revised submission, please increase total word count to 1,500 words to address the reviewers' comments, not including title, authors, tables, captions, or references. You may now use up to 3 figures or tables. Per the Journal of Safety Research guidelines, organization of material for empirical investigations should follow standard reporting format. This format can be easily expanded to full article length if you later wish to submit to the Journal of Safety Research for their special edition.

Deadline for revised submissions is Thursday, May 15, 2014, midnight PDT.

Please submit using Word (not PDF) format.

The word counts below are a suggestion to keep the material balanced. Deviations are allowed.

Problem [200 words]

Driver behavior is essential in the process of designing new Advanced Driving Assistance and Multimedia systems. It is often studied using classical statistical approach in naturalistic driving studies [1]. The full potential of data mining methods remains to be explored, which was the purpose of this study. This paper is based on data from a specific FOT (Field Operationnal Test) which took place in June 2013 near Grenoble (a city in south of France). It is part of SCORE@F, a pilot project for design of cooperative systems regarding car-to-X communications [10]. During a 1 h trip, 30 different drivers were presented on-board messages on an Ipad integrated on the dashboard. To evaluate data mining methods potential for driver behavior study, we chose a specific message concerning maximum speed limit. Available CAN data such as pedal position or speed were used to estimate whether the driver had a reaction or not and to describe frequent behavior patterns when the message appeared on the Ipad. This approach gives a different perspective for driver behavior study which will be complementary to usual naturalistic driving studies [2],[3]

Method [400 words]

To take maximum advantage of the data, we used a three-step method (1) data visualization and analysis; (2) state sequence analysis; (3) sequential pattern mining. For (1), data were cleansed to identify the relevant points of interest (speed limit message here) and perform an appropriate sampling (i.e. ± x time of recorded data before and after the message). Then we explored the sampled dataset with different data visualization tools (GPS tracks, multiple time series, boxplot). For (2), we converted the data into qualitative coding (i.e the percentage of the throttle actuation translated into states such as “strong actuation”, “slight actuation”, etc.). This data coding gives a representation of driver‘s behavior as state sequences, usually easier to interpret. Finally, through step (3), we mined frequent driving behavior patterns. We focus on the events representing state changes (for instance, steady states are removed from the sequences). Frequent subsequences included in these sequences were extracted by a sequence mining algorithm. Several methods could be used in this way (SPADE, PrefixSpan, FreeSpan, GSP…): refer to [4,5] for a literature review. Sequence data-mining techniques initially came from marketing application (market basket analysis) and their use in industrial contexts is fairly innovative (see for instance [6] for an application to manufacturing processes). To demonstrate their potential of application for driver behavior analysis, we performed a case-study with the R language and the ad hoc TraMineR library [7]. We decided to use this library because of its efficient mining algorithm implementation and its off-the-shelf results visualization tools. The detail of the implemented algorithm is provided in [8]. In our research context, such a mining aims at discovering archetypal behaviors (e.g. statistically significant) that could put forth the possible influences of the displayed messages. Constraints can be added in this search: maximal time span during which a subsequence should occur, maximum time between two transition etc.

Results [350 words]

For the sake of brevity, we only present steps (2) and (3) results, the most original part of our work. Fig. 1 shows how the states distribution of a variable are visualized.

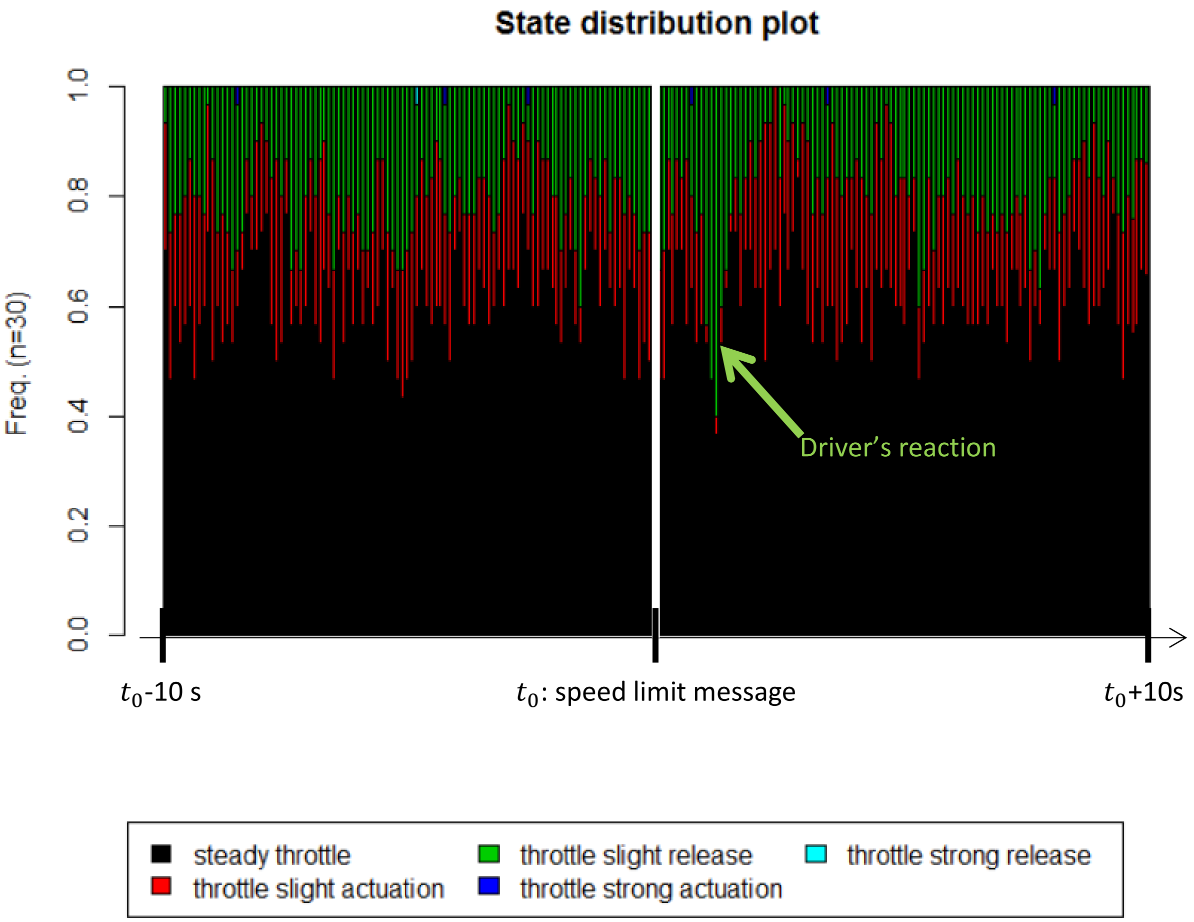

Figure 1 : Visualization of the states of distribution

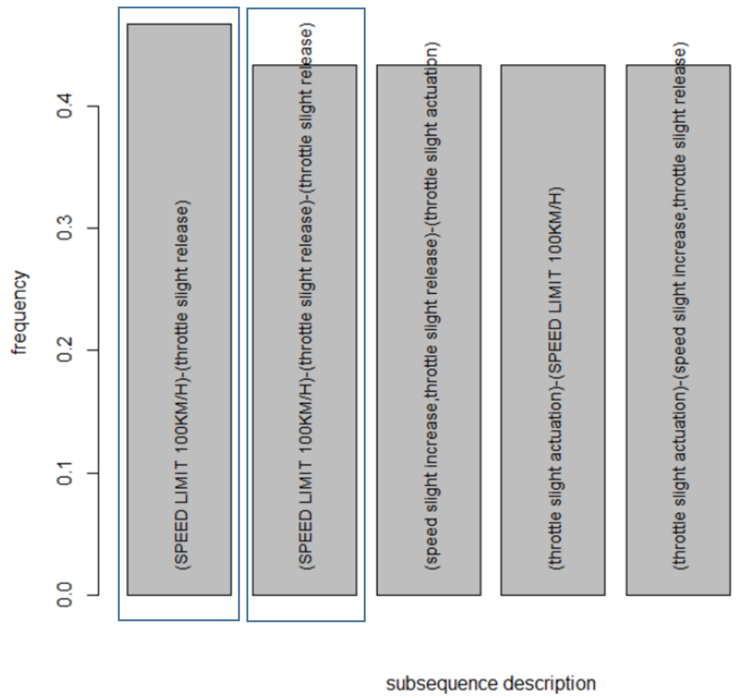

Figure 1 : Full driver's events sequences mining (extract)

It displays the distribution of the throttle actuation sequence of 30 drivers, along a 200ms timestamp. The variable states were coded as explained above. A message "Warning: speed limit" was displayed at $t=t_0$. A strong increase of the states "throttle slight release" was observed a few times after the message: an impact of the message on driver's behavior can be supposed. Displaying such information for all recorded variables provided first insights about driver's behavior. Then, step (3) provided a more systematic mining of the data. Full driver's change events sequences were mined. These sequences were coded as follow: x_0-(event$_1$, event$_2$, …, event$_m$)-x_1-(event$_1$, event$_2$, …, event$_{n,}$)-x_t-…, where events into brackets represent simultaneous events and x_1, …, x_t the elapsed times between events. For instance, an extract of a sequence could be: "100-(speed increase, throttle slight actuation)-100-(speed decrease, throttle slight release, turning indicator left on)-100". These sequences were long (potentially several hundreds of events per driver). To reduce the computational cost of the mining, we focused on subsequences that may occur on the proximity of the speed limit message (t_0) and tracked frequent subsequences with the following constraints: three events maximum, 500ms maximum between two transitions. Of course, other settings could also be tested for further insights. A sample of the results is shown in Fig. 2. We were interested in observing the behavior following the display of the message. For instance (boxed in blue), we found that the subsequence "("message")-("throttle slight release")" occur in 47% of the cases, "("message")-("throttle slight release")-("throttle slight release")" in 43% of the cases, etc. This method reveals to be a powerful tool to automatically extract archetypal behavior following a message.

Of course, this is only a first result. We now have to perform deeper tests for a full evaluation of the potential of this method. First of all, a bigger sample is expected for further studies, for more statistical significance. Furthermore, results for other type of Ipad message should be analyzed as well. Anyway, this is a promising method for driver behavior analysis, allowing automatic extraction of behavioral pattern without any a priori knowledge.

Discussion [300 words]

The study can be completed in many ways. First, adjustments should be tested to optimize the method: sampling frequency, states and events coding, pattern mining constraints etc. Second, we are working on another important improvement: completing the method with a discriminant analysis. The objective is to identify distinctive behaviors among subgroups of the whole population. TraMineR offers straightforward functions to perform such an analysis. However, a question must be addressed: how should the population be categorized? Two approaches are considered: 1) using qualitative knowledge about drivers (age, sex…); 2) inferring behavioral subgroups from the data. To explore the latter, we tested several statistical approaches, such as usual clustering methods or longitudinal studies.

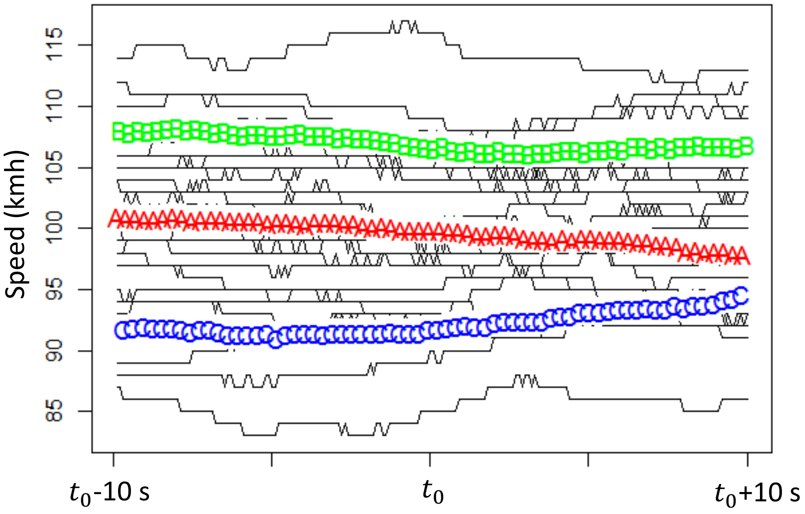

Figure 2 : Speeds over time clustered with the KmL algorithm

For example, fig. 3 shows how three subgroups are constituted, by applying a k-means longitudinal clustering algorithm (R library KmL) to the speed profiles of the drivers, before the message (the assumption is that speed patterns make it possible to identify homogenous subgroups). Three groups are characterized: high speed with slight acceleration before the message (A), medium speed with deceleration (B), low speed with acceleration (C). State sequences representation and frequent subsequences mining can then be applied to discriminate the analysis according to these subgroups. Significant differences are observed: for instance, the throttle slight release after the message is particularly noticeable for drivers from the B group. These results should now be interpreted and validated on a larger sample of drivers. Mixing qualitative information about the drivers and statistical clustering will certainly lead to promising insights for discriminant analysis in the context of sequence mining.

Algorithmic data mining approach therefore reveals to be a very useful way to complete usual statistical naturalistic driving studies: both approaches are complementary. The future of quantitative naturalistic driving studies could be grounded on developing synergies between both approaches, to address a research problem according to multiple points of views. As an example, SCOOP@ F project will study potential for IT data mining systems on a larger scale. A particular attention will be given to driver inattention. This data mining approach will lead to better understanding of the impact of messages on driver behavior.

Another upcoming project is UDRIVE. This European NDS includes research questions about the study of driver behavior during secondary tasks. Data mining methods will offer a useful complement to classical statistical study (ANOVA, Relative risks study, Odds ratio)

Summary [250 words]

Naturalistic driving studies are the main way to study driver behavior in normal driving. They help car manufacturers to design safe multimedia and driving assistance systems. This work is based on SCORE@F data, a French pilot project for design of cooperative systems regarding car-to-X communications. SCORE@F was set up to evaluate stakes and benefits regarding fatalities and severe injuries addressed by these systems. In our work, these data were used in a more specific way: our study evaluates data-mining methods potential for driving behavior study.

The SCORE@F FOT consists in different rides where on-board information messages (traffic, road safety, service etc.) were presented to drivers on an Ipad. CAN data, GPS data and information messages were recorded and synchronized. This paper specifically focuses on speed limit message scenario among the different on-board information messages. After a first exploration of data, we propose a data coding process (20 sec around the time of message) highlighting frequent driver behavior for this specific scenario in terms of throttle pedal actuation. Then through a sequence data-mining algorithm, we were able to extract archetypal behaviors linked to this message. In addition, clustering methods and longitudinal studies give more precise evaluation of driving behavior. Developing synergies between usual statistical approach and such data mining approach will expand the possibilities of driver's behavior understanding for the future naturalistic driving studies like SCOOP@F or UDRIVE naturalistic driving study.

References

[1] F.Faber et al. (2011)

EuroFOT Deliverable D6.2 Analysis methods for user related aspects and impact assessment on traffic safety, traffic efficiency and environment

[2] Saint Pierre G., Tattegrain H., Val C. (2014), « Impact evaluation of speed regulation systems using naturalistic driving data: The EuroFOT example », Transport Research Arena 2014, Paris, La Défense

[3] NHSTA (2010) An Analysis of Drive Inattention Using a Case-Crossover Approach On 100-Car Data: Final Report

4] N.R. Mabroukeh et C.I. Ezeife (2010) A taxonomy of sequential pattern mining algorithms, ACM Computing Surveys 43: 1

[5] J. Han et al. (2007) Frequent pattern mining: current status and future directions, Data Mining and Knowledge Discovery 15: 1

[6] Kerdprasop et Kerdprasop (2013) Performance analysis of complex manufacturing process with sequence data mining technique, International Journal of Control and Automation 6: 3

[7] http://mephisto.unige.ch/traminer/

[8] F. Masseglia, Algorithmes et applications pour l'extraction de motifs séquentiels dans le domaine de la fouille de données : de l'incrémental au temps réel, PhD. thesis, Université de Versailles Saint-Quentin en Yvelines

[9] http://christophe.genolini.free.fr/kml/

[10] LAB, CEREMA (2014) Rapport d'impact sur la securite routière et le trafic routier
projet score@f, livrable L531