Traffic safety: non-linear causation for injury severity

M. Mougeot[1] & R. Azencott[2]
[1] *LPMA / UMR 7599, Denis Diderot University (P7), France*
[2] *Department of Mathematics, University of Houston, USA*

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. 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 entropy contribute to rigorously quantify the influence of causation factors on injury 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. The results show that additional safety gains potential are expected from intelligent speed adaptation systems, collision warning and collision avoidance systems.
Keywords: risk analysis, safety, mutual information, conditional entropy.

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. Information on traffic accidents in Europe are today 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. Extraction of significant injury causation factors hidden in massive databases allows a better understanding and

WIT Transactions on the Built Environment, Vol 117, © 2011 WIT Press
www.witpress.com, ISSN 1743-3509 (on-line)
doi:10.2495/SAFE110221

determination of the crash generating issues. New preventive actions can emerge from in depth investigations of accidents data, with one objective to reduce the rate and severity of accidents [1]. This is an important challenge for expecting safety benefits in the future.

Ordered probit or logit models have been used to analyze injury severity frequencies [2]. The selection of explanatory variables is often 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 [3]. For continuous variables, the correlation coefficient ρ^2 is a long-standing measure of statistical dependency, and is often used in accidents analysis [4]. However, 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. For databases with a large number of descriptive fields, prior knowledge of functional relationships between variables is never directly available and consequently, linear assumptions, can be totally inappropriate to measure statistical dependencies.

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. Mutual information ratios can be computed for discrete, continuous and discrete-continuous variables [5]. MI provides a powerful extension of the classical correlation coefficient and of Cramer's V measure without requiring any constraint on variable nature or linearity. In this paper, we first introduce the general concept of MI and present some analytical and computational developments, then we show how we have adapted this approach to variable selection and we illustrate this in the domain of accidentology by selecting the most informative variables that explain injury severity in a large dataset, the GIDAS database [6].

2 Mutual information

Mutual information, based on conditional entropy, quantifies the relationship between two random variables X and Y. For example, Y could 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 $H_X = -\sum_{i=1}^{m} p_i \log(p_i)$ with the convention, $0 \log(0) = 0$. If X is deterministic, its entropy is minimal, and H_X=0. This is because knowing the values taken by X in random trials brings no new information, since X is constant. But if X follows a uniform distribution, its entropy is maximal, $H_X = -\log(m)$, since any new value of X, which has a constant probability to occur, bring new information.

For both 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: $H_{X,Y} = -\sum_{j=1}^{p}\sum_{i=1}^{m} p_{ij}\log(p_{ij})$.

Conditional entropy $H_{Y/X}$ quantifies the average information provided by discovering the actual value of Y when the value of X is already known, and is defined by $H_{Y/X} = -\sum_{j=1}^{p}\sum_{i=1}^{m} p_{ij}\log(p_{j/i})$ where $p_{j/i}$ 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 new information about the value of Y.

Mutual information is based on conditional entropy and is a measure of statistical dependency between two variables X and Y. $I_{X,Y}$ quantifies the average amount of information on the actual value of Y provided by the knowledge of the actual value of X: $I_{X,Y} = H_Y - H_{Y/X}$. 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: $R_{X,Y} = \frac{I_{X,Y}}{H_Y}$. For two independent variables X and Y, prior knowledge of X doesn't provide any information on Y and $R_{X,Y} = 0$. If a deterministic functional relationship exists between X and Y, the prior knowledge of X completely determines the value of Y and the mutual information ratio is then 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.

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. Joint probabilities can be estimated by $\hat{p}_{ij} = \frac{1}{N}\sum_{k=0}^{N} v_{ij}^k$ where $v_{ij}^k = 1$ when $X = \alpha_i$ and $Y = \beta_j$ for observation k and $v_{ij}^k = 0$ otherwise. The plug-in estimate of the mutual information ratio is then $\hat{I}_{X,Y} = \hat{H}_Y - \hat{H}_{Y/X}$ with $\hat{R}_{X,Y} = \frac{\hat{I}_{X,Y}}{\hat{H}_Y}$ and $\hat{H}_{X,Y} = -\sum_{j=1}^{p}\sum_{i=1}^{m} \hat{p}_{ij}\log(\hat{p}_{j/i})$.

Theoretical results 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 bounded by $\log(m)/\sqrt{mN}$. Confidence intervals can then be computed for MIR coefficients [7].

Below we show how we adapted this approach to factor selection.

2.2 Factor selection using mutual information ratio

Factor selection: Let us consider 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 p different factors as follow. Mutual information ratios are first independently computed between Y and all X_j, $1 \leq j \leq p$. Each MIR coefficient lies between 0 and 100%, and evaluates the percentage of information on the value of Y that is provided by X. To compare the influence level of a given factor X_j on a severity indicator Y, the MIR coefficients $R_{X_j,Y}$ are ordered by decreasing magnitude. $X_{(1)}$ then denotes the factor with the largest MIR, which has the highest predictive power for Y:

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

Selection of factor group: Mutual information can also be computed for multivariate factors. 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 the previous equations. The selection of a group G_k of k factors, among p, that have the highest joint predictive power for Y, can be done as above for single factors, and hence select the group G_k^0 of k factors with the highest MIR ratio $R(G_k^0, Y)$. Among all groups of k factors, the group G_k^0 shows the highest predictive power and best explains the Y values. Finding the best group of k factors among p factors is generally computationally not feasible. Therefore, in the multivariate framework, we specially developed a greedy algorithm based on MIR to select the smallest group with the highest predictive power. The following pseudo code details the algorithm for multivariate variables selection based on MIR (table 1).

Table 1: Greedy algorithm based on MIR for selecting a group of factors of small size with the highest predictive power.

Notation: Y is the target variable, $X_1, \ldots, X_p$ the p factors.
Initialisation: $Z_0 = \{\}$; $G_0 = \{\}$; $J_0 = 1 \ldots p$;
choose $K \in \{1..p\}$; K size of the multivariate group of selected factors
Algorithm:
for $k = 1$ to K do
$j_0 = \text{ArgMax}_{j \in J_{k-1}}\, MIR(Y, U_k(j))$ with $U_k(j) = [Z_{k-1}; X_j]$;
$G_k = [G_{k-1}; j_0]$; $\quad Z_k = [Z_{k-1}; X_{j_0}]$; $\quad J_k = J_{k-1} - \{j_0\}$;
end
G_K is the multivariate group of size K with high predictive power on Y.

This approach provides an efficient way of constructing increasing hierarchies of causation factors for a given severity indicator Y. In the following, this Mutual Information Ratio method is applied to GIDAS data in order to extract groups of factors with a high predictive power on injury severity.

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*-) conducts an large German In-Depth Accident Study (GIDAS). In the areas of Hanover and Dresden, personal injury caused by traffic accidents are systematically reported by the police and the fire department stations. Annually, approximately 2,000 traffic accidents are recorded 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 details and the motions of the vehicles and their occupants are reconstituted. The collision processes and the resulting injuries are generally dependent on the technical background 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 GIDAS was, at the end of 2006, around 14000 with the following per year distribution: 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, and, therefore, classical correlation analysis may be of limited use and information methods based on conditional entropy computation offer a more rigorous tool to explore association or causation relationships between variables. We have applied the MIR methodology presented above to GIDAS data, that includes 14000 observations, described by more than 800 fields. Data on all vehicles and people involved in a crash (when at least an injured people can be found) are stored in the database. A preliminary filtering treatment was first applied to the whole database in order to eliminate inappropriate values [7]. For this study, tests and analyzes were implemented using the R statistical programming software [R development Core Team]. All the code and functions used to compute the theoretical coefficients have been programmed using R standard language and are now available for future applications.

4.1 Injury severity indicators

The first analysis was focused on three indicators of injury severity for different body parts (Y variable): Maximum Injury Severity (MAIS), Head Injury Severity and leg injury Severity [8]. To be short, we only present here the results for MAIS. In the GIDAS database, MAIS values fall into 7 categories $0 \ldots 6$, corresponding

to 7 possible values for the maximum severity of injuries. In order to analyze whether accidents led to severe, moderate or to absence of injury, the 7 initial modalities of MAIS were regrouped into 3 categories. The 3 labels *Safe*, *Slightly Injured* and *Severely Injured* denoted respectively accidents with no injury (MAIS tag = 0), accidents with 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 (or no) injury" accidents (MAIS tag ≤ 2), for a total of 11586 available observations.

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 and to estimate and compare the causation strengths between potential causation factors and the injury severity descriptors. In a second step, a combination of causation factors with the highest power to predict injury severity was computed. A list of potential causation factors was first prepared by the German BAST institute, based on expert knowledge. Factors describing collision, environment, human characteristics, safety, site of accident or vehicle characteristics were chosen (see table 2 for a complete description). These factors were of different types: continuous, discrete or nominal. Six factors are linked to accidents with *collision*: the initial speed of the collision (continuous), the kind of opponent (6 categories), the main damage to the car (7 categories), the type (7 categories) and kind of accident (10 categories) and whether or not a rollover happened (binary). *Environmental factors* include: the speed limit (17 categories),

Table 2: Association factors used for MAIS outcome descriptor.

Variable	Description	Modalities
GENDER	Gender	(2) male/ female
PLACE	Place of the accident	(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...
ACCKIND	Kind of accident	(10) unfall/anfährt/...
LIMITSPEED	Speed limit at the scene	(17) 5 km/h/.../ 140 km/h
OPPONENT	Opponent	(7) Car HGV Bike Cyclist ...
AGE	Age of the driver	(8) $(0, 18]$, $(25, 30]$ $(30, 35]$ ...
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	(2) yes/no

the place (binary) and the time of the accident (3 categories). *Human effects* are analyzed through the following factors: the age of the driver (8 categories), its gender (binary), and its guiltiness (binary). The *Vehicle* is described by its age (continuous) and the airbag equipment (binary). *Seat belt* is described by the use, or not, of the seat belt (binary variable). Overall, a total of 15 factors were selected for the study: 13 categorial factors and two continuous variables (collision speed or car age), which were divided into 10 classes as previously described for the computation of MIR.

5 Results

In this section, mutual information ratios (MIR) were computed to estimate the causation strengths between the potentially causal factors and the accident outcome descriptor MAIS in the GIDAS database.

5.1 Impact factors for maximum injury severity

Each MIR value were computed using more than 8000 observations, depending on the proportion of missing values for the studied variables. As each specific MIR calculation involves only a subset S of variables, all records presenting missing values for one or several variables in S were temporarily removed to compute the coefficient. The MIR coefficients were estimated using the coarser categories (*Safe*, *Slightly Injured*, *Severe Injured* for MAIS. These coefficients are sorted by decreasing order of magnitude and evaluate how well MAIS is explained by each potential causation factor (Figure 1).

The results are presented in Figure 1. All MIR coefficients lie theoretically between 0 and 100%. MIR allows ranking on a same graph continuous (collision speed or car age) and discrete variables (others) which can then be easily compared for the strength of their relationship to the variable to be explained.

In the case of the MAIS indicator, this analysis shows that the most influent factor is OPPONENT with a MIR around 28%. Accident KIND appears in second position (MIR $=$ 13%), and accident TYPE comes in third position (MIR $=$ 10%). The SEATBELT factor appears towards the middle of the list with a small MIR (1.95%). At first sight, this is surprising since the usage of a seat belt is considered to be an important factor affecting injury severity of vehicle traffic accidents. Note however that today, drivers and passengers are required by law to use their seat belt. We accordingly observe here that 97% of the available observations in our database correspond to the use of seat-belt. MIR coefficient is here overwhelmingly determined by cases where seat belt is used, and hence only partially reflects the intrinsic risk associated to the absence of a seatbelt. We observe a similar distribution with rollover accidents, which are less frequent in the database.

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

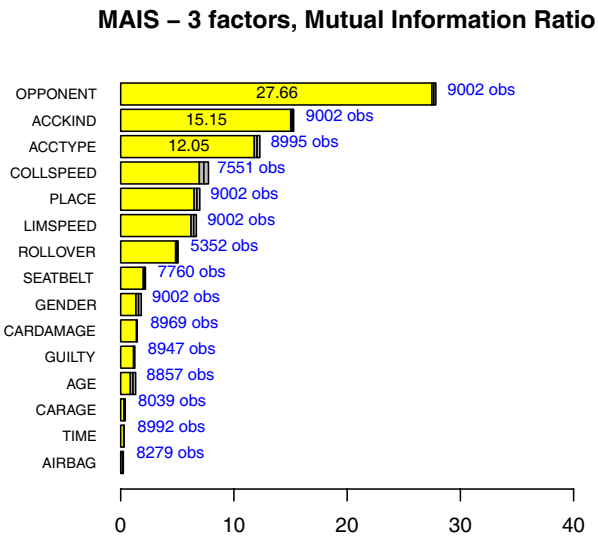

Figure 1: Monovariate MIR for MAIS (%). The MIR coefficient computed for each single factor (displayed on the left) is represented by the length of the horizontal bar. The number of joint observations used is displayed on the right. A confidence interval at a 95% confidence level is displayed at the right end of each bar.

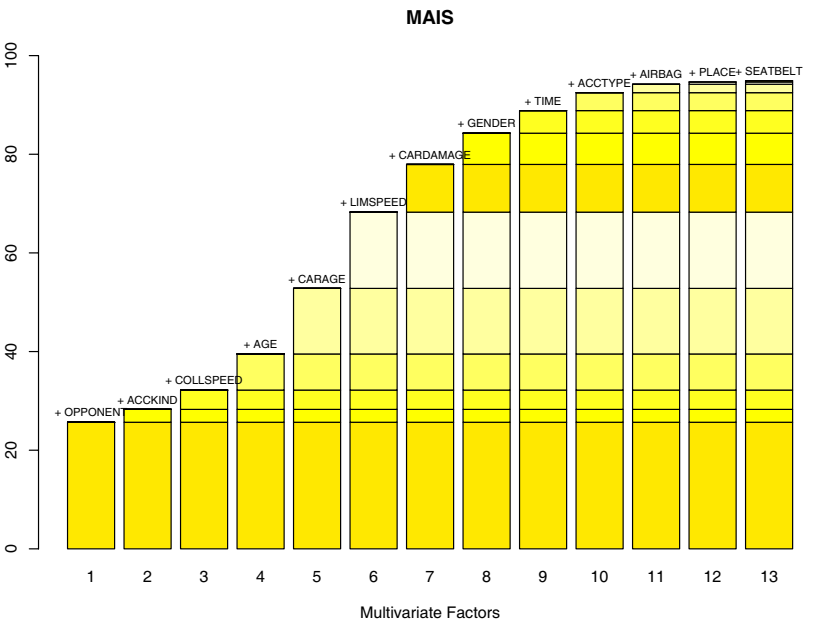

Figure 2: Multivariate MIR for MAIS descriptor (%).

5.2 Group of factors for explaining maximum injury severity

Multivariate analysis was then conducted to analyze which groups among a given number 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 computed as a function of the number of potential causation factors (Figure 2), using the greedy algorithm previously introduced.

The 3rd column of Figure 2 indicates, for instance, that the group of 3 factors (OPPONENT, Accident KIND and Collision SPEED) has a joint MIR of 38%; this group of factors 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 1). In the multivariate analysis, Accident TYPE does not appear in the group of 3 factors and is replaced by the factor Collision SPEED (which was in 4th position for the single factor analysis, after the type of accident). This is essentially due to the sizable redundancy between accident KIND and TYPE, as can be analyzed from their pairwise MIR, which is equal to 52%.

6 Discussion and conclusions

The proposed statistical methodology is applicable to accident research and allows a better understanding and determination of the injury determinants. Our results show that additional safety can be expected from collision warnings, collision avoidance systems, automatic crash notification systems and intelligent speed adaptation. Benefits of technology-based safety measures can be expected using statistics analysis and the safety gains are even higher for higher injury severity levels. The MIR coefficients estimated for MAIS in our study confirmed by objective computation the empirical knowledge of BAST Experts about the main injury severity causation factors in accidents (with regard to the list of factors analyzed here).

Factors selection using multivariate MIR yields groups of factors of minimal size, with no redundancy and that best explained the injury severity. 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 strong correlation between variables (e.g. accident kind and accident type in our case). The results help then to focus on a small key influent parameters.

A theoretical strong advantage of MIR analysis is that it does not require to specify a functional form of dependency such as correlation or Cramer indicator. In a classical regression analysis, the estimated relationship between the predictor and the factors can be erroneous if the model is miss-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

with regard to independent and dependant factors. On the contrary, due to the 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: $Y = F_S(X_{(1)}, \ldots, X_{(k)})$. The empirical relationship F_S naturally depends on the data set S of observations used. In a more complete study, we have demonstrated how accident data analysis can be useful to select the most relevant variable for model building. 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 [9].

MIR is a powerful method for identifying the strength of the relationships between variables of different natures without any constraint on the distribution laws. It is a very useful way to select the most pertinent variables that may be included in predictive models.

Acknowledgements

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

References

[1] Hautzinger, H., Gromping, U., Kreiss, J., Mougeot, M., Pastor, C., Pfeiffer, M. & Zangmeister, T., Statistical methods for traffic accident causations studies in Europe. *TRACE European project N 027763, WP 75,* 2008.

[2] Milton, J., Shankar, V. & Mannering, F., Highway accident severities and the mixed logit model: an explanatory empirical analysis. *Accident Analysis and Prevention*, **40(1)**, pp. 260–266, 2008.

[3] Yau, K., Risk factors affecting the severity of single vehicle traffic accidents in Hong Kong. *Accident Analysis and Prevention*, **36,** pp. 333–340, 2004.

[4] Huang, Y., Che, J., DeArmond, S., Cigularov, K. & P.Y., P.C., Roles of safety climate and shift work on perceived injury risk: a multi-level analysis. *Accident Analysis and Prevention,* **39,** pp. 1088–1096, 2007.

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

[6] GIDAS. *http://www-gidas-org.*

[7] Mougeot, M. & Azencott, R., Information theoretic methods for accident causation studies and prediction of injuries, n 027763, wp7-st 2.2. *European Project, TRACE*, 2007.

[8] Mougeot, M. & Azencott, R., Information theoretical methods dedicated to accident analysis for gidas database. *European Symposium on Accident Research Proceedings*, Hannover, 2008.
[9] Mougeot, M. & Azencott, R., Injury severity analysis based on mutual information for in depth investigation of accident database, technical document. *European Project, TRACE*, 2009.