Information theoretical methods dedicated to accidents analysis for GIDAS database.

Mathilde Mougeot [1,2,3], Robert Azencott[3,4]

[1] Modal'X,Université ParisX, 200 avenue de la République 92001 Nanterre , France.
[2] LPMA / UMR 7599, Université Denis Diderot, 75251 PARIS Cedex 05, France.
[3] Mathematics Depart CMLA, ENSC, 61 avenue du président Wilson, 94 Cachan
[4] Mathematics Department, University of Houston, Houston, Texas 77204-3008, United states

ABSTRACT

Nowadays, traffic accidents are recorded in historical databases. Regarding the huge quantity of data, the use of data mining tools is essential to help Experts, for automatically extracting relevant information in order to establish and quantify relations between severity and potential factors of accidents. An innovative approach is here proposed for an in depth investigation of real world accidents data base. Mutual information ratio based on conditional entropies is used to quantity the association strength between an accident outcome descriptor (injury severity) and other potential association factors. Information theoretic methods help to select automatically groups of factors mostly responsible of the severity of accident.

This work was conducted in the framework of the European project TRACE (Traffic ACcidents in Europe)[1]

Keywords: mutual information, conditional entropy, risk analysis.

NOTATION

MIR: Mutual Information Ratio

INTRODUCTION

Nowadays, traffic accidents are progressively reported and stored, through many fields, in historical database. In the GIDAS database devoted to German traffic accidents, more than 800 fields are potentially defined to describe an accident and more than 2000 new accidents are stored each year. Investigating relevant accident causations hidden in huge databases is an important goal for improving our knowledge on traffic accident and traffic safety. New preventive actions can also emerged from in depth investigations of real world accidents, with one objective, to reduce, in the future, rate and severity of accidents. This study focuses on injury severities. One of the main objective is here to find the main accident causes impacting the severity of the accidents.

The relation strength between injury severity and other variables can be quantified or modelled statistically. Depending on the nature of the variables, the association strength is measured differently. For continuous variables, the correlation coefficient, ρ, is a long-standing measure to evaluate the statistical dependence between variables and this coefficient is quite used in accidentology ([Huang et al., 2007]). For categorical data, the Cramer's V based on the X^2 statistics is mostly used to quantify the association between two variables. Both association coefficients are driven by some specific underlying hypothesis. Correlation coefficients are known to measure only linear dependence between variables. If the relation is not linear, then the use of this type of coefficient is definitely not the most efficient. For qualitative variables, in case of sparse contingency tables, the Cramer's V indicator, based on X^2 test, can also be inappropriate. When investigating large data bases, prior knowledge of functional relationships between variables is never directly available and consequently, the use of correlation coefficients, based on linear assumptions, can be totally inappropriate to measure statistical dependencies.

[1] Authors thank Claus Pastor from BAST Institute for useful discussions and relevant comments during TRACE project.

Mutual information (MI), introduced by Shannon (1949) is a measure of statistical dependence which is able to catch complex relation between variables, even in cases of non linear dependence [Billingsley, 1965; Cover et al., 1991]. Mutual information ratio can be computed within discrete, continuous and discrete-continuous variables [Brillinger 2004], and provides also a powerful extension to the classical correlation and Cramer's V measures.

GIDAS DATA BASE

In Germany, since 1999, a consortium of two institutes (BAST, Federal Highway Research Institute and FAT, German Association for Research on Automobile-Technique) drives an important project of German In-Depth Accident Study [GIDAS]. For this purpose, teams of physicians and technicians collect information on personal injury accidents. In the area of Hanover and Dresden, all personal injury traffic accidents occurring are reported continuously by the police and the fire department stations. Accidents are selected according to a defined random procedure and then are carefully described following a given protocol. A detailed description of the investigation methodology can be found in [GIDAS]. Annually, approximately 2,000 traffic accidents are recorded in this way and the information is stored in an historical database. The "GIDAS" database is now the biggest and most complete In-Depth accident survey and data collection in Europe. In order to avoid distortions in the data structure of accidents recordings by different teams, the data are weighed annually through comparison with the officially recorded accident structure. This ensures that the present accident data are regarded as representative for the investigation area of the cities and administrative districts of Hanover and Dresden. The number of available observations in GIDAS database was at the end of year 2006 around 14 000 with the following per year repartition: 1999 (1018); 2000 (1987); 2001 (1906); 2002 (1643); 2003 (1806); 2004 (1849); 2005 (2007); 2006 (1737).

Accident outcome descriptor

For the current analysis, two accident outcome descriptors have been chosen: the maximum accident severity, (MAIS) and the accident Injury Severity of the head region (HWS).

Maximum Injuries Severity (MAIS)

In GIDAS database, MAIS original distribution is defined over 7 categories {0,1,2,3,4,5,6} which correspond to different injury severities: 0 corresponds to non injury, and 1 to 6 to more and more severe injuries (Figure 1).

Figure 1 : MAIS distribution for GIDAS data. Original and agregated distribution of data.

The original MAIS distribution is agregated into three categories in order to analyze accidents leading to "not injured", "slightly injured" and "severe and fatal injured". The "Safe" label corresponds to observations with no injury (label 0), the "slightly Injured" label corresponds to observations with some injuries (labels 1 and 2), the "severe injured" corresponds to labels higher than 3. Most of the accidents stored in the database lead to no injury (rate 60%) or to minor injuries (rate 74%, for 0 and 1 labels).

Head injuries (HWS)

In GIDAS database, Head injuries are stored in the output variable "HWS", defined over 7 categories as for MAIS.

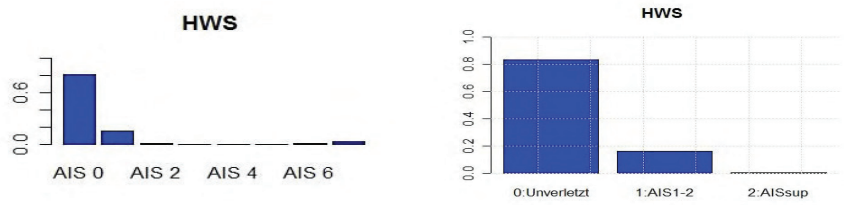

Figure 2: Head injuries distribution for GIDAS data. Original and agregated distribution of data.

Figure 2 shows that a large majority of accidents (80%) led to no injury for the head. HWS variable is aggregated into three categories to study safe, slightly or severe injured people relatively to the head. Histograms are built with 11586 observations.

Potential association factors

The choice of factors, used for this study, has been selected in collaboration with the German BAST institute. One of the objectives of this application is to focus on specific factors, to measure and compare the association strengths between factor and outcome, and then to determine which combination of factors is most influent on the outcome. For MAIS and HWS, the goal is especially to determine statistically the impact of each pre selected factors on the severity of the injuries, and which smallest group of factors can explain the injuries severity distribution given GIDAS data. Factors are listed in Table 1 for the analysis of accident injury severity (MAIS, HWS).

Variable (tag name)	Description	Number of modalities and brief description
GENDER	Gender	(2) male/ female.
PLACE	Place of the accident (urban/rural)	(2) urban/ rural.
TIME	Time of the day	(3) day/night/dawn
COLLSPEED	Initial speed of collision	Continuous
SEATBELT	Seat belt usage	(2) belted/ unbelted
ACCTYPE	Type of accident	(7) F/AB/EK/UES/RV/LV/SO
ACCKIND	Kind of accident	(10) unfall/ anfährt/…
LIMITSPEED	Speed limit at the accident scene	(17) 5 km/h/…/ 140 km/h
GUILTY	Responsible or not for the accident	(2) yes/no
OPPONENT	Opponent	(7) others Car HGV Bike Cyclist Pedest. Object
AGE	Age of the driver	(8) (0,18] , (25,30] (30,35] … (65,75] , (75,100]
AIRBAG	Use of the airbag	(2) AIRBAG /no AIRBAG
CARAGE	Age of the car at the date of the accident	continuous
DAMAGE	Main damage to the car (front, size, rear))	(7) Front Right Side … Bottom
ROLLOVER	Rollover (yes/no)	(2) yes/no

Table 1: Association factors used for MAIS or HWS outcome descriptor.

The following graph (figure 3) shows the empirical histogram computed for the different factors.

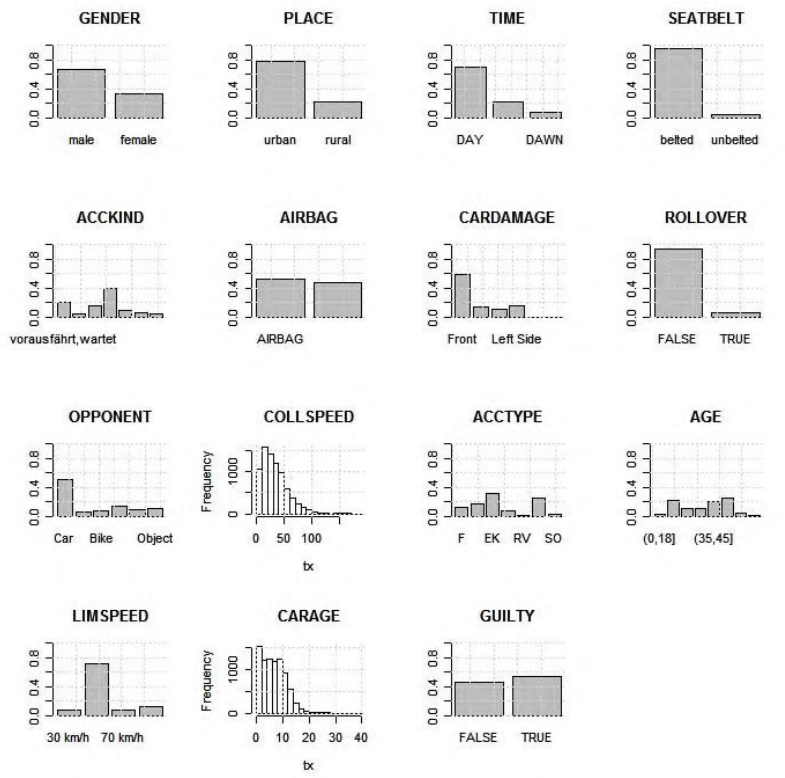

Figure 3: Histogram of potential association factors for MAIS descriptor.

MUTUAL INFORMATION

Mutual information, based on conditional entropy, quantifies the relation between two random variables X and Y. For example, Y can describe an accident gravity descriptor and X an accident causation factor.

The **Entropy** measures the average of information provided by the knowledge of a variable. For an X variable defined over a set of α_i modalities each of them with an occurrence probability p_i=Probability(X= α_i), $1 \le i \le m$, the entropy, H_X, is defined by:

$$H_X = -\sum_{i=1}^{m} p_i \log(p_i) \qquad [1]$$

By convention, $0\log(0) = 0$

If X is deterministic, the entropy is minimal, and H_X=0. The occurrence of any extra value of X brings no complementary information for the knowledge of X, which is, in this case, constant. On the opposite, for a uniform distribution, the entropy is maximal: H_X=m. All modalities of X, which have the same probability to occur, bring new information.

For two discrete variables X and Y, defined over a set of α_i and β_j modalities, with joint probability p_{ij}=Probability(X= α_i, X= β_j) , $1 \le i \le m$, $1 \le j \le p$, the **joint entropy,** $H_{X,Y}$ is defined by:

$$H_{X,Y} = -\sum_{j=1}^{p} \sum_{i=1}^{m} p_{ij} \log(p_{ij}) \qquad [2]$$

Conditional entropy

$H_{Y/X}$ measures the average of information brought by variable X for the knowledge of Y and is defined by:

$$H_{Y/X} = -\sum_{j=1}^{p} \sum_{i=1}^{m} p_{ij} \log(p_{j/i}) \quad [3]$$

pj/i denotes the conditional probability of $Y= \beta_j$ given $X= \alpha_i$. If X and Y are independent, then $H_{Y/X}= H_Y$: knowledge of X doesn't' bring any help nor information for the knowledge of Y.

Mutual information

Based on conditional entropy, Mutual information is a measure of statistical dependence between two variables X and Y. $I_{X,Y}$ quantifies the amount of information provided by the knowledge of variable X for the complementary knowledge of variable Y.

$$I_{X,Y} = H_{x} - H_{X,Y} \quad [4]$$

Normalized by the entropy of variable Y, the **mutual information ratio (MIR)**, $R_{X,Y}$, is a zero to one range measure of the dependence of X and Y.

$$R_{X,Y} = \frac{I_{X,Y}}{H_Y} \quad [5]$$

For two independent variables X and Y, prior knowledge of X doesn't provide any information for the knowledge of Y: $R_{X,Y}=0$. On the opposite, if a deterministic relation exists between X and Y then prior knowledge of X implies a specific value of Y, the mutual information ratio is also maximal: $R_{X,Y}=1$.

Estimation of Mutual Information Ratio

In most operational cases, theoretical distributions of jointly variables are not known and MIR should be estimated.

Considering, N independent realizations of (X, Y) available in an accident database, $v_{ij}(k)$ denotes the potential occurrence of (X, Y) for both modalities α_i and β_j and for realization k. if $v_{ij}(k) =1$, it means that (α_i, β_j) occurs at k, $v_{ij}(k) =0$ otherwise. The probability p_{ij} can be estimated by the maximum likelihood as follow:

$$\hat{p}_{ij} = \frac{1}{N} \sum_{k} v_{ij}^{k} \quad [6]$$

The plug-in estimate of the mutual information ratio is then ; with

$$\hat{H}_{X,Y} = -\sum_{j=1}^{p} \sum_{i=1}^{m} \hat{p}_{ij} \log(\hat{p}_{ij}) \quad [7]$$

$$\hat{R}_{X,Y} = \frac{\hat{I}_{X,Y}}{\hat{H}_{Y}} \quad [8]$$

Consistent estimation of $R_{X,Y}$ is computed by a bootstrap aggregating procedure [Efron & Tibshirani 1993]. The Mutual information ratio is computed using B replications of the same unit procedure. For each b replication, an estimation of $R_{X,Y}(b)$ is performed, for a subset of observations chosen at random from the original data set. $R_{X,Y}$ is estimated by averaging all unit estimations of $R_{X,Y}$ (b) over the B replications.

$$\hat{R}_{X,Y} = \frac{1}{B}\sum_{b} \hat{R}^{b}_{X,Y} \qquad [9]$$

A similar bootstrap procedure is used to compute confidence intervals.

Selection of factors using mutual information ratio

Given a specific injury severity outcome (Y) and p potential accident factors (X_1,…Xp), mutual information is used to estimate statistically the relation strength between Y and the factors.
Considering the p factors, mutual information ratios are computed, in a first step independently, for each single variable using equation (7). To compare the respectively influence of the different variables on the Y outcome, the coefficients, $R_{X1,Y}$, …$R_{Xp,Y}$ are sorted in decreasing order of magnitude. We note $X_{(1)}$, the variable associated with the largest MIR, which has the highest predictive power on Y.

$$R_{X(1),Y} = \max_j \left\{ R_{X_j,Y} \right\} \qquad [10]$$

Each coefficient lies between 0 and 100%, and measures the percentage of mutual information brought by X on Y entropy.

Mutual information can also be computed for multivariate factors [Joe 1989].
Let note $X=(X_{i1},\ldots X_{ik})$, a multivariate variable of k factors ($k \leq p$). Mutual information ratio is computed, in a similar way, using equations (6) & (7). It is also possible to compute mutual information ratio considering a fixed number of factors: k=1 or k= 2 or k=p. In order to select a subset of k factors, which, in combination with each other, have the highest predictive power for Y, the same procedure as the one described above for single factors is applied to select the group of k variables with the highest MIR. This sub-group of k factors best explains the Y outcome distribution.
This method provides also an efficient and rigorous way of constructing hierarchies of causality factors on a given variable Y. The selected causality factors also computed have a strong prediction power on the outcome descriptor Y, and can be used as input to a model.

APPLICATIONS FOR RISK FACTORS QUANTIFICATION

In the German GIDAS database, most of the variables are qualitative, we hence have a natural situation where classical correlation analysis may be of limited use, and information theoretic methods based on entropy computation offer a more rigorous exploration tool for association or causal relations. The previously methodology has been applied on GIDAS database, with, at the end of 2006, 14000 observations, described over more than 800 fields [Mougeot & Azencott 2007]. A first pretreatment has been applied on the whole database to eliminate inappropriate values. All the studies have been carried on using the statistical programming software R [R development Core Team]. All the codes have been developed using the R standard language.
Mutual information is used to estimate the relation strength between each outcome descriptor (MAIS, HWS) and the corresponding potential factors presented in the previous tables. The observations of GIDAS data base are used to estimate the relation strength. Each estimated coefficient is computed using more than 8000 observations, depending on the proportion of missing values. For this specific

study, given one MIR ratio, all missing values have been eliminated for the involved variables. In a first step, the MIR coefficients are estimated independently for each single factor, and ordered. Groups of multivariate factors which best explain the accident outcome is computed afterwards.

MAIS

The MIR coefficients are first estimated using all original categories of MAIS (7 modalities) , and then estimated again using aggregated factors of MAIS (“Safe”, “slightly” or “severe” Injured categories). MIR coefficients, which evaluate the association link between MAIS outcome descriptor and each one of the accident causation factors selected by BAST, are computed and sorted by decreasing order of magnitude (Figure 4).

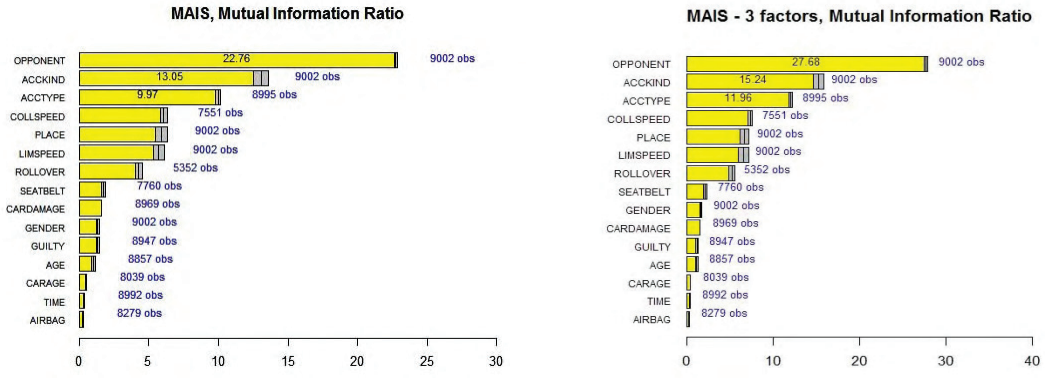

Figure 4 MIR for MAIS. Left: initial distribution. Right: aggregated distribution.

The presentation of the results is the following (Figure 4): considering an outcome descriptor (here MAIS), each MIR coefficient computed for one single factor is represented by a horizontally bar. The strength of the association, given by the MIR coefficient, is represented by the length of the horizontal bar (depending on the graph scale). The name of the tag of the corresponding factor is written on the left and Table 1 gives the description of corresponding tag names. The number of joint observations used for computing the coefficient is written on the right, and corresponds to the non missing values used to estimate each coefficient. On the right end of each bar, a confidence interval, computed by bootstrap, is represented for a 95% risk level. All MIR coefficients lie theoretically between 0 and 100%.

Considering, MAIS outcome descriptor, it appears that the type of OPPONENT during the accident is the most influent facto, with a MIR around 23%. As the number of initial modalities is reduced to aggregated classes (“safe”, “slightly” or “severe” injuries severity), this feature is even sharper and MIR value increases up to 28%. The Accident KIND appears in second position (13%; 16% for aggregated modalities), and the Accident TYPE in third position (10% or 12%). The association strength of all coefficients increases when computed from the original to the aggregated distribution.
The SPEED of collision, the PLACE and the limitation of speed obtain similar MIR coefficients. Although ROLLOVER accidents are quite rare, their impact on MAIS seems quite severe (MIR 5.8%).
The SEATBELT factor appears in the middle of the list and obtains a small coefficient (1.95%). SEATBELT usage is usually considered to be an important factor affecting the injuries severity of vehicle traffic accidents and, on a first view, this result seems to be contradictory with all knowledge about accidents causes and severity. Today, drivers and passengers are required by law to use their seat belt and the rule seems to be followed by most drivers: 97% of the available observations of GIDAS correspond to the use of seat-belt. So, statistically speaking, there is, today, no statistical variations for SEATBELT usage (or not), and this is confirm by the investigation of the real world accidents recorded in GIDAS (Figure 5).
In order to point out and to focus on, the severity of accidents due non seatbelt usage, we have artificially selected a subset of data in GIDAS with an equal proportion of observations corresponding to the usage (or not) of seatbelt . All observations corresponding to the non usage of seatbelt have

been taken (minor proportion), and have been completed with an equal proportion of observations, taken at random, corresponding to seatbelt usage. In order, to obtain, a robust estimation of the MIR coefficient, this procedure has been replicated 20 times, and the MIR coefficient has been averaged over all replications. For this artificial mixture of observations, the impact of the SEATBELT factor increases from 1.95% to 14%, which is quiet a high value (a 14% MIR corresponds to a second position in the ranking list). SEATBELT usage stays an important factor which is directly linked to injury severity. As today, a large majority of drivers wear their seatbelt; this factor appears to be less important in the real world accidents population.

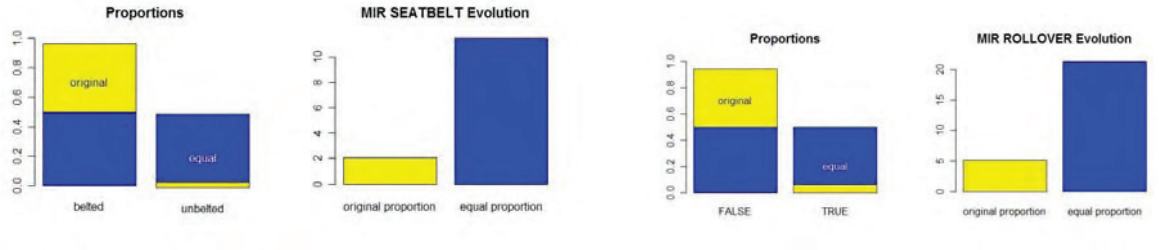

Figure 5: Seatbelt usage and Rollover accidents for original data (yellow) and artificially equaled proportion (blue). Impact on MIR.

If we compute the impact of rollover accident, using the same procedure as used before for the seatbelt factor, we observe that the MIR coefficients increases to 25%, which confirm the gravity of rollover accidents (figure 5).

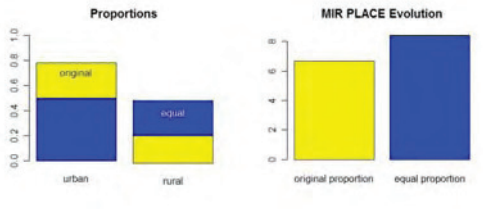

Figure 6: Accident place proportion for GIDAS data (yellow) and equaled proportion (blue). Impact on MIR.

On the opposite, urban and rural accident places do not have a strong impact on injuries severity even after re sampling (figure 6).

Multivariate analysis is then conducted to analyze for a given number of explanatory variables, which group of factors has the highest mutual information ratio, and best explains Maximum Injury Severity. The following graph presents, for MAIS outcome descriptor, the estimation of the highest MIR, as function of the number of factors (**Fehler! Ungültiger Eigenverweis auf Textmarke.**).

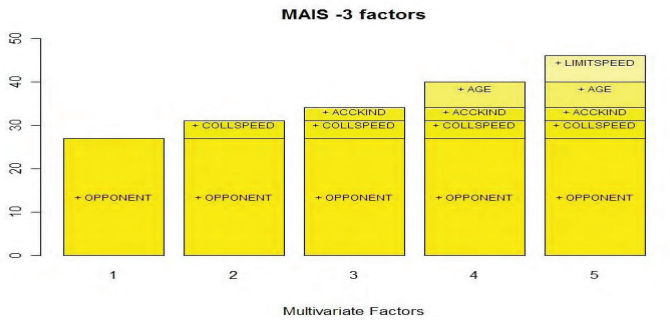

Figure 7: Mutual Information ratio for MAIS descriptor (aggregated modalities) in multivariate case.

For instance, the 3rd column indicates that the group of 3 factors (OPPONENT, Collision SPEED and Accident KIND) has a multivariate MIR of 38%; this group is associated with the highest predictive power for all groups of three factors (figure 4).

It is interesting to observe that, for the single factor analysis, OPPONENT, Collision SPEED and Accident KIND were respectively in first, second and third position, regarding the association strength level. In the multivariate analysis, the Collision SPEED, which was in 4th position for the single factor analysis, combined with the opponent factor, best explains injuries severity.
The previous results were conducted for maximum injury severity. It is possible to focus the analysis on specific injuries. A similar study is then conducted for head injuries.

HWS

Mutual Information ratios are then computed for head injuries for the same potential factors as for MAIS described in Table 1. As observed for MAIS outcome descriptor, the MIR coefficients estimated for HWS are sharper when computed for an aggregated distribution as for the original distribution (figure 8). The OPPONENT is, as for MAIS, the most influential factor explaining head injuries severity however the relation strength is smaller (12,5% as compared to 23%). The same holds true for the factors Accident KIND and TYPE which are again placed second and third. "GUILTY", who describes whether a driver has been held responsible for causing the accident, is now at 4th place. The GENDER becomes quite important for head injuries, probably indicating that women are more vulnerable than men in this case. The mainly damaged part of the car (DAMAGE) comes also into play, probably indicating that rear end collisions play a high role.

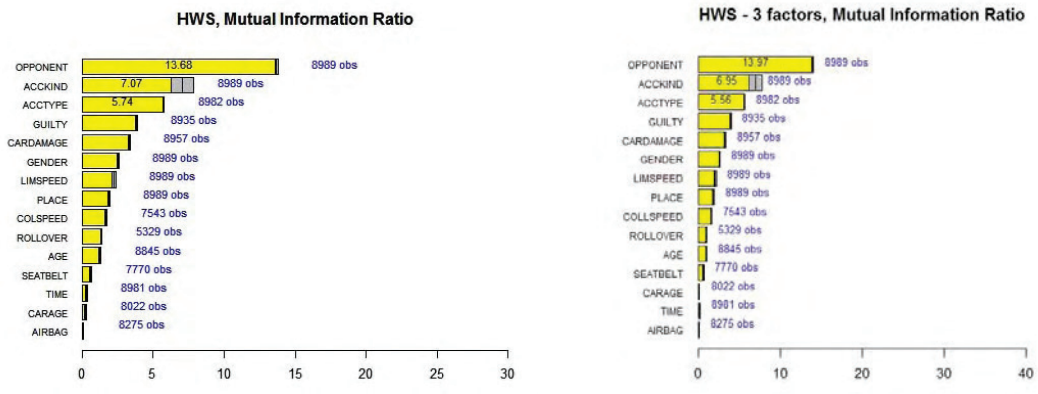

Figure 8: MIR for HWS. Left: initial distribution. Right: aggregated distribution.

Multivariate analysis is then conducted, as for MAIS, to select which group of factors has the highest mutual information ratio, and best explains head injury severity. Results are presented in the following figure (figure 9). Both factors, OPPONENT and GEGNER explain head and MAIS injuries severity.

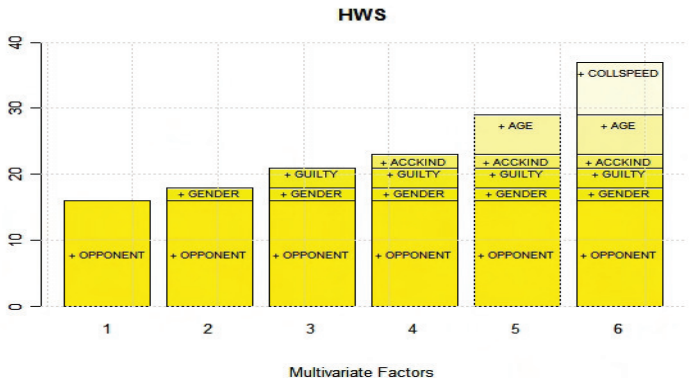

Figure 9: Mutual Information ratio for HWS descriptor in multivariate case

CONCLUSION

Mutual information ratio is used to compute the subset of most influent factors on a given accident outcome Y. Mutual information ratio is model independent and can be used also, before modeling, to select the most pertinent variables. It is also possible to use the selection of factors to design a model

to estimate Y given the previous selected variables. We have used Support Vector Machines to compute an empirical relation between Y and the group of factors selected by mutual information: i
The empirical relation F_S naturally depends on the data set S of observations used during learning. Using the previous factors selected by mutual information, prediction models have been elaborated using support vectors machines and gives quiet good results. A complete study of this work is available in [Mougeot & Azencott 2007].

REFERENCES

- Azencott R. (2006) Information theoretic methods and algorithms for accident causation analysis. Trace report WP7 Task 2.2, september 2006.
- Billingsley P. (1965) Ergodic Theory and Information, John Wiley.
- Brillinger D. (2004) Some data analysis using mutual information. Brazilian Journal of Probability and Statistics, 18, pp. 163-182.
- Brillinger D. & Guha A. (2006) Mutual information in the frequency domain. Journal of statistical planning and inference, 137, pp 1076-1084.
- Cover T. M. & Thomas J. A. (1991) Elements of Information Theory, John Wiley.
- Efron B, & Tibshirani R. (1993) Introduction to the bootstrap. Chapman and Hall.
- Joe, H. (1989) Relative entropy measures of multivariate dependence. J. American Statistical Association, 84, 157-164.
- GIDAS http://www.gidas.org
- Huang Y.H., Chen J.C., DeArmond S, Cigularov K and Chen P.Y. (2007) Roles of safety climate and shift work on perceived injury risk: a multi-level analysis. Accident Analysis and Prevention 39, 1088-1096.
- Joe H. (1989) Relative entropy measures of multivariate dependance. Journal of the American Statistiacl Association, Vol. 84, N° 405, pp. 157-164.
- Mougeot, M & Azencott R. (2007) Information theoretic methods for accident causation studies & prediction of injuries. European Project N° 027763-TRACE. WP7, ST 2.2.
- Mougeot, M & Azencott, R (2008) Risk factors quantification based on mutual information ratio for in depth investigation of real world accidents database (submitted).
- Pfeiffer M & Hautzinger H. (2007) Methodological problems and principles of establishing causality in traffic accident research. European Project N° 027763-TRACE. WP7, ST 2.1
- R Development Core Team (2007). R: a language and Environment for statistical computing. R foundation for statistical computing, http://www.r-project.org. ISBN 3-900051-07-0.
- Shannon C.E. (1948) A mathematical theory of communication. Bell system Tech. J., 27.