Injury Severity Analysis based on mutual information for in depth investigation of accident database.

Mathilde Mougeot[1], Robert Azencott[2]

July 9, 2010

[1] Université Paris-Diderot, CNRS LPMA, 175 rue du Chevaleret, 75013 Paris, France.

[2] Department of Mathematics, University of Houston, Houston, Texas 77204-3008, USA.

Corresponding author:
Mathilde Mougeot, mathilde.mougeot@univ-paris-diderot.fr

Abstract

In Europe traffic accidents are now widely recorded in national databases. 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 relationships between injury severity and potential causation factors. We present an innovative data mining approach for in depth investigation of causation in accidents databases. 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 relationships between observed variables. We successfully apply this approach to analyze causation factors in the German In Depth Accident Study database, 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 Accident Causation 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 result from a large number of causes, including human, vehicle, safety or environment factors. Informations on traffic accidents in Europe are todays stored in large databases that systematically record 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 databases is clearly a major task to address. Extraction of significant injury causation factors hidden in massive databases is an important goal for improving our knowledge of traffic accidents and traffic safety. New preventive actions can emerge from in depth investigations of accidents data, with the objective to reduce the rate and severity of accidents (Hautzinger et al, 2008).

In accident databases, the type and severity of injuries are essentially described by a small number of indicators, referring to injured body parts. But the list of potential causation factors for injury severity is very large. The link between accident descriptors, on one side, and injury severity, on the other side, needs to be quantified, or statistically estimated. Ordered probit or logit models have been used to analyze injury severity frequencies (Abdel, 2003; Yamamoto et al., 2004; Milton et al, 2008). 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), or by standard regression associated with Student's test to eliminate variables with no significant impact (Yau, 2004). Classification and regression trees have also been applied to establish a relationship between injury severity and accident descriptors (Chang et al.,2006) and to analyze the effects of road geometry and traffic characteristics on accident rates for rural roads (Karlafatis et al., 2002).

Depending on the nature of the variables involved, the strength of the dependency between accident descriptors and injury severity is measured differently. For continuous variables, the correlation coefficient ρ^2 is a long-standing measure of statistical dependency between two variables, and is often used in accidents analysis (Huang et al., 2007). For categorical data, statistical dependency is often quantified by Cramer $'$s V , based on the χ^2 statistics. The Cramer indicator provides a zero-to-one range value comparable to ρ^2. Moreover, dependency 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 relationships, 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 databases with a large number 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, the Cramer $'$s V indicator, based on χ^2 test, is also inappropriate, in the case of sparse contingency tables.

Mutual information (MI), introduced by Shannon (1949) is a measure of statistical dependency that is able to catch complex relationships between variables, even in case of non linear dependency (Billingsley, 1965; Cover et al., 1991). Mutual information ratios can be computed for discrete, continuous and discrete-continuous variables (Brillinger, 2004). MI provides a powerful extension of the classical correlation coefficient and of Cramer $'$s V measure. Mutual information has been used to evaluate the link between different kinds of variables. For instance, Granger et Lin (1994) have identify temporal lags for non-linear models and, in the spectral domain, MI has been used to infer frequency statistical dependency between seismic time series (Brillinger et Guha, 2006).

In this paper, we show how this method can successfully be used in the domain of accidentology in order to select the most informative variables that explain injury severity in a large dataset, without constraints on variable nature or linearity.

2 Mutual Information

Mutual information, based on conditional entropy, quantifies the relationships between two random variables X and Y. For example, let consider Y 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 example, for a variable that takes two modalities ($m = 2$), if only one modality is observed for all observations ($p_1 = 0$ or $p_1 = 1$), the entropy is minimum and equals 0. It means that no variability can be observed in the set of data. On the opposite, if both modalities are observed in equal proportions ($p_1 = p_2 = 0.5$), then the entropy is maximum and equals 1.

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 dependency 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 dependency 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 relationship exists between X and Y, then prior knowledge of X completely determines the value of Y, and the mutual information ratio is maximal: $R_{X,Y} = 1$. 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-Leibler distance between the joint distribution of (X, Y) and the product of its marginal X, Y (Brillinger, 2004).

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 relationship 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 relationship 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 set of real numbers $\mathbb{R}$, with a density function f(x) defined on the support $[A, B]$. The interval $[A, B]$ is split into m disjoint intervals J_1, J_2, …, J_m and an arbitrary point D_k is selected 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 above since U^m has a finite number of modalities. A similar discretization scheme can be applied to an arbitrary pair X, Y of continuous variables with a joint density, and this approach provides, as m tends to infinity, an explicit formula for the MIR of (X, Y).

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:

$$\hat{p}_{ij} = \frac{1}{N} \sum_{k=0}^{N} v_{ij}^k \tag{6}$$

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

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

and

$$\hat{R}_{X,Y} = \frac{\hat{I}_{X,Y}}{\hat{H}_Y} \tag{8}$$

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

2.2 Factor selection using mutual information ratio

Given a specific injury severity indicator Y and p potential causation factors $(X_1, \ldots, X_p)$, mutual information can be used to estimate and statistically compare the strength of the causal relationship between Y and the factors of the group. Mutual information ratios are first computed between Y and all X_j, $1 \leq j \leq p$, using equation (7). Each MIR coefficient lies between 0 and 100%, and evaluates the percentage of information on the value of Y which is provided by X .Then, 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. $X_{(1)}$ denotes the factor with the largest MIR, associated to the highest predictive power for Y.

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

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 G_k of k factors having the highest joint predictive power for Y, we proceed as above for single factors, and hence select the group G_k of k factors with the highest MIR ratio $R(G_k, Y)$. Among all groups of k factors , the group G_k best explains the Y values. Finding the best group of k factors among p factors is generally computationally infeasible. Hence, we proceed with a greedy algorithm. The following pseudo code details the algorithm for multivariate variables selection based on MIR (table 1).

This method provides also an efficient and rigorous way of constructing increasing hierarchies of causation factors for a given severity indicator Y. This method is applied to GIDAS data to extract group of factors with a high predictive power on injury severity.

2.3 MI and dependance analysis for accident data

Mutual information ratio is a non parametric measure of association between at least two variables, Y and X. Let us sketch some alternative approaches.

The MIR quantifies the level of general non linear functional dependency between X and Y, while the usual **correlation** $\rho_{X,Y}$ (equation 10) only quantifies the level of linear dependency between X and Y. Moreover correlations have no intrinsic meaning for categorical observations (Li, 1990). Recall that

```
Notations:
$Y$ is the target variable, $X_1, ....X_p$ the p factors.

Initialisation:
$Z_0 = \{\}$; $G_0 = \{\}$; $J_0 = 1...p$;
choose $K \in \{1..p\}$; K size of the multivariate group of selected factors

Algorithm:
for k=1 to $K$ do
    $j_0 = ArgMax_{j \in J_{k-1}} MIR(Y, U_k(j))$ with $U_k(j) = [Z_{k-1}; X_j]$;
    $G_k = [G_{k-1}; j_0]$;
    $Z_k = [Z_{k-1}; X_{j_0}]$;
    $J_k = J_{k-1} - \{j_0\}$;
end
$G_K$ is the multivariate group of size K with high predictive power on Y.
```

Table 1: g

reedy algorithm for to select multi variate factors using MIR criteria.

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

where σ_X and σ_Y are the standard deviation of X and Y. Note that arbitrary strictly monotone transformations of the variables X and Y do not affect the existence of a functional dependency between X and Y. Hence the fact that MIR(X,Y) is invariant under such monotone transformations of X and Y confers to MIR a strong intrinsic robustness. Indeed the correlation $\rho_{X,Y}$ can be arbitrarily modified by such monotone transformations, since the X_i, Y_j values occur explicitly in its definition. For instance a perfect correlation $\rho_{X,Y} = 1$ can be lowered arbitrarily by non linear transformation of one of the variables. Both Mutual Information ratios and Correlations can a priori be used for *ordering explanatory variables* by decreasing influence levels when one seeks to explain an injury severity descriptor Y by p potential explanatory variables $X_1 \ldots X_p$. The MIR approach seems nevertheless more generic and more robust. One major advantage is that MI can be applied to symbolic data (categories) as well as numerical data.

CART (Classification And Regression Tree) is a non parametric methodology to point out dependencies between variables (Breiman et al., 1998) which has been applied to analyze traffic injury severity. Karlafatis et al. (2002) used tree-based regression to analyze the effects of road geometry and traffic characteristics on accident rates for rural roads. Chang et al. (2006) have applied a CART model to establish a relationship between injury severity and accident descriptors. The algorithm identifies the most relevant factors and the associated categories (or the associated threshold for continuous variable) by partitioning the feature space of explanatory variables into a set of rectangles, and provides recursive binary trees.The construction of the tree is data-driven and based on local optimization. It should be note that following this methodology, the same explanatory variable X can appear many times in different position within the tree due to the possible split of the set of defined categories (see hereafter for illustration on accident data). Compared to CART, MIR generates a hierarchical list of ranking factors which best explains a target variable. Each factor appears only ones in the list. The construction of the list is global for each variable.

3 Accidents database

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). 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.

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. 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 factor 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 relationships 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 database. A preliminary filtering treatment has first been applied to the whole database, to eliminate inappropriate values (Mougeot et al., 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]. All the code and functions to compute the theoretical coefficients have been programmed using R standard language. No specific R toolboxes has been used for this application.

4.1 Injury severity indicators

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

4.1.1 Maximum Injury 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.

INSERT FIGURE 2

In order to analyze whether accidents led to severe, light of non injuries, the initial 7 modalities of MAIS have been regroup into 3 categories . The 3 labels *Safe*, *Slightly Injured* and *Severe Injured* denote respectively accidents with no injury (MAIS tag = MAIS0), accidents with some minor injuries (MAIS tag $\in \{1, 2\}$), and accidents with severe injuries (MAIS tag ≥ 3). In the database, a frequency of 60% is observed for "no injury" accidents , and a frequency of 74% for "slight injury" accidents (MAIS tag ≤ 1). Histograms are built with 11586 observations.

4.1.2 Neck Injury Severity(HWS)

In the GIDAS database, the variable HWS focus on Neck Injury Severity and has 7 modalities, as defined for MAIS.

INSERT FIGURE 3

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

4.1.3 Leg Injury Severity(AISBEIN)

In the GIDAS database, Leg Injury 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 3 broad categories labeled *Safe*, *Slightly Injured* and *Severe Injured* for the legs.

4.2 Potential causation factors for injury severity

A key objective of this study was to focus on a target list of potential causation factors for injury severity, to estimate and compare the causation strengths between potential causation factors and the injury severity descriptors, and to determine which combination of causation factors has the highest power to predict injury severity. A list of potential causation factors was first prepared by the German BAST institute. Groups of factors describing as collision, environment, human, safety, site and vehicle characteristics were chosen (table 2).

INSERT TABLE 2

Different type of factors are observed as continuous, discret or nominal (table 3).

INSERT TABLE 3

The *collision* is described by six factors : the initial speed of the collision (continuous), the kind of opponent (6 categories), the main damage to the car (7 catagories), the type (7 categories) and kind of accident (10 categories) and if it's a rollover accident (binary). *Environmental factors* take into account: the speed limit (17 categories), the place (binary) and the time of the accident (3 categories). *Human effects* are analyzed through following variables: the age of the driver (8 categories), its gender (binary), and its guiltiness (binary). The *Vehicule* is described by its age (continuous) and the airbag equipment (binary). *Safety* is described by the use (or not) of the seatbelt (binary variable).

Table 4 presents the initial target list of potential causation factors for the injury severity indicators MAIS, HWS and AISBEIN.

INSERT TABLE 4

This exploratory study was restricted to three injury severity indicators MAIS, HWS and AISBEIN, to better evaluate the practical impact of the mutual information approach. 15 factors have been selected for this study : 13 factors are categorial variables and the 2 remaining factors (COLLSPEED and CARAGE) are continuous and have been divided into 10 classes as described for the computation of MIR. Figure 13) shows the barplots or histograms of the selected factors.

INSERT FIGURE 13

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. MIR coefficients are separately estimated then sorted for each potential causation factor. We then successively determine groups of multivariate factors of different size which best explain the injury severity indicator.

5.1 MAIS

The MIR coefficients are first estimated using the 7 original modalities of MAIS, and then estimated again using only the coarser categories (*Safe*, *Slightly Injured*, *Severe Injured* for MAIS. The coefficients evaluate how well MAIS is explained by each potential causation factor. These coefficients are then sorted by decreasing order of magnitude (Figure 5).

INSERT FIGURE 5

The results are presented in Figure 5. Fixing an outcome descriptor such as MAIS, the MIR coefficient computed for each single factor is represented by the length of an horizontal bar. The tag name of the corresponding factor is displayed on the left and table 4 gives the list of all 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 for a 95% confidence level. All MIR coefficients lie between 0 and 100%.

For MAIS indicator, this analysis shows that the most influent factor OPPONENT with a MIR around 23%. When the 7 initial modalities are regrouped into coarser classes, this feature is even sharper and MIR increases to 28%. Accident KIND appears in second position (MIR = 13%), and accident TYPE comes in third position (MIR = 10%). All MIR coefficients increase when computed for coarser ternary distribution. SPEED of collision, PLACE, and SPEED LIMIT obtain similar MIR coefficients.

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 our database correspond to the use of seat-belts (Figure 6). So the MIR coefficient is here overwhelmingly determined by 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 52%. The MIR coefficients estimated for MAIS confirmed here by objective computation the knowledge of BAST Experts about the main injury severity causation factors in accidents,in the list of chosen fators.

Classification And Regression Tree have been computed for MAIS on the same data using *Tree* library of R software. Figure 8 shows the CART graph. We observe that variables opponent, collision speed, and accident kind are first selected for both methods. In the MIR methodology, each variable appears only one time, in the successive selection of factors. For CART, the optimization process split, during the analyze, the categories which can appears also at different levels.

INSERT FIGURE 8

5.2 HWS

Just as for MAIS, the Mutual Information Ratios are computed between HWS and the potential causation factors listed in Table 4.

INSERT FIGURE 9

As above, the MIR coefficients estimated for HWS are sharper when computed for a ternary distribution as for the original distribution (Figure 9). The OPPONENT type is, as for MAIS, the most influential factor explaining head injury severity however the causation strength is smaller (14% 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 10.

INSERT FIGURE 10

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

Just as for MAIS and HWS, the MIR are computed between AISBEIN and the potential causation factors listed in Table 4.

INSERT FIGURE 11

As for MAIS and HWS, the MIR coefficients are sharper when computed for a ternary distribution as for the original distribution (figure 11). 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 12

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 stronger influence on head injuries, but less for leg injuries.

6 Discussion and conclusions

In this study, we used Mutual Information Ratio to identify the risk factors that can influence the injury severity in traffic accidents. This methodology provides a useful framework that enable studying potentially influent factors with no constraint on the variable type. In the GIDAS database, the accident outcome descriptors (MAIS, HWS, AISBEIN) take categorical values. The original seven modalities of the descriptors have been merged into three coarser categories in order to analyze whether accidents led to severe, light or non injuries. Regarding the outcome descriptor, different types of variables have been considered: continuous variables (collision speed, age of the car), ordinal variables (speed limit, age), binary variables (gender, setbelt...) or nominal data (car damage, opponent,...). Mutual information ratios offer then a wide range of possibilities to study, in the same framework, causation links between variables of different nature with continuous distributions as well as finite sets of modalities.

Factors selection using multivariate MIR yields groups of factors of minimal size with no redundancy, that best explains the outcome descriptor. One main advantage of this approach is to intrinsically handle multi-collinearity factors. If a deterministic relationship exists between two factors, only one of them will be selected. This property is particularly useful when dealing with accidents because traffic data often show serious correlation between variables (e.g. accident kind and accident type in our case).

From a theoretical point of view, one strong advantage of MIR analysis is that it does not require to specify a functional form of dependency such as correlation or Cramer $'$ s indicator. In a classical regression analysis, the estimated relationship between the predictor and the factors can be erroneous, if the model is mis-specified. As well, in case of strong correlations between the factors, the estimation of the coefficients is less precise in a regression analysis which can lead to wrong interpretations between independent and dependant factors.

The analyze of GIDAS data shows that, for all accident outcome descriptors, opponent is the most critical factor determining injury severity in traffic accidents. Factors such as accident kind or accident type are respectively in second and third position when the influence of each factor is analyzed independently.

The MIR coefficients naturally depend on the histogram of factor modalities, and we show how factors for which a single modality is overwhelmingly represented as for example the seat belt factor, yield under-estimated causation strengths values. This can be solved, however, with adequate random re-sampling of the data, which enables the MIR coefficient to correctly estimate causation strength even when one single modality is omnipresent (see figure 6). 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).

Due to the major probabilistic properties of MIR, these mutual information ratios are very efficient to detect non linear causation links. 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: $Y = F_S(X_{(1)}, ..., X_{(k)})$. The empirical relationship F_S naturally depends on the data set S of observations used during learning. In preparatory analysis of accident data prior to model building, it has been validated that, because it is model independent, mutual information is a powerful tool to select the most relevant variables (Mougeot et Azencott, 2008). MIR appears then as a powerful method for identifying the strength of relationship between variables of different natures without constraints on the distribution laws. The most pertinent variables may then be included in predictive models.

References

[1] Abdel-Aty, M. (2003) Analysis of driver injury levels at multiple locations using ordered probit models. Journal of safety research 34, 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, 163-182.

[6] Brillinger, D., Guha, A. (2006) Mutual information in the frequency domain. Journal of statistical planning and inference, 137, 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] Efron, B., Tibshirani, R. (1993) Introduction to the bootstrap.Chapman and Hall.

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

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

[12] 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.

[13] 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.

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

[15] Hautzinger, H. ,Grmping, U., Kreiss, J.P., Mougeot, M., Pastor, C., Pfeiffer, M., Zangmeister T. (2008) Statistical methods for traffic accident causations studies in Europe. TRACE European project N 027763, W.P. 7.5.

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

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

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

[19] 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, 260-266.

[20] 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.

[21] Mougeot, M., Azencott, R. (2008) Information theoretical methods dedicated to accident analysis for GIDAS database. European Symposium on Accident Research Proceedings, Hannover.

[22] 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, 379-753.

[23] 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.

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

[25] 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.

[26] 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, 369-876.

[27] 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.

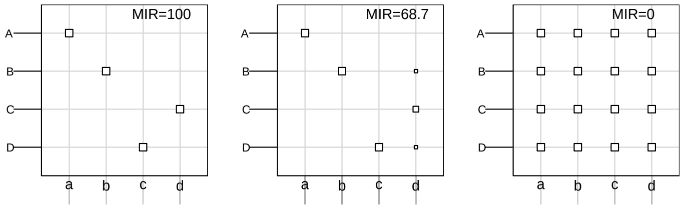

Figure 1: Mutual Information Ratio for some joint distributions. From left to right: $R_{X,Y} = 100\%$; $R_{X,Y} = 68\%$; $R_{X,Y} = 0\%$.

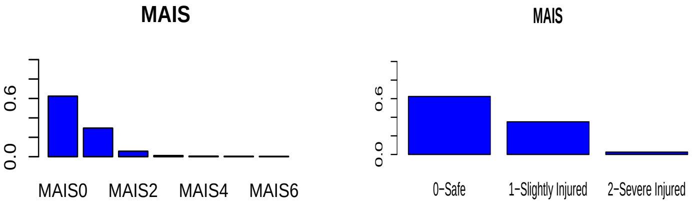

Figure 2: MAIS distribution for GIDAS data. Original and aggregated distribution of data.

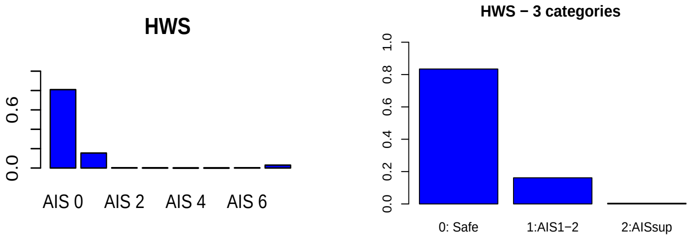

Figure 3: Head injuries distribution for GIDAS data. Original and aggregated distribution of data.

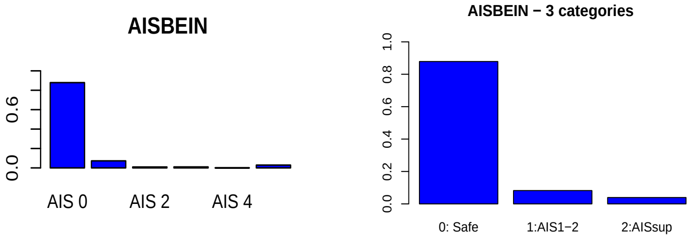

Figure 4: Leg injuries distribution for GIDAS data. Original and aggregated distribution of data.

Factors	nb	Description
Collision	5	collision speed, rollover, opponent, main damage, type of Acc., kind of Acc.
Environment	3	speed limit, place(urban/rural/highway), time (day/night/dawn)
Human	3	age, gender, guilty
Safety	1	seat belt
Vehicle	2	car age, airbag

Table 2: Association factors used for MAIS, HWS or AISBEIN outcome descriptor.

Continuous data	Collision speed, car age.
Binary data	gender, seatbelt, rollover, place, guilty
Ordinal data	speed limit, age,...
Nominal data	car damage, opponent, acc. type, acc. kind

Table 3: Type of Association factors used for MAIS, HWS or AISBEIN outcome descriptor.

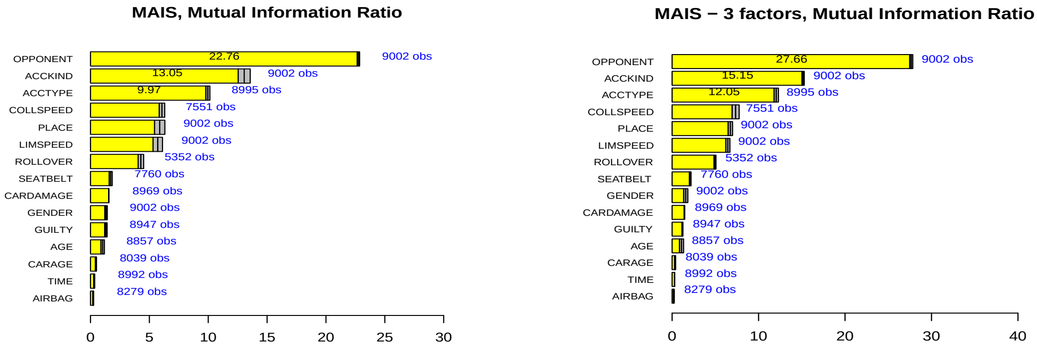

Figure 5: MIR for MAIS (%). Left: initial distribution. Right: MIR computed fo trinary distribution of *safe* and *injured* accidents.

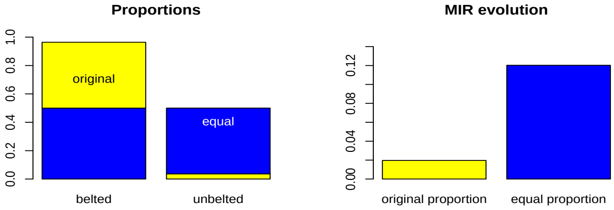

Figure 6: Seatbelt usage for GIDAS original data (yellow) or equaled proportion (blue). Corresponding Impact on MIR.

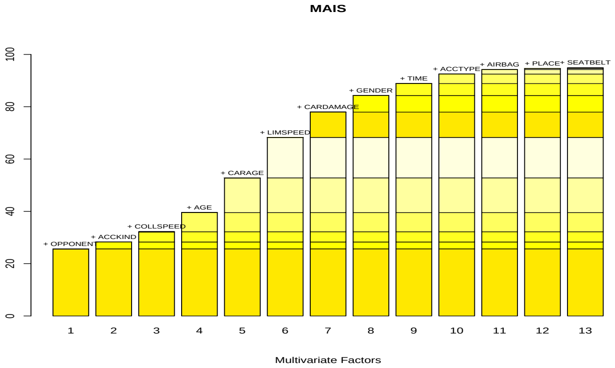

Figure 7: Multivariate MIR for MAIS descriptor (%).

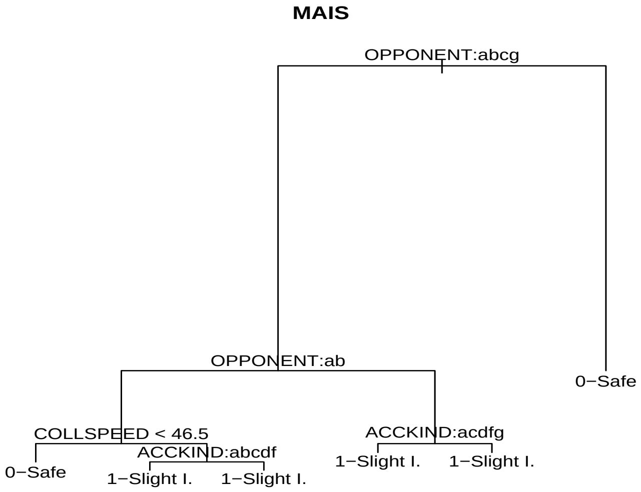

Figure 8: Classification and Regression Tree for MAIS descriptor.

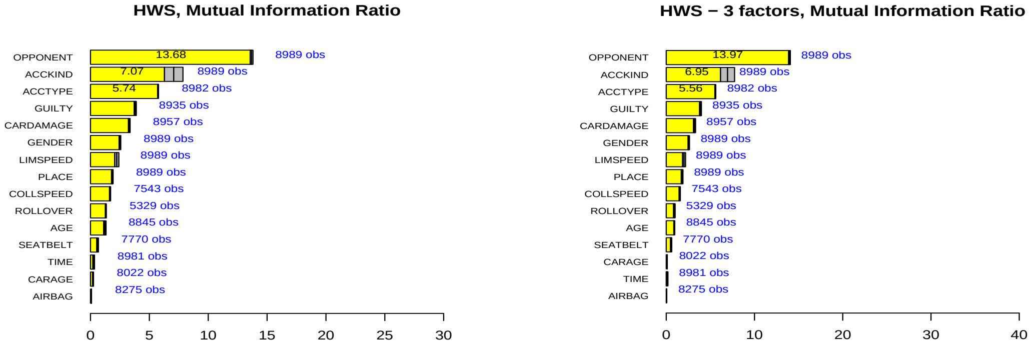

Figure 9: MIR for head injuries (%). Original and ternary distribution.

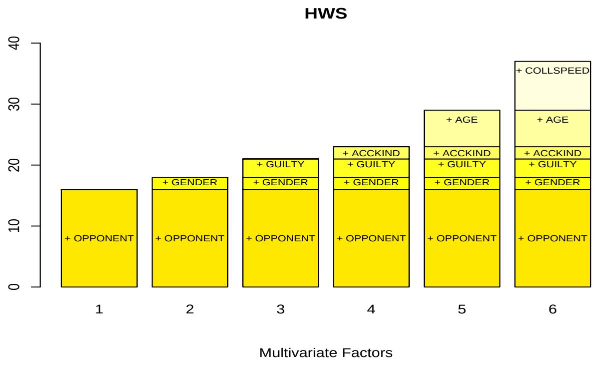

Figure 10: Multivariate Mutual Information Ratio for Head descriptor (%).

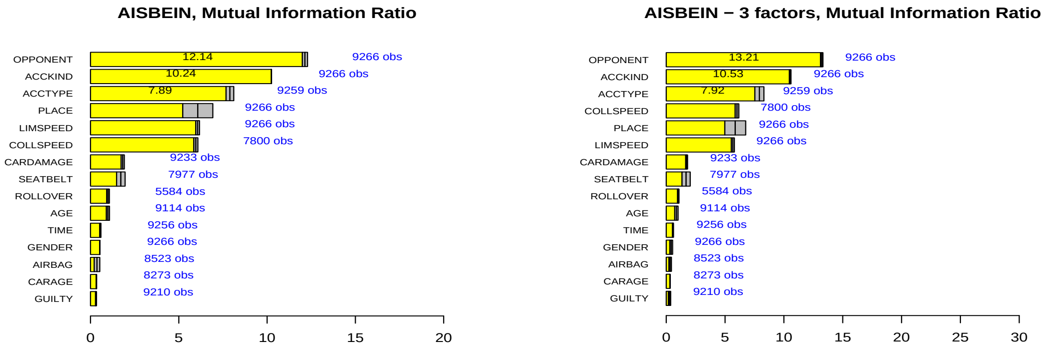

Figure 11: MIR for leg injuries (%). Original and ternary distribution.

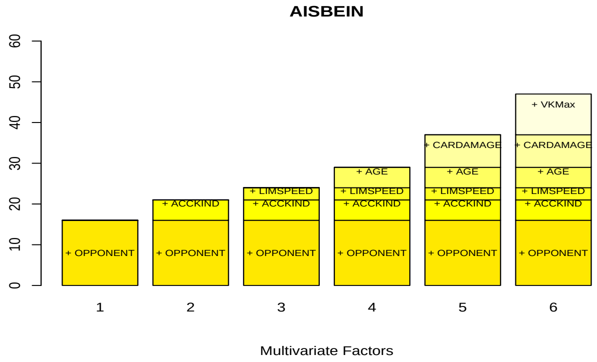

Figure 12: Multivariate Mutual Information Ratio for Leg descriptor (%).

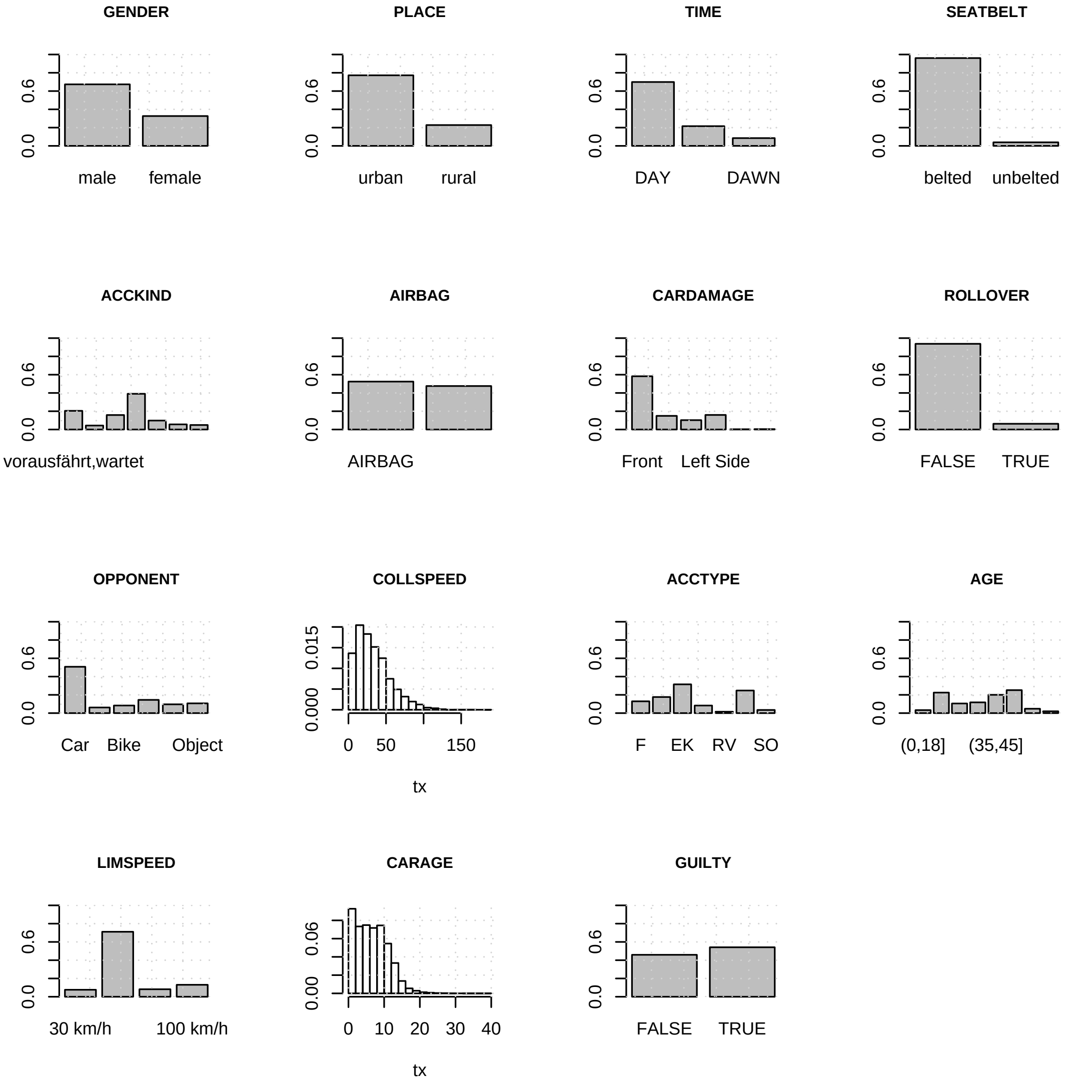

Figure 13: Histogram of potential association factors for MAIS, HWS and AISBEIN descriptor.

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

Table 4: Association factors used for MAIS or HWS outcome descriptor.