Risk factors quantification based on mutual information ratio for in depth investigation of real world accidents database.

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

June 17, 2008

[1] Modal'X \SEGMI, Université ParisX, 200 avenue de la République, 92001 Nanterre, France.

[2] Laboratoire de probabilités et modèles aléatoires, \UMR 7599, Université Denis Diderot, 75251 PARIS Cedex 05, France.

[3] Mathematics Department, University of Houston, Houston, Texas 77204-3008, United states

Corresponding author:

Mathilde Mougeot, mathilde.mougeot@u-paris10.fr,

ModalX\SEGMI, Université ParisX, 200 avenue de la République 92001 Nanterre, France.

Abstract

In Europe traffic accidents are now widely recorded in national data bases. In view of the massive amounts of accident data, the use of data mining tools is essential to sift truly relevant information, and to extract reliable relations between injuries severity and potential causation factors. We present an innovative data mining approach for in depth investigation of causation in accidents data bases. Classical statistical tools evaluate the strength of potential causal relationships by essentially linear techniques, or strongly rely on ad hoc specific models. We outline here how mutual information ratios (based on conditional entropies) contribute to rigorously quantify the influence of causation factors on accident outcome descriptors such as injury type and severity. Information theoretic methods

help to automatically select small groups of factors with high causation impact on accidents severity, with no hypothesis on underlying relations between observed variables. We successfully apply this approach to the analyze causation factors in the German In Depth Accident Study data base, which is one of the largest and most complete in depth accident survey and data collection in Europe.

Key Words: mutual information, conditional entropy, risk analysis.

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

1 Introduction

Traffic accidents are a major concern due to their economic and social costs, and above all, because accident injuries are often incapacitating or fatal. Accident injuries can be caused by a large number of factors, including human, vehicle, safety or environment factors. Traffic accidents in Europe are increasingly stored in large data bases, systematically recording many descriptive fields. In the German In Depth Accident Study (GIDAS) database, dedicated to traffic accidents in Germany, more than 800 fields are assigned to describe each accident and more than 2000 new accidents are stored each year. Intensive data mining on such data bases is clearly a major task to address. Extraction of significant accident causation factors hidden in massive databases is an important goal to improve our knowledge on traffic accidents and traffic safety. New preventive actions can also emerge from in depth investigations of accidents data, with one objective, to reduce the rate and severity of accidents.

The strength of potential causation relations between accidents descriptors and injury severity needs to be quantified, or statistically estimated. The types and severity of injuries are essentially described by a small number of indicators, refereing to the main injured body parts. But the list of potential causation factors for injuries severity is very large. Depending on the nature of the variables involved, the strength of their causal dependency is measured differently. For continuous variables, the correlation coefficient ρ^2 is a long-standing measure of statistical dependence between variables, and is often used in accidents analysis (Huang et al., 2007). For categorical data, statistical dependence is often quantified by Cramers

V , based on the χ^2 statistics. The Cramer indicator provides a zero-to-one range value comparable to ρ^2.

In modeling approaches, severities are expressed as deterministic or stochastic functions of explanatory variables such as vehicle descriptors, drivers characteristics, or road conditions. When severity is categorized by a few predetermined levels, disaggregate models have been applied to examine odds ratio (ODonnell et al, 1996). Ordered probit or logit models have been used to analyze injury severity frequencies (Abdel, 2003; Yamamoto et al., 2004; Milton et al, 2008). An important task prior or during severity modeling is to select the most relevant explanatory variables. Modeling methods are often expected to perform better when the set of explanatory variables increase. However the key explanatory variables may often constitute only a restricted subset of the available variables, and many variables may be irrelevant or even harmful to build a pertinent model. In the modeling approach, the selection of explanatory variables is mainly performed by stepwise regression associated with Bayesian Information Criteria (BIC) or Akaike Information Criteria (AIC) criteria, or by standard regression associated with Student's test to eliminate variables with no significant impact (Yau, 2004).

Dependence coefficients as well as modeling rely on specific underlying hypotheses. Correlation coefficients are known to measure only linear dependencies between variables. If variables are linked by non linear relations, then the use of correlation is definitely not the most efficient choice. During a stepwise or backward linear regression, variables are selected according to multivariate linear coefficients R^2. For categorical data, the χ^2 tests of dependence reaches its limits when the numbers of joint observations is small. For data bases with large numbers of descriptive fields, 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 (Li, 1990). For qualitative variables, in case of sparse contingency tables, the Cramers V indicator, based on χ^2 test, can also be inappropriate.

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 ratios can be computed within discrete, continuous and discrete-continuous variables (Brillinger, 2004), and provides also a powerful extension of the classical correlation and Cramers V measures. Mutual information has been successfully applied in numerous practical situations. It has been introduced to identify temporal lags for non linear models (Granger et Lin; 1994). In the spectral domain, MI has been used to infer frequency statistical dependence between seismic time series (Brillinger et Guha, 2006), but never to traffic accident analysis. In preparatory analysis of accident data prior to model building, we have validated that, because it is model independent, mutual information is a powerful tool to select the most relevant variables (Mougeot et Azencott 2007).

2 Mutual Information

Mutual information, based on conditional entropy, quantifies the relation between two random variables X and Y. For example, Y can be an injury severity descriptor and X a potential accident causation factor.

The **Entropy** measures the average quantity of information provided by the knowledge of the actual value of a random variable. For a random variable X with modalities α_i and occurrence probabilities $p_i = Probability(X = \alpha_i)$, $1 \leq i \leq m$, the entropy, H_X, is defined by:

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

with the convention, $0log(0) = 0$.

If X is deterministic, its entropy is minimal, and H_X=0: knowing the actual values taken by X in random trials brings no new information since X is constant. But if X has a uniform distribution, its entropy is maximal: $H_X = -log(m)$: all actual new values of X, which have the same probability to occur, bring new information.

For two discrete variables X and Y, with modalities α_i and β_j , and with joint probabilities $p_{ij} = Probability(X = \alpha_i, Y = \beta_j)$, $1 \leq i \leq m$, $1 \leq j \leq p$, the **joint entropy**, $H_{X,Y}$ is defined by:

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

Conditional entropy $H_{Y/X}$ quantifies the average information brought by discovering the actual value of Y when the value of X is already known, and is defined by:

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

$p_{j/i}$ denotes the conditional probability of $Y = \beta_j$ given that $X = \alpha_i$. If X and Y are independent, then $H_{Y/X} = H_Y$: knowing the value of X doesn't bring any information about the value of Y.

Mutual information, based on conditional entropy, is a measure of statistical dependence between two variables X and Y. $I_{X,Y}$ quantifies the average amount of information about the actual value of Y provided by the knowledge of the actual value of X.

$$I_{X,Y} = H_Y - H_{Y/X} \tag{4}$$

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.

$$R_{X,Y} = \frac{I_{X,Y}}{H_Y} \tag{5}$$

For two independent variables X and Y, prior knowledge of X doesn't provide any information on Y, and $R_{X,Y} = 0$. But if a deterministic functional relation exists between X and Y, then prior knowledge of X completely determines the value of Y, and the mutual information ratio is maximal: $R_{X,Y} = 1$.

Some illustrative examples are presented in figure 1. In these toy examples of joint distributions, X and Y each have 4 modalities: a, b, c, d for X and A, B, C, D for Y. The number of observations remains equal to 100 for all cases, but the proportion of deterministic coupling between modalities of X and Y ranges from case to case.

INSERT FIGURE 1

For the 1st joint distribution (left display), there is a one-to-one deterministic relation between X and Y, and MIR equals 100%. For the 2nd distribution (right display), given any modality of X, all modalities of Y are equally probable, and MIR equals 0; X modality brings no information to forecast any Y modality. For the last joint distribution (center display), most modalities of X have a one-to-one relation with a specific modality of Y, but for one X modality, the Y value remains ambiguous: in this case MIR equals 68%.

The random variables X and Y considered above take only a finite set of possible values (m modalities for X and p for Y). It is however possible to define the mutual information $I_{X,Y}$ for continuous random variables. Conditional entropy, from an actual set of continuous observations of X and Y, impose the discretization of the possible values of both variables X and Y. Consider a continuous random variable X, taking values in the space of real numbers $\mathbb{R}$, with a density function f(x) defined on the support $[A, B]$, it is possible to partition the interval $[A, B]$ into m disjoint intervals J_1, J_2, ..., J_m and to select an arbitrary point of the distribution D_k in each interval J_k ($1 \leq k \leq m$). The "discretized" random variable is defined as $U^m = D_k$ whenever the random value of X falls in J_k. The random variable U^m takes only a finite number of m values. As "m" tends to infinity, this classical discretization scheme provides, from the point of view of measure theory, a good approximating sequence of X by the sequence of random variables U^m.

The absolute entropy H_{U^m} can be computed as defined earlier since variable U^m takes only a finite number of modalities. Same methodology, used for Y, leads to define and compute MIR for continuous variables.

2.1 Estimation of Mutual Information

In operational cases, exact joint distributions of variables are naturally unknown and MIR must be estimated. Consider N independent observations of (X, Y) extracted from an accident database. Let $v_{ij}^k = 1$ when $X = \alpha_i$ and $Y = \beta_j$ for observation k and let $v_{ij}^k = 0$ otherwise. Joint probabilities can be

estimated as follow:

$$\hat{p}_{ij} = \frac{1}{N} \sum_{k=0}^{N} v_{ij}^{k}$$

The plug-in estimate of the mutual information ratio is then $\hat{I}_{X,Y} = \hat{H}_Y - \hat{H}_{Y/X}$ with

$$\hat{H}_{X,Y} = -\sum_{j=1}^{p} \sum_{i=1}^{m} \hat{p}_{ij} log(\hat{p}_{j/i}) \tag{6}$$

and

$$\hat{R}_{X,Y} = \frac{\hat{I}_{X,Y}}{\hat{H}_Y} \tag{7}$$

Theoretical results can be achieved to quantify the error between entropy and its empirical estimate using a set of observations. For a categorical variable X with m modalities and for a large number of N observations, the error between H_X and $\hat{H}_X$ can be approximated by a Gaussian random variable with zero mean and standard deviation equaled to $log(m)/\sqrt{(}mN)$ (Azencott 2006).

Here, we use a bootstrap aggregating procedure to compute a consistent estimation of $R_{X,Y}$. The Mutual information ratio is estimated using B replications of the same unit procedure. For each replication, an estimation of $R_{X,Y}$ is computed from a subset of observations (95%) chosen at random from the original data set and $\hat{R}^b_{X,Y}$ denotes the estimation of $\hat{R}_{X,Y}$ for replication b. $\hat{R}_{X,Y}$ is estimated by averaging the estimates $R^b_{X,Y}$ over the B replications.

$$\hat{R}_{X,Y} = \frac{1}{B} \sum_{b} \hat{R}^b_{X,Y} \tag{8}$$

From an asymptotic point of view, when the number of observations N tends to infinity, the bootstrap distribution tends to be equaled to the original distribution (Efron and Tibshirani 1993), and for GIDAS data, the large number of available observations ($N \simeq 11500$) leads to use the bootstrap aggregating procedure. A similar bootstrap procedure is used to compute confidence intervals. The confidence intervals are similar to the theoretical approach.

2.2 Selection of factors using mutual information ratio

Given a specific injury severity indicator Y and p potential causation factors $(X_1, \ldots, X_p)$, mutual information is used to estimate statistically the strength of the causal relationship between Y and this specific group of factors. First, we compute separately the mutual information ratios between Y and each X_j, $j \leq j \leq p$, using equation (7). To compare the individual influence levels of X_j on the severity indicator Y, the MIR coefficients $R_{X_j,Y}$ are ordered by decreasing magnitude. We denote $X_{(1)}$ the factor with the largest MIR, which has the highest predictive power for Y.

$$\hat{R}_{X_{(1)},Y} = max_{\{j\}} \left\{ R_{X_j,Y} \right\} \tag{9}$$

Each MIR coefficient lies between 0 and 100%, and evaluates the percentage of information on the value of Y which is provided by X .

Mutual information can also be computed for multivariate factors (Joe, 1989). Let $X = (X_{i_1}, \ldots, X_{i_k})$ be a multivariate variable regrouping k factors ($k \leq p$). The MIR of Y with respect to X is computed as above using natural extensions of equations (6) and (7). To select a group of k factors having the highest joint predictive power for Y, we proceed as above for single factors, and hence select the group G of k factors with the highest MIR ratio R(G,Y). Among all groups of k factors , the group G best explains the Y values. This method provides also an efficient and rigorous way of constructing increasing hierarchies of causation factors for a given severity indicator Y .

2.3 MIR and analyze of dependence for accident data

Mutual information ratio is a non parametric measure of association between at least two variables, Y and X. It can be applied to symbolic data (categories) as well as numerical data. In the bivariate case, mutual information is the Kullbak-Liebler distance between the joint distribution of (X, Y) and the product of its marginal X, Y (Brillinger, 2004). As MI measures the general dependence, the correlation function, ($\rho_{X,Y}$ equation 10), is restricted to measure the linear dependence between both variables and is restricted to numerical observations (Li, 1990) .

$$\rho_{X,Y} = \frac{\sum_{i,j} p_{i,j}(X_i - \bar{X})(Y_j - \bar{Y})}{\sigma_X \sigma_Y} \tag{10}$$

where σ_X and σ_Y are the standard deviation of X and Y.

MI is, in particular, invariant under strictly monotone transformation of the initial random variables. If two different strictly monotone functions are applied independently on both variables, the new MIR computed on the transformed variables doesn't change. On the opposite for the correlation function, as the probabilities of occurrence of observations $p_{i,j}$ are weighted by the values of the variables X_i, Y_j, if we consider two variables with a correlation coefficient equaled to 1 and if a strictly monotone transformation (not linear) is applied on one or both variable then the new correlation coefficient changes. However, both functions, Mutual Information or Correlation, can be used for *ranking variables* considering one target variable as an injury severity descriptor Y and p potential explanatory variables $X_1 \dots X_p$.

On a complementary point of view, **Association rules mining** focuses on *ranking attributes* for specific modalities of X and Y and have been applied to the study of traffic accident data (Wang, et al. 2005; Marukatat 2006). Given a specific modality β_j of the target variable Y, it is possible to evaluate the occurrence of $Y = \beta_j$, given modality α_i of X by the conditional probability: $Probability(Y = \beta_j / X = \alpha_i) = \frac{p_{i,j}}{p_{i+}}$. Association rule mining compares the conditional probability to the probability of occurrence of $Y = \beta_j$ without any condition of X: $Probability(Y = \beta_j) = p_{+j}$ and computes γ_{ij}, ratio defines as :

$$\gamma_{i,j} = \frac{Probability(Y = \beta_j / X = \alpha_i)}{Probability(Y = \beta_j)} = \frac{p_{i,j}}{p_{i+}p_{+j}} \tag{11}$$

where p_{i+}, p_{+j} are the marginal probabilities of X and Y for modalities α_i and β_j.

This association rule coefficient ranges between 0 and infinity. $\gamma_{i,j} > 1$ means that the probability of occurrence of modality β_j of Y given modality α_i of X is greater compared to the marginal distribution (than without any condition on X): there is then an attraction between the 2 modalities $Y = \beta_j$ and $X = \alpha_i$. On the opposite, for $\gamma_{ij} < 1$, a repulsion between both modalities can be showed: an occurrence of $Y = \beta_j$ tends to annihilate an occurrence of $X = \alpha_i$. This coefficient is, in particular, involved in

the computation of the statistical χ^2 distribution. When a dependence relation is observed and tested between two variables, the *attraction-repulsion* ratio helps to localize the dependency through the joint modalities of the two variables.

Classification And Regression Trees (CART) are a non parametric methodology to point out the dependencies between variables (Breiman et al., 1998). With the capacity to automatically search for the best explanatory variable and split-point to achieve the best fit, CART have been successfully applied to the analysis of traffic injury severity. Karlafatis et al., (2002) applied hierarchical tree-based regression to analyze the effects of road geometry and traffic characteristics on accident rates for rural roads. Chang et al., (2006) has applied a CART model to established the relationship between injury severity and accident variables. CART identify group of hierarchical factors by partitioning the feature space into a set of rectangles. The construction of a tree is data-driven and based on local optimization. CART becomes nowadays a very popular method which can produce nice recursive binary trees. When analyzing the tree, a same variable can appear many times from top to bottom of the tree, the result is then not so easily readable and it it quiet difficult to evaluate globally the impact of a variable.

Compared to CART, Mutual information ratio provides a list of hierarchical factors which best explains a target. The construction of the list is global and do not depend on local optimization. Moreover, risk factors and confidence intervals are available for each factor or each group of factors.

2.4 Modeling of injuries severity

Since mutual information ratios are model independent, they can be used, prior to modeling, to select the most relevant group G of explanatory variables to predict a given accident outcome Y. One can then construct a model to predict Y outcome given the group G of selected variables.

$$Y = F_S(X_{(1)}, ..., X_{(k)}) \tag{12}$$

The empirical relation F_S naturally depends on the data set S of observations used during learning. As mutual information is a variables selection tool, given the group of key causation factors selected by mutual information, we have used Support Vector Machines to compute empirical predictive relations

between using a data set of training examples form GIDAS (Mougeot et Azencott 2006).

3 Accidents 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 German In-Depth Accident Study (GIDAS). The accidents units, composed of a team of experts on duty, respond to any traffic accident with injuries in the region of Hanover. In a detailed investigation, the team acquires both technical data from the accident site and medical/injury data from the people involved. In the areas of Hanover and Dresden, personal injury traffic accidents are systematically reported by the police and the fire department stations. Annually, approximately 2,000 traffic accidents are recorded in this way and the information is stored in an historical database. 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 (cities and administrative districts of Hanover and Dresden). The accidents are recorded by each team daily with alternating shift times so that a uniform distribution between day and night and between the different days of the week is ensured.

Standardized classification systems are used to describe the severity of injuries, such as AIS (Abbreviated Injury Scale). Each accident is analyzed in detail and the motions of the vehicles and occupants reconstructed. The geographical distribution of the investigation areas correlates well to the one of Federal Republic of Germany as a whole. In both, approximately 90% of the area can be regarded as rural and 10% urban, so that the distribution between rural and urban built-up areas is similar. Since collisions processes are generally dependent on technical background conditions and the resulting injuries often affected by these conditions, GIDAS investigations can be used for most aspects of passive and active safety.

The GIDAS database is now the largest and most complete In-Depth accident survey and data collection in Europe. The number of available observations in the GIDAS database was, at the end of 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).

4 Applications to risk factors quantification

In the GIDAS database, most variables are qualitative, we hence have a natural situation where classical correlation analysis may be of limited use, and information theoretic methods based on conditional entropy computation offer a more rigorous tool to explore association or causation relations between variables. We have applied to GIDAS data the MIR methodology outlined above, with, at the end of 2006, 14000 observations, described by more than 800 fields. All vehicles, and people involved in a crash data (when at least an injured people can be found) are stored in the data base. A preliminary filtering treatment has first been applied to the whole database, to eliminate inappropriate values (Mougeot and Azencott 2007). For our whole study, tests and analyzes have been implemented by programs we developed using the R statistical programming software [R development Core Team]. No specific R toolboxes has been used for this application. All the code and functions to compute the theoretical coefficients have been programmed using R standard language.

4.1 Injuries severity indicators

We have, for the moment, focused our exploratory causation analysis on three indicators of injury severity for different body parts (Y variable) : Maximum Injuries Severity (MAIS), Head Injury Severity(HWS) and leg injuries Severity (AISBEIN).

4.1.1 Maximum Injuries Severity (MAIS)

In the GIDAS database, MAIS values fall into 7 categories $0 \dots 6$, corresponding to 7 possible values for the maximum severity of injuries. MAIS0 corresponds to "non injured" accidents. MAIS1-MAIS2 corresponds to "slightly injured" accidents and MAIS3+ to "fatally injured".

INSERT FIGURE 2

We regroup the original 7 modalities of MAIS into 2 categories in order to analyze wether accidents led to injuries or not. The 2 labels *Safe* and *Injured* respectively denote accidents with no injury (MAIS tag = MAIS0) and accidents with some injuries ($MAIStag \geq 1$). In the database, "no injury" accidents have frequency 60%, and "minor injuries" accidents ($MAIStag \leq 1$) have frequency 74%. Histograms are built with 11586 observations.

4.1.2 Head Injuries Severity(HWS)

In the GIDAS database, Head Injuries Severity is recorded by the variable HWS, which has 7 modalities, as defined for MAIS.

INSERT FIGURE 3

Figure 3 shows that a large majority of accidents (80%) lead to no head injury. Histograms are built with 11586 observations. As above, we split the 7 modalities of HWS into 2 broad categories labeled *Safe* and *Injured*.

4.1.3 Leg Injuries Severity(AISBEIN)

In the GIDAS database, Leg Injuries Severity is recorded by the variable AISBEIN, which has 7 modalities, as defined for MAIS and HWS.

INSERT FIGURE 4

Figure 4 shows that a large majority of accidents (80%) lead to no leg injury. Histograms are also built with 11586 observations. As above, we split the 7 modalities of AISBEIN into 2 broad categories labeled *Safe* and *Injured*.

4.2 Potential causation factors for injuries severity

In this study, our target list of potential causations factors was prepared according to existing expert judgments communicated by the German BAST institute. A key objective of this study was to focus on a target list of potential causation factors for injuries severity, to estimate and compare the causation strengths between potential causation factors and the severity descriptors, and to determine which combination of causation factors has the highest power to predict injuries severity. This exploratory study was restricted to the severity indicators MAIS, HWS and AISBEIN, to better evaluate the practical impact of our mutual information approach. We present in Table 1, our initial target list of potential causation factors for the 2 injuries severity indicators MAIS and HWS. 15 factors have been selected for this study: 13 factors are categorial variables and the 2 remaining factors (COLLSPEED and CARGE) are continuous and have been divided in 10 classes for the computation of MIR (figure 12).

INSERT TABLE 1

5 Results

In this section, mutual information ratios (MIR) are computed to estimate the causation strengths between potential factors and accident outcome descriptors (MAIS, HWS, AISBEIN). Each MIR is computed using more than 8000 observations, depending on the proportion of missing values for the studied variables. Each specific MIR involves only a precise small set S of variables, and to compute this MIR coefficient, we temporarily eliminate all records having missing values for some of the variables in S. First, the MIR coefficients are separately estimated for each potential causation factor, and then ordered. For each one of the 2 variables (MAIS, HWS, AISBEIN), we then successively determine by which group of multivariate factors we can best explain it.

5.1 MAIS

The MIR coefficients are first estimated using the 7 original modalities of MAIS, and then estimated again using only the coarser binary categories (*Safe* or *Injured* for MAIS. These MIR coefficients evaluate how well MAIS is explained by each potential causation factor in the BAST target list. We then sort these coefficients by decreasing order of magnitude (Figure 5).

INSERT TABLE 5

We present the results as follows (Figure 5). Fix an outcome descriptor (such as MAIS), the MIR coefficient computed for each single factor is represented by the length of a horizontal bar. The tag name of the corresponding factor is displayed on the left and Table 1 gives the list of all these tag names. The number of joint observations used for computing the MIR is displayed on the right. At the right end of each bar, we display a confidence interval for the MIR value, computed by bootstrap at a 95% confidence level. All MIR coefficients lie between 0 and 100%.

For the MAIS indicator, this analysis shows that the most influent factor explaining maximum injury severity is the OPPONENT type, with a MIR around 23%. When the 7 initial modalities are regrouped into binary classes (Safe versus Injuries), this feature is even sharper and its MIR value increases to 31%. The accident KIND appears in 2nd position (MIR = 13% , and MIR = 16% for binary modes), and the accident TYPE comes in 3rd position (MIR = 10% , and MIR = 12.5% for binary modes). All MIR coefficients increase when computed for the coarser binary distribution. SPEED of collision, PLACE, and SPEED LIMIT obtain similar MIR coefficients.

The SEATBELT factor appears in the middle of the list with a small MIR (1.95%). At first sight, this is surprising since SEATBELT usage is considered to be an important factor affecting injury severity of vehicle traffic accidents. Recall that today, drivers and passengers are required by law to use their seat belt, so that 97% of the observations in GIDAS correspond to the use of seat-belts (Figure 6). So the MIR coefficient is here overwhelmingly determined by the cases where seatbelt is used, and hence reflects

only partially the intrinsic risk associated to the absence of seatbelt.

To focus on the severity of accidents due non seatbelt usage, we have artificially selected a random set of GIDAS data with equal proportions of "seatbelt use" and "no seatbelt use" (Figure 6). The small proportion (3%) of accidents records with non usage of seatbelt have all been retained, and have been completed with an equal proportion of observations, taken at random among the numerous accident records with corresponding to seatbelt usage. To obtain a robust estimation of the MIR, this procedure has been replicated 20 times, and the MIR has been averaged over all replications. For this specific mixture of observations, better suited to evaluate the impact of the SEATBELT factor, the MIR increases from 1.95% to 14%, which is a quite high value, corresponding to a 2nd position in the ranked list of causation factors. SEATBELT usage remains an important causation factor directly linked to injury severity. Since only a very small minority of drivers do not wear seatbelts, the proportion of accidents where this factor becomes really active remains extremely small.

INSERT FIGURE 6

ROLLOVER accidents are quite rare, and their impact on MAIS is high (MIR 5.8%), but the intrinsic risk associated to ROLLOVER is much higher. To compute the severity impact of ROLLOVER, we use the same procedure as for SEATBELT, and select an artificial random sample of accidents, with 50% of ROLLOVER cases. We observe that the MIR coefficients increases to 27%, which confirms the exceptional gravity of rollover accidents.

The GENDER variable has fairly small MIR, and hence does not seem to have a strong impact on MAIS.

Multivariate analysis is then conducted to analyze for a given number of explanatory variables, which group of factors has the highest mutual information ratio with MAIS, and hence best explains Maximum Injury Severity. The following graph presents, for MAIS, the highest MIR feasible as function of the number of potential causation factors (Figure 7).

INSERT FIGURE 7

For instance, the 3rd column indicates that the group of 3 factors (OPPONENT, Collision SPEED and Accident KIND) has a joint MIR of 38%; this group has the highest predictive power for all groups of 3 factors. It is interesting to observe that, for the single factor analysis, OPPONENT, Accident KIND and Accident TYPE were respectively in 1st, 2nd and 3rd position, regarding the association strength level (figure 5). In the multivariate analysis, Collision SPEED, which was in 4th position for the single factor analysis, replaces Accident TYPE in the most predictive combination of 3 factors. This is essentially due to the sizable redundancy between accident KIND and TYPE, as can be seen from their pairwise MIR which is equal to 52The MIR coefficients estimated for MAIS confirmed here by objective computation the knowledge of BAST Experts about the main injury severity causation factors in accidents.

5.2 HWS

We now present our analysis of head injuries causation factors. Just as for MAIS, the Mutual Information Ratios are computed between HWS and the potential causation factors listed in Table 1.

INSERT FIGURE 8

As above, the MIR coefficients estimated for HWS are sharper when computed for a binary distribution as for the original distribution (Figure 8). The OPPONENT type is, as for MAIS, the most influential factor explaining head injuries severity however the causation strength is smaller (12,5% as compared to 23%). The same holds true for the factors Accident KIND and TYPE which are again placed 2nd and 3rd. GUILTY, which records the driver's responsibility, is now at 4th place. The driver's GENDER becomes a quite important factor for head injuries, indicating that women are more vulnerable than men from this point of view. The mainly damaged part of the car (DAMAGE) comes also into play, probably reflecting that rear end collisions play a high role in the occurrence of severe head injuries.

Multivariate analysis is then conducted, as above, to select which group of factors has the highest mutual information ratio, and hence best explains head injury severity. Results are presented Figure 9.

INSERT FIGURE 9

The two factors, which jointly best explain head injuries, are: OPPONENT type and GENDER. Observe that GENDER, which as a single factor influencing HWS was in 5th position, is now the factor, which in combination with OPPONENT type, best explains head injuries (considering more than 9.000 observations of GIDAS database). This result confirm a fact known to experts, namely that, in traffic accidents, women are more vulnerable than men for head and neck injuries.

5.3 AISBEIN

We now present our analysis of leg injuries causation factors. Just as for MAIS and HWS, the MIR are computed between AISBEIN and the potential causation factors listed in Table 1.

INSERT FIGURE 10

As for MAIS and HWS, the MIR coefficients are sharper when computed for a binary distribution as for the original distribution (figure 10). Comparing MAIS, HWS and AISBEIN analyses, we observe that the same subsets of factors are associated with the highest MIR ratio. The type of OPPONENT, the TYPE and KIND of accident, as defined in GIDAS, are the strongest factors.

INSERT FIGURE 11

For the multivariate analysis, we observe the same analogy. The group of tree factors is the same for MAIS and AISBEIN. GENDER seems to have a strongest influence on head injuries, but less for legs injuries.

6 Conclusion

In this study, major causation factors impacting injuries severity have been identified and ranked, corresponding causation strengths have been estimated, by analysis of the GIDAS accidents database, which offers one of the largest accident survey and data collection in Europe. Our Mutual Information approach has proved to be quite efficient for selecting and ranking potential causation factors. The MIR coefficients naturally depend on the histogram of factor modalities, and we show how factors for which a single modality is overwhelmingly represented, may get underestimated causation strengths and association strength values. We have also shown how prior adequate random re sampling of the data enables the MIR coefficient to correctly estimate causation strength even when one single modality is omnipresent. This feature was illustrated for both seat belt and rollover factors, and confirms that rollover accidents as well as the non usage of seat belts lead to serious injuries (even if the actual proportion of accidents in which these factors were active is very small).

Mutual information ratios offer then a wide range of possibilities to study causation links between variables having continuous distribution or finite sets of modalities.

Due to the major probabilistic properties of MIR, these mutual information ratios are very efficient to detect non linear causation links. Since they are radically "model" independent, the MIR coefficients and ranking can be used for variables selection prior to statistical modeling. Using as inputs the causation factors selected by mutual information ratios, prediction models have then been constructed using support vectors machines , with good performance. [Mougeot and Azencott 2007].

References

[1] Abdel-Aty, M., 2003 Analysis of driver injury levels at multiple locations using ordered probit models. Journal of safety research 34, pp 597-603.

[2] Azencott, R., 2006 Information theoretic methods and algorithms for accident causation analysis. European Trace report WP7 Task 2.2, september 2006.

[3] Billingsley, P., 1965 Ergodic Theory and Information, John Wiley.

[4] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1998 Classification and Regression Trees. Chapman Hall/CRC.

[5] Brillinger, D., 2004 Some data analysis using mutual information. Brazilian Journal of Probability and Statistics, 18, pp. 163-182.

[6] Brillinger, D., Guha, A., 2006 Mutual information in the frequency domain. Journal of statistical planning and inference, 137, pp 1076-1084.

[7] Chang, L.Y., Wang, H.W., 2006 Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accident Analysis and Prevention 38, 1019-1027.

[8] Cover, T. M., Thomas, J. A., 1991 Elements of Information Theory, John Wiley.

[9] Cox, D.R., Wermuth, N. 2001 Some statistical aspects of causality. In: European Sociological review, 17, pp65-74. ISSN 1437-4110.

[10] Cox, D.R., Wermuth, N., 2004 Causality: a statistical view. IN: International Statistical review, 72, pp 285-305.

[11] Efron, B., Tibshirani, R., 1993 Introduction to the bootstrap.Chapman and Hall.

[12] Joe, H., 1989 Relative entropy measures of multivariate dependence. J. American Statistical Association, 84, 157-164.

[13] Geman, D., Jedynak, B., 2000 Model-based Classification Trees. Technical document.

[14] GIDAS, http://www.gidas.org

[15] Granger, C.W.L., Lin, J.L., 1994 Using the mutual information coefficient to identify lags in nonlinear models. J. Time Series Anal. 15, 371-384.

[16] Huang, Y.H., Chen J.C., DeArmond, S. ,Cigularov, K., 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.

[17] Joe, H., 1989 Relative entropy measures of multivariate dependence. Journal of the American Statistiacl Association, Vol. 84, N 405, pp. 157-164.

[18] Karlaftis, M.G., Golias, I., 2002 Effects of road geometry and traffic volumes on rural roadwayaccident rates. Accident Analysis and Prevention 34, 357-365.

[19] Li, W., 1990 Mutual information functions versus correlation functions. Journal of Statistiacl physics. 60, 5/6.

[20] Marukatat, R., 2006 Structure-Based Rule Selection Framework for Association Rule Mining of Traffic Accident Data. CIS 2006: 231-239.

[21] Milton, J.C., Shankar, V.N., Mannering, F.L., 2008 Highway accident severities and the mixed logit model: an explanatory empirical analysis. Accident Analysis and Prevention, Volume 40, Issue 1, Pages 260-266.

[22] Mougeot, M., Azencott, R., 2007 Information theoretic methods for accident causation studies and prediction of injuries. European Project N 027763-TRACE. WP7, ST 2.2.

[23] O'Donnell, C.J., Connor, D.H., 1996 Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice. Accident Analysis and Prevention. 28, 6, pp. 379-753.

[24] Pastor, C., 2006 Self Organizing maps for accident causation analysis European Project N 027763-TRACE. WP7, ST 2.2.

[25] 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

[26] 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.

[27] Shannon, C.E., 1948 A mathematical theory of communication. Bell system Tech. J., 27.

[28] Wang, H., Parrish, A., Smith, R.K., Vrbsky S., 2005 Improved variable and value ranking techniques for mining categorical traffic accident data. Expert Systems with Applications 29, 795-806.

[29] Yamamoto, T., Shankar, V., 2004 Bivariate orderd-response probit model of driver's and passager's injury severities in collisions with fixed objects. Accident Analysis and Prevention, 36, pp. 369-876.

[30] Yau, K.K.W, 2004 Risk factors affecting the severity of single vehicle traffic accidents in Hong Kong. Accident Analysis and Prevention 36, 333-340.

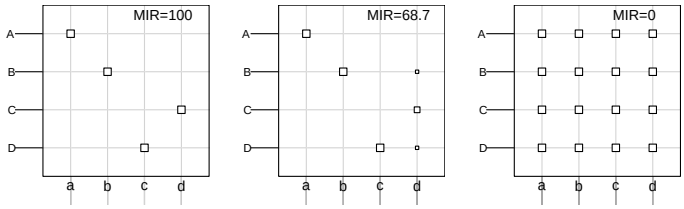

Figure 1: Mutual Information Ratio for some joint distributions. From left to right: $R_{X,Y} = 100\%$; $R_{X,Y} = 68\%$; $R_{X,Y} = 0\%$.

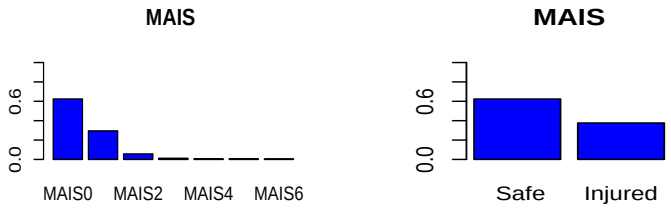

Figure 2: MAIS distribution for GIDAS data. Original and binary distribution of data.

Variable	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/ anfhrt/
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..
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	continuous
DAMAGE	Main damage to the car	(7) Front Right Side Bottom
ROLLOVER	Rollover (yes/no)	(2) yes/no

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

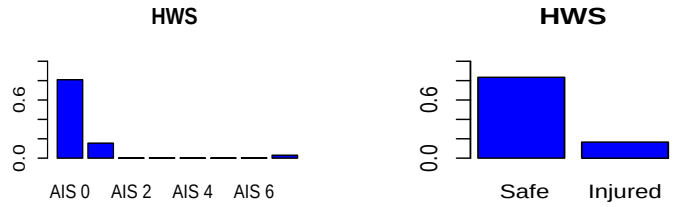

Figure 3: Head injuries distribution for GIDAS data. Original and binary distribution of data.

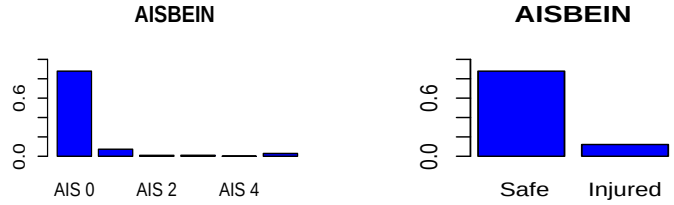

Figure 4: Leg injuries distribution for GIDAS data. Original and binary distribution of data.

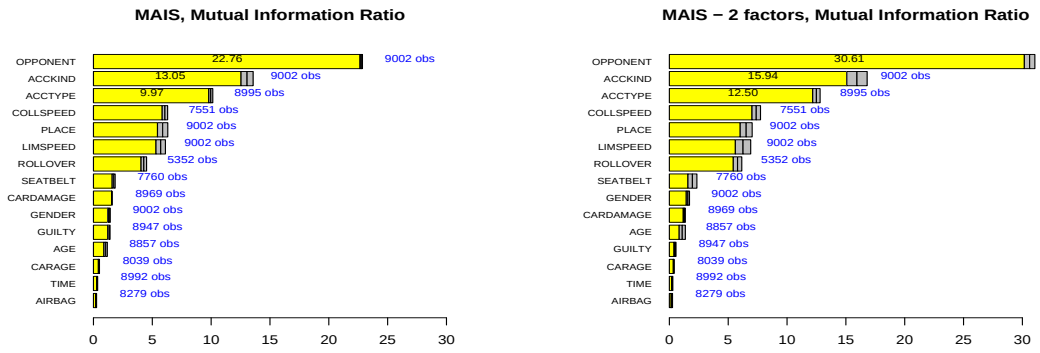

Figure 5: MIR for MAIS. Left: initial distribution. Right: MIR computed fo binary distribution of *safe* and *injured* accidents.

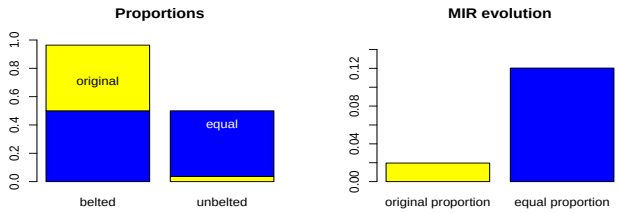

Figure 6: Seatbelt usage for GIDAS original data (yellow) or equaled proportion (blue). Corresponding Impact on MIR.

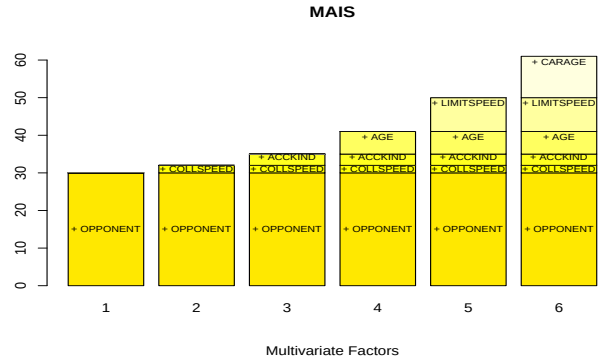

Figure 7: Multivariate MIR for MAIS descriptor.

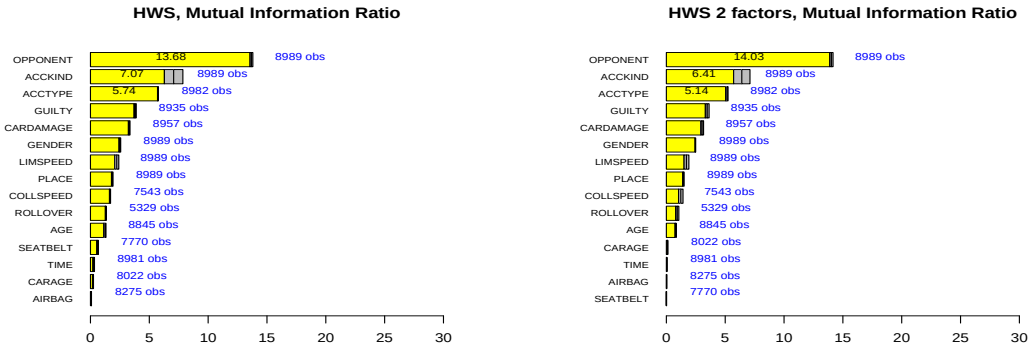

Figure 8: MIR for head injuries. Original and binary distribution.

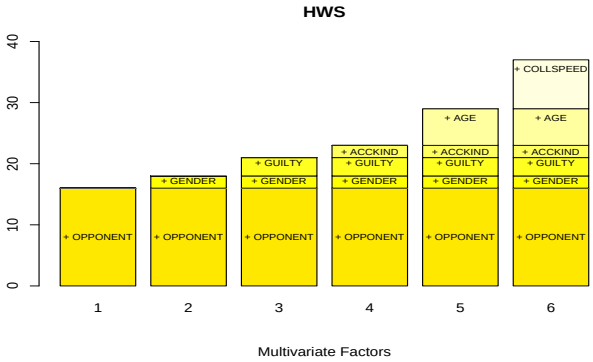

Figure 9: Multivariate Mutual Information Ratio for Head descriptor.

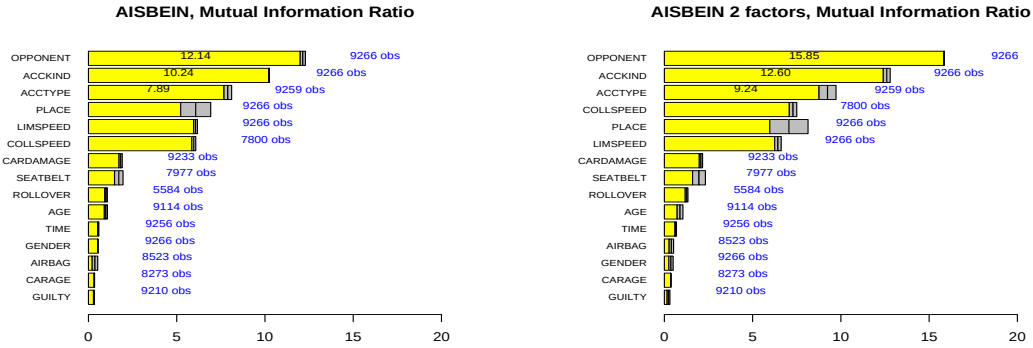

Figure 10: MIR for leg injuries. Original and binary distribution.

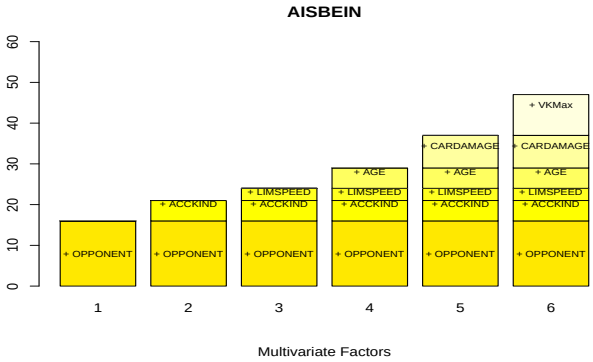

Figure 11: Multivariate Mutual Information Ratio for Leg descriptor.

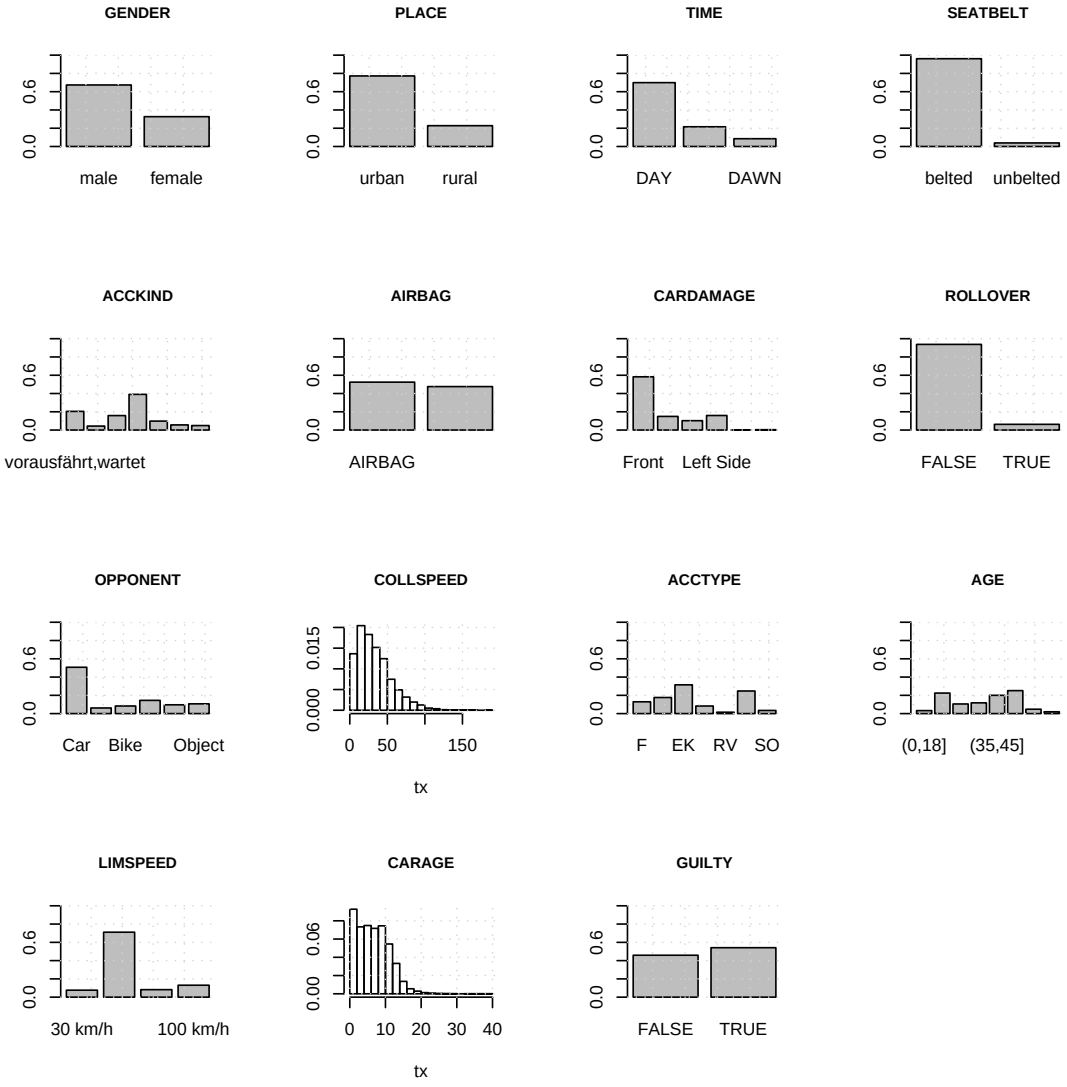

Figure 12: Histogram of potential association factors for MAIS, HWS and AISBEIN descriptor.