# Contextual ranking by passive safety of generational classes of light vehicles

Zaïd Ouni, Christophe Denis, Cyril Chauvel, Antoine Chambaz

# Contextual ranking by passive safety of generational classes of light vehicles

Z. Ouni[1,2], C. Denis[3], C. Chauvel[2], A. Chambaz[1]
[1] Modal'X, Université Paris Ouest Nanterre
[2] Laboratoire d'Accidentologie et de Biomécanique
[3] LAMA, Université Paris-Est Marne-la-Vallée

July 13, 2016

**Abstract**

Each year, the BAAC (Bulletin d'Analyse des Accidents Corporels) data set gathers descriptions of traffic accidents on the French public roads involving one or several light vehicles and injuring at least one of the passengers. Each light vehicle can be associated with its "generational class" (GC), a raw description of the vehicle including its date of design, date of entry into service, and size class. In two given contexts of accident, two light vehicles with two different GCs do not necessarily offer the same level of safety to their passengers. The objective of this study is to assess to which extent more recent generations of light vehicles are safer than older ones based on the BAAC data set.

We rely on "scoring": we look for a score function that associates any context of accident and any GC with a real number in such a way that the smaller is this number, the safer is the GC in the given context. A better score function is learned from the BAAC data set by cross-validation, under the form of an optimal convex combination of score functions produced by a library of ranking algorithms by scoring. An oracle inequality illustrates the performances of the resulting meta-algorithm. We implement it, apply it, and show some results.

Keywords: car safety, ensemble learning, oracle inequality.

## 1 Introduction

### 1.1 Background

In 2015, preventing *traffic accidents* (we will simply write *accident*s in the rest of the article) and limiting their often tragic aftermaths is a worldwide, European, French priority for all the actors involved in road safety. The stakes are high. According to the European Commission's statistics [13], 25,700 people died on the roads of the European Union in 2014. For every fatality on Europe's roads there are an estimated four permanently disabling injuries such as damage to the brain or spinal cord, eight serious injuries and 50 minor injuries.

Vehicles obviously play a central role in road activity. Therefore, enhancing road safety notably requires to apprehend vehicles from the angle of accidentology, the study and analysis of the causes and effects of accidents, from the early stage of their design to the late stage of their life on the road. Of course, road safety is one of the keys to the design process when models of vehicles are conceived, developed and validated in research departments and laboratories. Yet, eventually, the analysis of real-life accidents is paramount to evaluating their real road safety.

Active and passive safeties are the two faces of the same coin. Active safety refers to the prevention of accidents by means of driving assistance systems which may guarantee, for instance, better handling and braking. A necessary complement to active safety, passive safety refers to the protection of occupants during a crash, by means of components of the vehicle such as the airbags, seatbelts and, generally, the physical structure of the vehicle. From now on, we focus on the *passive safety* (when not stated otherwise, *safety* will now stand for *passive safety*) and on the need of experts in accidentology for a methodology to better monitor, internally, the safety of generational classes of vehicles based on real-life accidents data.

### 1.2 Safety ratings

For twenty years, safety ratings have been an influential tool for the assessment and improvement of aspects of the safety of vehicles and their crash protective equipment [12]. There are two types of safety ratings. On the one hand, predictive safety ratings assess the safety of vehicles based on crash tests. Introduced in 1995, the New Car Assessment Program (NCAP) [17] has spawned many similar predictive safety ratings, among which the European Euro NCAP. The NCAP safety rating is a five-star score. Three intermediate scores quantify the protection of adults (drivers and passengers), children, and pedestrians in different crash scenarios. An additional intermediate score quantifies the effectiveness of driver assistance systems meant to enhance the active safety of the vehicle. The final five-star score is calculated as a weighted average of the four intermediate scores, ensuring that none of them is under-achieving. On the other hand, retrospective safety ratings assess the safety of vehicles based on real-life accidents from police and insurance claim data. The origin of retrospective safety ratings can be traced back to 1975 and the U.S. Department of Transportation's first annual census of motor vehicle fatalities and its statistical analysis. The Swedish Folksam Car Safety Rating System is the main retrospective safety rating in Europe [12]. For each model of vehicle, a measure is computed of how high is the risk of fatality or injury in the event of a crash. It is obtained under the form of a weighted average of a collection of intermediate risks. It has been shown that there is a strong correlation between Folksam and Euro NCAP safety ratings [22, and references therein].

In this article, we elaborate a novel safety rating of generational classes of *light vehicles* (we will simply write *vehicles* in the rest of the article). The safety rating is retrospective because its construction exploits real-life accidents data. It is also predictive, but in the usual statistical sense: it is possible to extrapolate a safety ranking for a synthetic generational class of vehicles even in the absence of data relative to it. Moreover, it is contextual: the safety ranking is conditioned on the occurrence of an accident in any given context. Before giving more details about our methodology, let us now briefly present the data that we use to elaborate it.

### 1.3 Data

We use the French national file of personal accidents called BAAC data set. BAAC is the acronym for the French expression *Bulletin d'Analyse d'Accident Corporel de la Circulation*, which translates to *form for the analysis of bodily injury resulting from an accident*. Every accident occurring on French public roads and implying the hospitalization or death of one of the persons involved in the accident *should* be described using such forms by the police forces. An example of blank BAAC form is given in Figure 4. Once filled in, a BAAC form describes the conditions of the accident. It tells us *when*, *where*, and *how* the accident occurred. It gives anonymous, partial description(s) of *who* was the driver (or were the drivers, in case more than one vehicle are involved) and, if applicable, *who* were the passengers. It reports *what* was the severity of injury incurred by each occupant.

In addition to these national data, fleet data should allow to associate a generational class

(GC) with every vehicle from the BAAC data set. However, one third of the vehicles cannot be found in the fleet data. Usually caused by wrongly copying a long alpha-numerical code, this censoring is fortunately uninformative. A GC consists of seven variables: date of design, date of entry into service, size class (five categories, based on interior passenger and cargo volumes and architecture), and four additional variables (either categorical or numerical). It gives a raw technical description of the vehicle.

In the rest of the article, we focus on accidents involving one or two light vehicles. When possible, the BAAC data are associated with the GC data. We call BAAC* data set the resulting collection of observations.

It is suggested in the first paragraph of this subsection that the BAAC data set is plagued by under-reporting (see the "*should*"). The pattern of under-reporting is analyzed in [1, 2, 3, 4] by comparing BAAC data with a road trauma registry covering a large county of 1.6 million inhabitants. The analysis reveals that the reporting of fatalities is almost complete. On the contrary, the reporting of non-fatal casualties is rather low, and strongly biased. Overall, the under-reporting rate is estimated to an average 38%, with a large variability depending on the general conditions of the accidents. We do not try to correct the bias. Put in other words, we investigate safety rankings from the angle of accidents in the BAAC* data set and not from that of accidents on French public roads (see the closing discussion in Section 7 on this matter).

### 1.4 Methodology

In two given contexts of accident, two vehicles with two different GCs do not necessarily offer the same level of safety to their passengers. We elaborate, study, encode and apply a statistical algorithm to assess to which extent more recent generations of vehicles are safer than older ones based on the BAAC* data set. Just like the above safety ratings, our algorithm relies on the "scoring" principle: it looks for a score function that associates any context and any GC with a real number in such a way that the smaller is this number, the safer is the GC in the given context of accident. Such score-based ranking procedures have already been considered in the literature [see for instance 14, 11, 10, and references therein]. Tailored to the problem at stake, our procedure innovates in two respects at least. First, it deals with the fact that data arising from a single accident seen from the points of view of its different actors are dependent. Second, it relies on the cross-validation principle to build a better score function under the form of an optimal convex combination of score functions produced by a library of ranking algorithms by scoring, following the general super learning methodology introduced in [35, 29].

### 1.5 Organization of the article

Section 2 presents the BAAC* data set and a model for its distribution. Section 3 formalizes statistically the challenge that we take up. It is cast in terms of the distribution $P$ of an accident seen from the point of view of one of its actors. The definition of $P$ is a by product of that of $\mathbb{P}$, the distribution of an accident seen from the points of view of all its actors, from which our data set is sampled. Section 4 shows how to infer features of $P$ from observations drawn from $\mathbb{P}$ by weighting. Section 5 describes the construction of a meta-algorithm for ranking by super learning and provides theoretical background to motivate its use. Section 6 summarizes the specifics of the application, illustrates properties of the inferred meta-algorithm and how it can be used. Section 7 is a closing discussion. Finally, an appendix gathers some technical material, including proofs of our main results.

# 2 Data and their distribution

## 2.1 Modelling

We observe a sample of $n$ data-structures $\mathbb{O}^1, \ldots, \mathbb{O}^n$. Each of them describes the scene, circumstances, and aftermath of an accident involving one or two vehicles.

Set $1 \leq i \leq n$.

1. Let $K_i$ be the number of vehicles involved in the accident described by $\mathbb{O}^i$, $K_i = 1$ if one single vehicle is involved and $K_i = 2$ two vehicles are involved.
2. If $K_i = 1$, let $J_1^i$ be the number of occupants of the single vehicle involved. If $K_i = 2$, let $J_1^i$ and $J_2^i$ be the numbers of occupants of the first and second vehicles. The choice of what we call the first and second vehicles is made in a such a way that it is uninformative.
3. If $K_i$=1, then $\mathbb{O}^i$ decomposes as $\mathbb{O}^i = (O_{11}^i, \ldots, O_{1J_1^i}^i)$. For convenience, we will also use the alternative notation $\mathbb{O}^i = \mathbb{O}_1^i$. If $K_i = 2$, then $\mathbb{O}^i$ decomposes as $\mathbb{O}^i = (\mathbb{O}_1^i, \mathbb{O}_2^i)$ with $\mathbb{O}_1^i = (O_{11}^i, \ldots, O_{1J_1^i}^i)$ and $\mathbb{O}_2^i = (O_{21}^i, \ldots, O_{2J_2^i}^i)$.

   For each $1 \leq k \leq K_i$ and $1 \leq j \leq J_k^i$, $O_{kj}^i$ describes the accident from the point of view of the $j$th occupant of the vehicle labelled as $k$. The choice of what we call the first to $J_k^i$th occupants is also made in such a way that it is uninformative.
4. Set $1 \leq k \leq K_i$ and $1 \leq j \leq J_k^i$. Data-structure $O_{kj}^i$ decomposes as $O_{kj}^i = (Y_{kj}^i, Z_{kj}^i)$.
   (a) Component $Z_{kj}^i$ indicates the severity of injuries incurred by the $j$th occupant of vehicle $k$ in accident $i$. It equals one if the injury is fatal (occupant dead within 30 days of the accident) or severe (occupant hospitalized for more than 24 hours) and zero if the injury is light (occupant hospitalized for less than 24 hours) or the occupant is unharmed.
   (b) Component $Y_{kj}^i = (W_{kj}^i, \Delta_k^i, \Delta_k^i X_k^i)$ gathers the context $W_{kj}^i$ of accident from the point of view of the $j$th occupant of vehicle $k$ in accident $i$, a missingness indicator $\Delta_k^i \in \{0, 1\}$ and its product $\Delta_k^i X_k^i$ with the GC $X_k^i$ of the vehicle labeled $k$.
      - GC $X_k^i$ of vehicle $k$ in accident $i$ gives a raw technical description of the vehicle. It consists of seven variables: date of design, date of entry into service, size class, and four additional variables (either categorical or numerical). Size class is a five-category variable. Its levels are "supermini car", "small family car", "large family car", "executive car" and "minivan".
      - GC $X_k^i$ may be missing, in which case $\Delta_k^i = 0$. Otherwise, $\Delta_k^i = 1$.
      - Section A.1 describes in details the content of $W_{kj}^i$. In particular, $W_{kj}^i$ includes $J_k^i$.

Let $J_{\max}$ be the maximal number of occupants of a vehicle. It follows from the uninformativeness of the labellings that there exists a finite collection of distributions $\{\widetilde{P}_{kj}^i : 1 \leq k \leq 2, 1 \leq j \leq J_{\max}\}$ such that, conditionally on $K^i$,

- if $K^i = 1$ then, conditionally on $J_1^i$, $O_{11}^i, \ldots, O_{1J_1^i}^i$ are identically distributed and drawn from $\widetilde{P}_{1,J_1^i}^i$;
- if $K^i = 2$, then $\mathbb{O}_1^i$ and $\mathbb{O}_2^i$ follow the same distribution; moreover, conditionally on $J_k^i$, $O_{k1}^i, \ldots, O_{kJ_k^i}^i$ are identically distributed and drawn from $\widetilde{P}_{2,J_k^i}^i$ (for both $k = 1, 2$).

If $K^i = 2$, then it also holds that, conditionally on $(J_1^i, J_2^i)$, $O_{k1}^i, \ldots, O_{kJ_k^i}^i$ are identically distributed (for both $k = 1, 2$).

## 2.2 Assumptions

We now derive a second collection of distributions $\{P^i_{kj} : 1 \leq k \leq 2, 1 \leq j \leq J_{\max}\}$ from $\{\widetilde{P}^i_{kj} : 1 \leq k \leq 2, 1 \leq j \leq J_{\max}\}$, where each $P^i_{kj}$ is characterized by intervening on $\widetilde{P}^i_{kj}$. The characterization of $P^i_{kj}$ uses a generic random variable $\widetilde{O}_{k1}$ drawn from $\widetilde{P}^i_{kj}$, which decomposes as $\widetilde{O}_{k1} = (Y_{k1}, Z_{k1})$ with $Y_{k1} = (W_{k1}, \Delta_k, \Delta_k X_k)$. The absence of a superscript $i$ is a notational reminder of the fact that $\widetilde{O}_{k1}$ is a generic random variable as opposed to an observation.

The characterization goes as follows. Each distribution $\widetilde{P}^i_{kj}$ gives rise to a distribution $P^i_{kj} = \widetilde{P}^i_{kj}(do(\Delta^i_k = 1))$ characterized as the distribution of $O_{k1}$ generated by this three-step procedure: *(a)* draw $\widetilde{O}_{k1}$ from $\widetilde{P}^i_{kj}$, *(b)* set $O_{k1} = \widetilde{O}_{k1}$, *(c)* replace the components $\Delta_k$ and $\Delta_k X_k$ of $O_{k1}$ with 1 and $X_k$, respectively. The difference between $P^i_{kj} = \widetilde{P}^i_{kj}(do(\Delta^i_k = 1))$ and $\widetilde{P}^i_{kj}$ is that the former imposes non-missingness of $X_k$.

We make the following four assumptions.

**A1.** Conditionally on $K^i = 2$ and $(J^i_1, J^i_2, \Delta^i_1, \Delta^i_2)$, $O^i_{k1}, \ldots, O^i_{kJ^i_k}$ are drawn from the conditional distribution of $\widetilde{O}_{k1}$ given $\Delta_k = \Delta^i_k$ under $\widetilde{P}^i_{2,J^i_k}$ (for all $1 \leq i \leq n$ and $1 \leq k \leq 2$).

**A2.** The distribution $P^i_{kj} = \widetilde{P}^i_{kj}(do(\Delta^i_k = 1))$ coincides with the conditional distribution of $\widetilde{O}_{k1}$ given $\Delta_k = 1$ under $\widetilde{P}^i_{kj}$ (for all $1 \leq i \leq n$, $1 \leq k \leq 2$, and $1 \leq j \leq J_{\max}$).

**A3.** The observations $\mathbb{O}^1, \ldots, \mathbb{O}^n$ are independent and follow the same distribution $\mathbb{P}$, hence $P^1_{kj} = \ldots = P^n_{kj} = P_{kj}$ for all $1 \leq k \leq 2$ and $1 \leq j \leq J_{\max}$.

**A4.** We know beforehand the conditional probabilities $\pi(j_1) = \mathbb{P}(\Delta_1 = 1 | K = 1, J_1 = j_1)$ and $\pi(j_1 j_2) = \mathbb{P}(\Delta_1 = 1 | K = 2, (J_1, J_2) = (j_1, j_2))$ for all $1 \leq j_1, j_2 \leq J_{\max}$.

Note that each $\pi(j_1, j_2)$ in **A4** also equals $\mathbb{P}(\Delta_2 = 1 | K = 2, (J_1, J_2) = (j_1, j_2))$ because the choice of what we call the first and second vehicles is made in a such a way that it is uninformative.

Under **A1**, $(O^i_{11}, \ldots, O^i_{1J^i_1})$ is conditionally independent from $(J^i_2, \Delta^i_2)$ given $K^i = 2$ and $(J^i_1, \Delta^i_1)$, and vice versa. With **A1**, we thus neglect the information that $(J^i_2, \Delta^i_2)$ may convey on the shared marginal conditional distribution of $O^i_{11}, \ldots, O^i_{1J^i_1}$ given $K^i = 2, (J^i_1, J^i_2, \Delta^i_1, \Delta^i_2)$ and, symmetrically, the information that $(J^i_1, \Delta^i_1)$ may convey on the shared marginal conditional distribution of $O^i_{21}, \ldots, O^i_{2J^i_2}$ given $K^i = 2, (J^i_1, J^i_2, \Delta^i_1, \Delta^i_2)$. Typically, knowing that $J^i_2$ is large makes it more likely that the second vehicle be larger, heavier, and more powerful; this may say something about the common marginal conditional distribution of $O^i_{11}, \ldots, O^i_{1J^i_1}$ given $K^i = 2, (J^i_1, J^i_2, \Delta^i_1, \Delta^i_2)$, a piece of information that we assume negligible.

Assumption **A2** supposes that missingness of a GC is uninformative. This is true if $(\widetilde{O}_{kj} \setminus (\Delta_k, \Delta_k X_k), X_k)$, the data-structure $\widetilde{O}_{kj}$ deprived of $\Delta_k$ with $X_k$ substituted for $\Delta_k X_k$, is independent from $\Delta_k$ under $\widetilde{P}^i_{kj}$ for all $1 \leq i \leq n$, $1 \leq k \leq 2$, and $1 \leq j \leq J_{\max}$.

With **A3**, we model our data-structures as independent draws from the observational experiment of distribution $\mathbb{P}$. Under **A3**, it is possible to test the validity of **A2** from the data.

Introduce the mixture

$$P = \sum_{j=1}^{J_{\max}} \mathbb{P}(K = 1, J_1 = j) P_{1j} + \sum_{j=1}^{J_{\max}} \mathbb{P}(K = 2, J_1 = j) P_{2j}. \tag{1}$$

Under **A2** and **A3**, $P$ is the shared distribution of every component $O_{kj}$ of $\mathbb{O}$ drawn from $\mathbb{P}$ under the constraint that the GC $X_k$ be observed. In other words, $P$ is the distribution of a random variable *fully* describing the scene, circumstances, and aftermath of an accident from the point of view of one of its actors. Assumption **A4** allows to infer features of $P$ based on

sampling from $\mathbb{P}$, see lemma 1 in Section 4. We actually estimate the conditional probabilities in **A4** based on a validation data set, see Section 6.1, and treat our estimators as deterministic proportions. Because the sample size of the validation data set is very large, our estimators are very accurate. We acknowledge that our assessments of performance may nevertheless be slightly overly optimistic. See [20] for the correction of the asymptotic distribution of the likelihood ratio statistics when nuisance parameters such as the probabilities in **A4** are estimated based on an external source.

# 3 Statistical challenge: contextual ranking by safety of generational classes

Our main objective is to learn to rank GCs by safety in different contexts of accident. We are now ready to formalize this statement.

Denote $\mathcal{Y}$ and $\mathcal{O} = \mathcal{Y} \times \{0,1\}$ the sets where $Y$ and $O = (Y,Z)$ take their values when $O$ is drawn from $P$. Formally, our objective is to build from the data a function/ranking rule $r : \mathcal{Y}^2 \to \{-1,1\}$ and to assert that, for every $(y,y') \in \mathcal{Y}^2$, where $y, y'$ both consist of a context and a GC, $y$ is safer than $y'$ if and only if $r(y,y') = 1$.

Let $P^{\otimes 2}$ denote the distribution of $(O,O')$ with $O$ and $O'$ independently sampled from $P$. The statistical performance of a ranking rule $r$ can be measured through its ranking risk $E_{P^{\otimes 2}}(L^0(r,O,O'))$, where the loss function $L^0$ maps any ranking rule $\rho : \mathcal{Y}^2 \to \{-1,1\}$ and two independent draws from $P$ denoted $O = (Y,Z)$ and $O' = (Y',Z')$ to $L^0(\rho,O,O') = \mathbf{1}\{(Z - Z')\rho(Y,Y') > 0\}$. This choice is motivated as follows:

- if $Z = Z'$, then $Y$ and $Y'$ are equally safe or unsafe, and $L^0(\rho,O,O') = 0$, while no ranking can be interpreted as incorrect;
- if $(Z,Z') = (1,0)$, then $Y$ proves less safe than $Y'$ and $L^0(\rho,O,O') = 1$ is equivalent to $\rho(Y,Y') = 1$, which does not imply a correct ranking;
- symmetrically, if $(Z,Z') = (0,1)$, then $Y$ proves safer than $Y'$ and $L^0(\rho,O,O') = 1$ is equivalent to $\rho(Y,Y') = -1$, which does not imply a correct ranking

For self-containedness, we now recall three classical results about ranking [14, 11, 10] (the easy proofs are given in Section A.2). The take-home message essentially consists in the following facts: *(a)* there exists an optimal rule which takes the specific form of a "scoring" rule associated with the conditional expectation of $Z$ given $Y$, *(b)* the difficulty of the ranking can be expressed in terms of Gini's mean difference coefficient, and *(c)* there is a strong link between the ranking risk and the area under the curve.

Introduce $Q_0(Y) = P(Z = 1|Y)$ and the ranking rule $r_0$ characterized by

$$r_0(Y,Y') = 2\mathbf{1}\{Q_0(Y) \leq Q_0(Y')\} - 1. \tag{2}$$

The ranking rule $r_0$ is a "scoring" rule: to rank $y, y' \in \mathcal{Y}$ based on $r_0$, it is sufficient to evaluate $Q_0(y)$ and $Q_0(y')$, then to assess whether $Q_0(y) \leq Q_0(y')$ or not. Its ranking risk satisfies

$$E_{P^{\otimes 2}}\left(L^0(r_0,O,O')\right) = \mathrm{Var}_P(Z) - \frac{1}{2}E_{P^{\otimes 2}}\left(|Q_0(Y) - Q_0(Y')|\right). \tag{3}$$

The scoring rule $r_0$ is optimal in the sense that, for every ranking rule $r : \mathcal{Y}^2 \to \{-1,1\}$, it holds that

$$0 \leq E_{P^{\otimes 2}}\left(L^0(r,O,O')\right) - E_{P^{\otimes 2}}\left(L^0(r_0,O,O')\right). \tag{4}$$

This inequality still holds when $r_0(Y,Y')$ is chosen arbitrarily in (2) for couples $(Y,Y')$ such that $Q_0(Y) = Q_0(Y')$. Equality (3) teaches us that the optimal risk is upper-bounded by $1/4$,

and that the difficulty of the ranking problem depends on the concentration properties of $Q_0(Y)$ through Gini's mean difference $E_{P^{\otimes 2}}(|Q_0(Y) - Q_0(Y')|)$.

The ranking risk is closely related to the area under the curve, see Section A.2. In this paragraph, let $s : \mathcal{Y} \to [0,1]$ be a scoring function such that $P^{\otimes 2}(s(Y) = s(Y')) = 0$ and let $r_s : \mathcal{Y}^2 \to \{-1, 1\}$ be the corresponding scoring rule given by $r_s(y, y') = 2\mathbf{1}\{s(y) \leq s(y')\} - 1$. In particular, the RHS expression in (4) can be easily bounded to yield the following stronger version of (4):

$$0 \leq E_{P^{\otimes 2}}\left(L^0(r_s, O, O')\right) - E_{P^{\otimes 2}}\left(L^0(r_0, O, O')\right) \leq 2E_P\left(|Q_0(Y) - s(Y)|\right). \tag{5}$$

Moreover, the ranking risk of $r_s$ satisfies

$$1 - \frac{E_{P^{\otimes 2}}\left(L^0(r_s, O, O')\right)}{2P(Z=1)P(Z=0)} = P^{\otimes 2}\left(s(Y) \geq s(Y')|Z = 1, Z' = 0\right) = \mathrm{AUC}_s \leq \mathrm{AUC}_{Q_0}. \tag{6}$$

Since the optimal ranking rule $r_0 = r_{Q_0}$ defined in (2) is a scoring rule, we will restrict our search for a ranking rule to the set of scoring rules.

# 4 Shifting from the comprehensive description of an accident to the coarser description from the point of view of one of its actors

Our approach to the contextual ranking of GCs by safety relies on the inference of quantities that write as $E_P(f_1(O))$ and $E_{P^{\otimes 2}}(f_2(O, O'))$ for some integrable functions $f_1 : \mathcal{O} \to \mathbb{R}$ and $f_2 : \mathcal{O} \times \mathcal{O} \to \mathbb{R}$, where $\mathcal{O} \times \mathcal{O}$ is the set of values that $(O, O')$ can take when it is sampled from $P^{\otimes 2}$. The following lemma shows that it is possible to relate $E_P(f_1(O))$ to the expectation under $\mathbb{P}$ of a random variable $\mathcal{W}(f_1)(\mathbb{O})$ deduced from $f_1$ by appropriate weighting. This technical device is often used when the observations at hand are not drawn from the distribution of interest itself (here, $\mathbb{O}^1, \ldots, \mathbb{O}^n$ are sampled from $\mathbb{P}$ and not $P$). One of the most typical example is case-control studies [33, 9] when one wishes to infer features of the population distribution which cannot be described as features of the conditional distributions of cases or controls without assuming that the population distribution belongs to a very specific parametric model. The easy proof of Lemma 1 is presented in Section A.3.

**Lemma 1.** *Let $f_1(O)$ be a real-valued random variable such that $E_P(f_1(O))$ is well-defined. It gives rise to the real-valued random variable $\mathcal{W}(f_1)(\mathbb{O})$ characterized by $\mathbb{O}$ drawn from $\mathbb{P}$ and*

$$\mathcal{W}(f_1)(\mathbb{O}) = \frac{1}{K}\sum_{k=1}^{K} \frac{\mathbf{1}\{\Delta_k = 1\}}{J_k \pi(J_1 \ldots J_K)} \sum_{j=1}^{J_k} f_1(O_{kj}),$$

*where $\pi(J_1 \ldots J_K)$ equals either $\pi(J_1)$ if $K = 1$ or $\pi(J_1 J_2)$ if $K = 2$, see* **A4**. *It holds that $E_P(f_1(O)) = E_{\mathbb{P}}(\mathcal{W}(f_1)(\mathbb{O}))$.*

Thus, denoting $\mathbb{P}_n$ the empirical distribution $\mathbb{P}_n = n^{-1}\sum_{i=1}^n \mathrm{Dirac}(\mathbb{O}^i)$, it appears that

$$E_{\mathbb{P}_n}(\mathcal{W}(f_1)(\mathbb{O})) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{W}(f_1)(\mathbb{O}^i) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{K^i}\sum_{k=1}^{K^i} \frac{\mathbf{1}\{\Delta_k^i = 1\}}{J_k^i \pi(J_1^i \ldots J_{K^i}^i)} \sum_{j=1}^{J_k^i} f_1(O_{kj}^i) \tag{7}$$

is an estimator of $E_P(f_1(O))$ based on the observations $\mathbb{O}^1, \ldots, \mathbb{O}^n$ which are independently drawn from $\mathbb{P}$. The rationale of the definition of $\mathcal{W}(f_1)$ is easy to explain in light of (7): to

estimate $E_P(f_1(O))$ based on $\mathbb{P}_n$, it is possible to use every $O^i_{kj}$ with an observed GC (hence the indicators $\mathbf{1}\{\Delta^i_k = 1\}$), provided that we properly balance the contributions of each $\mathbb{O}^i$ depending *(a)* on the number of vehicles involved in the accident (see the factor $(K^i)^{-1}$), *(b)* on how many actors contribute their own description of the single accident summarized by $\mathbb{O}^i$ (see the factor $(J^i_k)^{-1}$), and *(c)* on how likely it is to observe a GC given the number of occupants in each vehicles involved (see the factor $\pi(J^i_1 \dots J^i_{K^i})^{-1}$). Note that our unique assumption on how the components $O^i_{kj}$ of $\mathbb{O}^i$ depend on each other is **A1** (**A2** specifies how the components of $O^i_{kj}$ depend on each other). In particular, if $K^i = 1$, *i.e.* if a single vehicle is involved in the accident summarized by $\mathbb{O}^i$, then we make literally no assumption on the dependency structure of $\mathbb{O}^i = (O^i_{11}, \dots, O^i_{1J^i_1})$.

The counterpart to Lemma 1 focusing on $E_{P^{\otimes 2}}(f_2(O, O'))$ does not deserve to be stated in a lemma. We will simply exploit that if $O_{11}$ and $O'_{11}$ are the first components of $\mathbb{O}$ and $\mathbb{O}'$ drawn independently from $\mathbb{P}$, then $E_{P^{\otimes 2}}(f_2(O, O')) = E_{\mathbb{P}^{\otimes 2}}(f_2(O_{11}, O'_{11}))$ (similar to $P^{\otimes 2}$, the notation $\mathbb{P}^{\otimes 2}$ will not be used in the rest of the article). From an empirical point of view, we will estimate $E_{P^{\otimes 2}}(f_2(O, O'))$ with the $U$-statistics

$$\frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} f_2(O^i_{11}, O^j_{11}).$$

# 5 Building a meta-algorithm for ranking

Section 5.1 first presents the elaboration of a meta-algorithm in a general framework. An oracle inequality shows the merit of the approach. Section 5.2 focuses on the elaboration of a meta-algorithm for ranking.

In Section 5.1 (and in Section A.4 as well), given a measure $\mu$ and a $\mu$-integrable function $f$, we use the shorthand notation $\mu f = \int f d\mu$ for clarity of exposition.

## 5.1 General presentation and oracle inequalities

Say that we are interested in estimating a particular feature/parameter of $P$, and that we know several approaches to do so. Instead of choosing one of them, we advocate for considering the whole collection of them, seen as a library of algorithms, and combining them into a meta-algorithm drawing data-adaptively the best from each of them. Many methods have been proposed in this spirit, now gathered under the name of "ensemble learners" [see 34, 38, 7, 8, 19, to cite only a few seminal works, with an emphasis on methods using the cross-validation principle]. We choose to rely on the super learning methodology [35, 29]. Its specifics are described in Section 6.2. The rest of this section and the next one is not specialized to super learning but applies to it.

Let $\mathcal{M}$ be a set of probability distributions on $\mathcal{O}$ such that $P \in \mathcal{M}$, where $P$ is defined in (1). Let $\Psi : \mathcal{M} \to \Theta$ be a mapping/parameter from $\mathcal{M}$ to a parameter set $\Theta$. We denote $\theta_0 = \Psi(P)$ the parameter evaluated at the truth $P$. We assume that it is identifiable in the sense that there exists a loss function $\ell$ mapping any $\theta \in \Theta$ and $(o, o') \in \mathcal{O}^2$ to $\ell(\theta, o, o') \in \mathbb{R}$ in such a way that

$$\mathcal{R}(\theta_0) = P^{\otimes 2}\ell(\theta_0) \in \min_{\theta \in \Theta} P^{\otimes 2}\ell(\theta), \tag{8}$$

where we use the shorthand notation $\ell(\theta)$ for the function from $\mathcal{O}^2$ to $\mathbb{R}$ given by $\ell(\theta)(o, o') = \ell(\theta, o, o')$. It is required that the loss function $\ell$ be *symmetric*: for all $\theta \in \Theta$, $(o, o') \in \mathcal{O}^2$, $\ell(\theta)(o, o') = \ell(\theta)(o', o)$.

Let $\widehat{\Psi}_1, \ldots, \widehat{\Psi}_{\mathcal{K}_n}$ be $\mathcal{K}_n$ algorithms for the estimation of $\theta_0$. For each $1 \leq k \leq \mathcal{K}_n$, for each subset $\{\mathbb{O}^i : i \in S\}$ of the complete data set and related empirical measure $\mathbb{P}_n^S = (\text{card}(S))^{-1} \sum_{i \in S} \text{Dirac}(\mathbb{O}^i)$ , $\widehat{\Psi}_k[\mathbb{P}_n^S] \in \Theta$ is an estimator of $\theta_0$. We want to determine which of the algorithms better estimates $\theta_0$. The cross-validation principle is the key both to determining the better algorithm and to evaluating how well we perform in selecting it.

Let $B_n \in \{0,1\}^n$ be a random vector indicating splits into a training sample, $\{\mathbb{O}^i : 1 \leq i \leq n, B_n(i) = 0\}$, and a validation sample $\{\mathbb{O}^i : 1 \leq i \leq n, B_n(i) = 1\}$. The vector $B_n$ is drawn independently of $\mathbb{O}^1, \ldots, \mathbb{O}^n$ from a distribution such that $n^{-1} \sum_{i=1}^n B_n(i) = p$, for $p \in ]0,1[$ a deterministic proportion bounded away from 0 and 1. For notational simplicity, we choose $p$ so that $np$ be an integer. Then, given $B_n$, $\mathbb{P}_{n,B_n,0} = (n(1-p))^{-1} \sum_{i=1}^n \mathbf{1}\{B_n(i) = 0\}\text{Dirac}(\mathbb{O}^i)$ and $\mathbb{P}_{n,B_n,1} = (np)^{-1} \sum_{i=1}^n \mathbf{1}\{B_n(i) = 1\}\text{Dirac}(\mathbb{O}^i)$ are, respectively, the training and validation empirical measures.

For each $1 \leq k \leq \mathcal{K}_n$, the risk of $\widehat{\Psi}_k[\mathbb{P}_{n,B_n,0}]$ is assessed through

$$\frac{1}{np(np-1)} \sum_{1 \leq i \neq j \leq n} \mathbf{1}\{B_n(i) = B_n(j) = 1\} \ell(\widehat{\Psi}_k[\mathbb{P}_{n,B_n,0}], O_{11}^i, O_{11}^j),$$

a $U$-statistic that we simply denote $P_{n,B_n,1}^{\otimes 2} \ell(\widehat{\Psi}_k[\mathbb{P}_{n,B_n,0}])$. This empirical assessment and notation are justified by the fact that $(O_{11}^i, O_{11}^j)$, the pair consisting of the first components of $\mathbb{O}^i$ and $\mathbb{O}^j$, respectively, is drawn from $P^{\otimes 2}$. Thus, the cross-validated risk of the algorithm $\widehat{\Psi}_k$ is

$$\widehat{\mathcal{R}}_n(k) = E_{B_n}\left(P_{n,B_n,1}^{\otimes 2} \ell(\widehat{\Psi}_k[\mathbb{P}_{n,B_n,0}])\right)$$

and the cross-validation selector is

$$\widehat{k}_n = \underset{1 \leq k \leq \mathcal{K}_n}{\arg\min}\, \widehat{\mathcal{R}}_n(k). \tag{9}$$

The performances of $\widehat{\Psi}_{\widehat{k}_n}$ relative to $\theta_0$ are evaluated by the loss-based dissimilarity $\widetilde{\mathcal{R}}_n(\widehat{k}_n) - \mathcal{R}(\theta_0)$, where $\mathcal{R}(\theta_0)$ given by (8) is the optimal risk and, for each $1 \leq k \leq \mathcal{K}_n$,

$$\widetilde{\mathcal{R}}_n(k) = E_{B_n}\left(P^{\otimes 2} \ell(\widehat{\Psi}_k[\mathbb{P}_{n,B_n,0}])\right) \tag{10}$$

is the true cross-validated risk of the algorithm $\widehat{\Psi}_k$. In the following proposition, we show that $\widehat{\Psi}(\widehat{k}_n)$ performs essentially as well as the benchmark (oracle) selector

$$\widetilde{k}_n = \underset{1 \leq k \leq \mathcal{K}_n}{\arg\min}\, \widetilde{\mathcal{R}}_n(k). \tag{11}$$

**Proposition 2.** *Assume that there exist $\alpha \in [0,1]$ and two finite constants $c_1, c_2 > 0$ such that*

$$\sup_{\theta \in \Theta} \sup_{(o,o') \in \mathcal{O}^2} |\ell(\theta)(o,o') - \ell(\theta_0)(o,o')| \leq c_1, \quad \textit{and} \tag{12}$$

$$\sup_{\theta \in \Theta} \frac{\text{Var}_P\left(E_{P^{\otimes 2}}\left[(\ell(\theta) - \ell(\theta_0))(O,O') \middle| O\right]\right)}{E_{P^{\otimes 2}}\left((\ell(\theta) - \ell(\theta_0))(O,O')\right)^\alpha} \leq c_2. \tag{13}$$

*Set $\delta > 0$ and $c_3 = 16\left(\left(\frac{4(1+\delta)^2 c_2}{\delta^\alpha}\right)^{1/(2-\alpha)} + 65(1+\delta)c_1\right)$. It holds that*

$$E_{\mathbb{P}}\left(\widetilde{\mathcal{R}}_n(\widehat{k}_n) - \mathcal{R}(\theta_0)\right) \leq (1+2\delta) E_{\mathbb{P}}\left(\widetilde{\mathcal{R}}_n(\widetilde{k}_n) - \mathcal{R}(\theta_0)\right) + c_3 \frac{\log(1+4\mathcal{K}_n)}{(np)^{1/(2-\alpha)}}. \tag{14}$$

The proof of Proposition 2 essentially relies on [36]. It is given in Section A.4.

### 5.2 The special case of ranking

We now turn to the elaboration of a meta-algorithm for ranking. Earlier results can be found for instance in [11] (see the oracle inequality in Corollary 8 for a ranking rule obtained by minimizing an empirical risk over a class of rules) and in [32] (see the oracle inequality in Corollary 9 for a ranking rule obtained by aggregating a given set of rules with exponential weights).

In the framework of ranking, we define $\Theta$ as the set of functions mapping $\mathcal{Y}$ to $[0,1]$. The parameter $\Psi$ is characterized by $\Psi(P')(Y) = P'(Z = 1|Y)$ for all $P' \in \mathcal{M}$ (in particular, $\theta_0 = \Psi(P) = Q_0$). We choose the loss function $\ell$ characterized over $\Theta \times \mathcal{O}^2$ by

$$\ell(\theta, o, o') = L^0(r_\theta, o, o') \tag{15}$$

where $r_\theta : \mathcal{Y}^2 \to \{-1, 1\}$ maps any $(y, y') \in \mathcal{Y}^2$ to $r_\theta(y, y') = 2\mathbf{1}\{\theta(y) \leq \theta(y')\} - 1$ (in particular, $r_{\theta_0} = r_0$). By (4), condition (8) is met and $\ell$, which is symmetric, does identify $\theta_0$. With this choice of loss function, the construction of the meta-algorithm is driven by the fact that we are eventually interested in ranking.

The following corollary of Proposition 2 shows that the meta-algorithm built for the purpose of ranking performs essentially as well as the benchmark oracle selector under a margin condition on $Q_0$. In particular, it is thus theoretically justified to resort to super learning, as described in Section 5.2, to build a ranking meta-algorithm from a collection of single ranking algorithms. To the best of our knowledge, Proposition 3 is a new result.

**Proposition 3.** *Assume that there exist $\alpha \in [0,1]$ and a constant $c_2 > 0$ such that, for all $y \in \mathcal{Y}$,*

$$E_P[|Q_0(y) - Q_0(Y)|^{-\alpha}] \leq c_2. \tag{16}$$

*Set $\delta > 0$, $c_1 = 1$ and let $c_3$ be the same constant as in Proposition 2. Then inequality (14) is valid when $\ell$ is given by (15).*

It is easy to verify that (12) holds with $c_1 = 1$ when $\ell$ is given by (15). Proposition 7 in [11] guarantees that (16) implies (13). The detailed proof is given in Section A.4.

As underlined in [11], (16) is rather weak. When $\alpha = 0$, it actually poses no restriction at all, but the rightmost term in (14) decreases in $n^{-1/2}$, a slow rate. Moreover, if the distribution of $Q_0(Y)$ under $P$ is dominated by the Lebesgue measure on $[0,1]$ with a density upper-bounded by $c_4 > 0$ then, for every $0 < \alpha < 1$, (16) holds with $c_2 = 2c_4/(1-\alpha)$ by Corollary 8 in [11]. As $\alpha$ gets closer to one, the rightmost term in (14) decreases faster, at the cost of a larger constant $c_3$.

## 6 Application

### 6.1 A few facts

On the one hand, the 2011 BAAC* data set consists of 16,877 reports of accidents. There are 7,716 one-vehicle and 9,161 two-vehicle accidents reported in it. On the other hand, the 2012 BAAC* data set consists of 15,852 reports of accidents. There are 7,025 one-vehicle and 8,827 two-vehicle accidents reported in it.

We exploit the 2011 BAAC* data set to build our meta-algorithm by super learning. The weights used in the process, see Lemma 1, are estimated based on the 2012 BAAC* data set. The 2012 BAAC* data set is also used to illustrate our application.

Based on it, we infer the conditional probability distribution given $K = 1$ of $J_1$ (the number of occupants of the sole vehicle involved in a one-vehicle accident) and the conditional probability

| | daylight | | driver's age | | under influence of alcohol | |
|---|---|---|---|---|---|---|
| $S$ | yes | no | 20-24 | 50-54 | yes | no |
| $\widehat{P}(Z=1\|W\in S)$ | 0.29 | 0.38 | 0.31 | 0.28 | 0.58 | 0.29 |

| | urban area | | |
|---|---|---|---|
| $S$ | outside | small | large |
| $\widehat{P}(Z=1\|W\in S)$ | 0.45 | 0.14 | 0.04 |

Table 1: Estimates of conditional probabilities of the form $P(Z = 1|W \in S)$. Depending on the choice of $S$, they correspond to the conditional probabilities that an occupant of a vehicle involved in an accident be severely or fatally injured *(a)* given that the accident occurred in daylight or not (columns 1-2 of top table), *(b)* given that the accident occurred outside urban areas, or in a small urban area, or in a large urban area (columns 1-3 of bottom table), *(c)* given that the driver was between 20 and 24 years old, or between 50 and 54 years old (columns 3-4 of top table), and *(d)* given that the driver was under the influence of alcohol, or not (columns 5-6 of top table).

distribution given $K = 2$ of $\{J_1, J_2\}$ (the pair of numbers of occupants of the vehicles involved in two-vehicle accidents), see Table 8. It appears that for a vast majority of one-vehicle accidents (approximately 99% of them), there are no more than five occupants in the car. Moreover, in 54% of the two-vehicle accidents, the sole occupants of the two vehicles are their drivers. In 27% of the two-vehicle accidents, one of the two drivers is accompanied by one person and the other driver is by oneself. A vast majority of the two-vehicle accidents (approximately 99% of them) involve one and one to five, two and two to five, or twice three occupants. The inference based on the 2011 BAAC* data set yields similar results.

We also estimate the conditional probabilities that an occupant of a vehicle involved in an accident be severely or fatally injured *(a)* given that the accident occurred in daylight, or not, *(b)* given that the accident occurred outside urban areas, or in a small urban area, or in a large urban area, *(c)* given that the driver was between 20 and 24 years old, or between 50 and 54 years old, and *(d)* given that the driver was under the influence of alcohol, or not. All the probabilities can be written $P(Z = 1|W \in S)$ for a well-chosen subset $S$ of the set where $W$ drawn from $P$ takes its values. We report their estimates in Table 1.

### 6.2 Library of algorithms and resulting super learning meta-algorithm

The meta-algorithm built by super learning relies on $\mathcal{K} = 49$ base algorithms. The algorithms are derived from 10 main methodologies for the estimation of the regression function $Q_0$. Each algorithm corresponds to a particular choice of tuning parameters and/or to a subset of the components of the explanatory variable $Y$. Table 2 lists the different methodologies and how we tune them. The coding is performed in the language `R` [30]. It greatly benefits from packages contributed by the community, first and foremost the `SuperLearner` package [28].

The main function of the package (`SuperLearner`) can be given a loss function and a model to combine algorithms (through its `method` argument). Instead of giving the loss function $L^0$ introduced in Section 3, we give the smooth approximation $L^\beta$ to it characterized by $L^\beta(r_s, O, O') = 1 - \text{expit}((Z - Z')(s(Y) - s(Y'))/\beta)$, where $\text{expit}(x) = 1/(1 + \exp(-x))$ (all $x \in \mathbb{R}$) and $\beta$ is a fine-tune parameter (we set $\beta = 1/30$), see [24] for a comparison of different AUC-maximization techniques. Moreover, we specify that we want to identify the best convex combination of the $\mathcal{K} = 49$ base algorithms provided as inputs, not the best single one — this is one of the key idea of super learning. This statement is easily clarified using the

terms of Section 5. Denote $\widehat{\psi}_1, \ldots, \widehat{\psi}_{\mathcal{K}}$ the $\mathcal{K}$ base algorithms and $\mathcal{A}_{n,\mathcal{K}}$ a net over the simplex $\Sigma^{\mathcal{K}} = \{a = (a_1, \ldots, a_{\mathcal{K}}) \in \mathbb{R}_+^{\mathcal{K}} : \sum_{k=1}^{\mathcal{K}} a_k = 1\}$ with cardinality $\mathcal{K}_n = \mathcal{O}(n^{\mathcal{K}})$ and such that, for all $a \in \Sigma^{\mathcal{K}}$, there exists $a' \in \mathcal{A}_{n,\mathcal{K}}$ with $\|a - a'\| \leq 1/n$. The $\mathcal{K}$ base algorithms give rise to $\mathcal{K}_n$ algorithms $\widehat{\Psi}_a = \sum_{k=1}^{\mathcal{K}} a_k \widehat{\psi}_k$ $(a \in \mathcal{A}_{n,\mathcal{K}})$. Identifying the best convex combination of $\widehat{\psi}_1, \ldots, \widehat{\psi}_{\mathcal{K}}$ amounts to inferring which element of $\{\widehat{\Psi}_a : a \in \mathcal{A}_{n,\mathcal{K}}\}$ better estimates $Q_0$ for the sake of ranking. In practice, there is no need to specify $\mathcal{A}_{n,\mathcal{K}}$, the numerical optimization being carried out over $\Sigma^{\mathcal{K}}$ itself. Finally, the law of the random splitting vector $B_n$ implements $V = 10$-fold cross-validation: the $n$ observations are arbitrarily gathered in $V = 10$ non-overlapping groups $\{\mathbb{O}^i : i \in I_\nu\}$ $(\nu = 1, \ldots, V)$, and $B_n$ is such that, with probability $V^{-1} = 1/10$, $B_n(i) = \mathbf{1}\{i \in I_\nu\}$ for all $1 \leq i \leq n$.

Let $a_n \in \mathcal{A}_{n,\mathcal{K}}$ be the vector of weights that characterizes the meta-algorithm resulting from the numerical optimization. Only six of its $\mathcal{K} = 49$ components are larger than $10^{-3}$. Say for convenience that they are the six first components of $a_n$. They correspond to random forest applied to all variables ($a_{n,1} \approx 39.7\%$), multivariate adaptive polynomial spline regression applied to all variables ($a_{n,2} \approx 22.0\%$), logistic regression with LASSO penalization applied to all variables ($a_{n,3} \approx 20.9\%$), multivariate adaptive polynomial spline regression applied to factors only ($a_{n,4} \approx 11.8\%$), random forest applied to factors only ($a_{n,5} \approx 4.2\%$), and tree based ranking ($a_{n,6} \approx 1.4\%$).

The purpose of Figure 1 is to give an idea of how the meta-algorithm assigns scores to a GC in a context. The figure is obtained as follows. For each accident from the 2012 BAAC* data set and for each vehicle involved in it with a known GC, we compute the scores assigned by the meta-algorithm to the GC in the contexts of the accident seen from the points of view of all the vehicle's occupants. By separating the resulting scores depending on whether the occupants were "unharmed or slightly injured" (group 0) or "severely or fatally injured" (group 1), we thus obtain two sets of scores. Figure 1 represents the empirical cumulative distribution functions (CDFs) of the two sets of scores. The empirical CDF of scores from group 0 dominates that of scores from group 1, an illustration of the fact that GCs in contexts with more dramatic aftermaths (group 1) tend to get higher scores than GCs in contexts with less dramatic aftermaths (group 0).

Figure 2 presents the empirical ROC curve of our meta-algorithm. Formula (6) shows that the ranking risk is closely related to the AUC. The derivation of a confidence interval for the AUC of our meta-algorithm is computationally prohibitive because of the need to estimate, by bootstrap, the variance of the point estimator of the AUC. Instead, we derive a point estimate and 95%-confidence interval for the cross-validated AUC [see 23, Section 5] of our meta-algorithm, obtaining a point estimate of 82.8% and the confidence interval $[82.4\%, 83.2\%]$ (with $V = 5$ folds).

### 6.3 Illustration

**Ranking eight synthetic GCs in seven synthetic contexts of accident.** For the sake of illustration, we first arbitrarily select an accident from the 2012 BAAC* data set. Its description is reported in Table 3. Second, we arbitrarily characterize eight GCs to rank in seven synthetic contexts of accident. The eight GCs are partially presented in Table 4.

Arbitrarily made up, the synthetic GCs are not obtained by averaging a collection of GCs with common date of design, date of entry into service and size class. Thus, none of them can be interpreted as a typical representant of a certain class of light vehicles.

The seven contexts of accidents are derived from the context described in Table 3, see Table 5. To obtain the first two synthetic contexts, we only modify the hour at which the accident occurred and the light condition, setting them to either 11:00AM and daylight (scenario

| methodology (`R` package) | tuning |
|---|---|
| bagging classification trees (`ipred` [27]) | applied to all variables, or only numeric variables, or only factors; with or without random forest variable importance screening |
| generalized additive models (`gam` [18]) | `deg.gam` set to 1, 2, 3, 4; applied to numeric variables only, with or without stratification by size class |
| generalized boosted regression models (`gbm` [31]) | `interaction.depth` set to 1, 2; applied to all variables, or to numeric variables only, or to factors only; with or without random forest variable importance screening `screen.randomForest` |
| logistic regression with LASSO penalization (`glmnet` [15]) | applied to all variables, or only numeric variables, or only factors; with or without random forest variable importance screening `screen.randomForest`; selection of regularization parameter by cross-validation |
| $k$-nearest neighbors (`kknn` [37]) | `k` set to $5, 7, \ldots, 23, 25$; applied to numeric variables only, with stratification by size class |
| multivariate adaptive polynomial spline regression (`polspline` [21]) | applied to all variables, or only numeric variables, or only factors; with or without random forest variable importance screening `screen.randomForest` |
| neural network (`nnet` [37]) | applied to numeric variables only |
| random forest (`randomForest` [25]) | applied to all variables, or only numeric variables, or only factors; with or without random forest variable importance screening `screen.randomForest` |
| support vector machine (`svm` [26]) | `nu` set to 0.05, 0.01, 0.1, 0.2 |
| tree based ranking (`treeRank` [6]) | applied to numeric variables only |

Table 2: Library of algorithms combined by super learning.

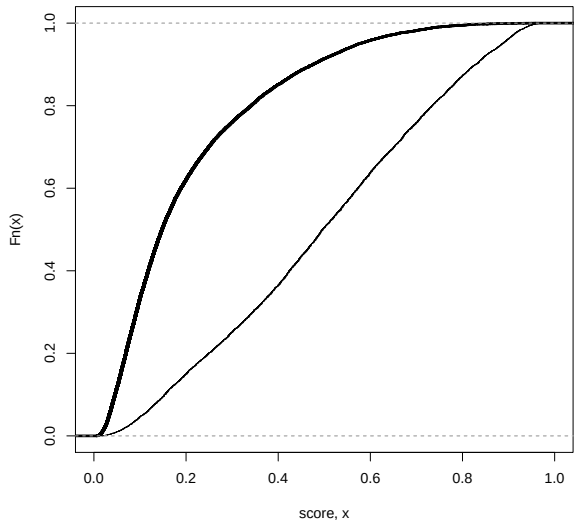


Figure 1: Empirical cumulative distribution functions of scores assigned by the meta-algorithm to GCs in some contexts derived from the 2012 BAAC* data set. See the last but one paragraph of Section 6.2 for details. The top curve corresponds to scores of GCs in contexts of accidents seen from the points of view of occupants who were unharmed or slightly injured (group 0). The bottom curve corresponds to scores of GCs in contexts of accidents seen from the points of view of occupants who were severely or fatally injured (group 1). One reads that 5% only of scores from group 1 are smaller than 0.1. In comparison, 35% of scores from group 0 are smaller than 0.1. One also reads that the 90%-quantiles of scores from groups 0 and 1 equal 0.48 and 0.82, respectively.

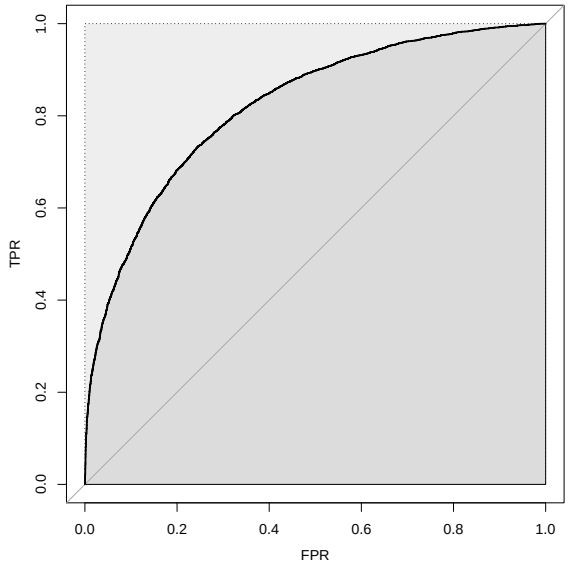


Figure 2: Empirical ROC curve of our meta-algorithm. The estimated value of the cross-validated AUC (with $V = 5$ folds) equals 82.8%, with $[82.4\%, 83.2\%]$ as 95%-confidence interval.

| | |
|---|---|
| *General* | Two vehicles were involved in the accident. There was only one driver in the vehicle of interest. |
| *When and where* | The accident occurred at 5:00PM, on a Thursday of May 2012, outside urban areas. It was daylight, the weather was clear. |
| *What roadway* | The accident did not occur at an intersection. The roadway was straight, its profile level, its surface condition dry. The infrastructure is unknown to us. |
| *What collision* | The vehicle was not responsible of collision. The collision was head-on, with a left-front half initial contact point. The second vehicle involved in the accident was hit. It is unknown to us if a fixed obstacle was hit too. |
| *Which driver* | The driver was a retired male, aged 57. His seatbelt was fastened. He was not driving under the influence of alcohol. He owned the vehicle he was driving, and his driving license was valid. |
| *Which occupant* | The occupant of interest is the driver himself. |

Table 3: Description of a context of accident arbitrarily selected from the 2012 BAAC* data set.

| | generational class (GC) | | |
|---|---|---|---|
| GC code | date of design | date of entry into service | size class |
| S1 | 1983 | 1995 | small family car |
| S2 | 1998 | 2006 | small family car |
| S3 | 2005 | 2009 | small family car |
| L1 | 1995 | 2001 | large family car |
| L2 | 2002 | 2007 | large family car |
| L3 | 2008 | 2010 | large family car |
| M1 | 1994 | 1998 | minivan |
| M2 | 2002 | 2005 | minivan |

Table 4: Eight synthetic GCs. We only report the dates of design, dates of entry into service, size classes, and give each GC a code for future reference. The above GCs are not obtained by averaging a collection of GCs with common date of design, date of entry into service and size class, so none of them can be interpreted as a typical representant of a certain class of light vehicles.

"daylight: yes") or 10:00PM and dark (scenario "daylight: no"). To obtain the next two contexts, we only modify the location of accident, setting it to either large urban area (scenario "urban area: yes") or outside urban areas (scenario "urban area: no"). To obtain the next two contexts, we simply modify the age and occupation of the driver, setting them to either 20 and student (scenario "driver's age: 20") or 50 and professional driver (scenario "driver's age: 50"). To obtain the last two contexts, we simply modify the variable specifying if the driver was under the influence of alcohol, setting it to either yes (scenario "under influence: yes") or no (scenario "under influence: no"). This does result in seven different contexts of accident regrouped in eight scenarios, because the scenarios "urban area: yes" and "under influence: no" coincide.

We underline that the accident and its aftermaths are seen from the point of view of the driver of the vehicle. We compute the scores given to each GC in every context by the meta-algorithm elaborated by super learning with the library presented in the previous subsection. The numerical values are reported in Table 6.

Its is expected by experts that a more recent GC should be safer than an older one within each size class. Inspecting the scores for each combination of size class and scenario yields that in 18 out of 21 combinations, the scores do decrease as the dates of design increase. The three

| scenario | modifications |
|---|---|
| daylight (yes/no) | hour and light condition set to either 11:00AM and daylight (daylight: yes) or 10:00PM and dark (daylight: no) |
| urban area (yes/no) | location of accident set to either large urban area (urban area: yes) or outside urban areas (urban area: no) |
| driver's age (20/50) | age and occupation of driver set to either 20 and student (driver's age: 20) or 50 and professional driver (driver's age: 50) |
| under influence (yes/no) | driver under the influence of alcohol set to either yes (under influence: yes) or no (under influence: no) |

Table 5: Seven synthetic contexts of accident regrouped in eight scenarios. The scenarios "urban area: no" and "under influence: no" coincide.

combinations where the scores do not decrease as expected correspond to L1, L2 and L3. In the three divergent scenarios, "urban area: yes", "driver's age: 20" and "driver's age: 50", L2 and L3 are assessed safer than L1 (as expected) but L2 is assessed safer than L3 (unexpected).

It is also known by experts that driving under the influence of alcohol is far more dangerous than driving sober. Inspecting the last two columns of Table 6 reveals that, in the context described in Table 3, every GC is assessed safer when driven sober relative to under the influence of alcohol. Likewise, it is known by experts that driving in a large urban area is generally safer than driving outside urban areas. Inspecting the third and fourth columns of Table 6 reveals that, in the context described in Table 3, every GC is assessed safer when driven in a large urban area relative to outside urban areas.

Consider now the pairs of scenarios "daylight: yes/no" and "driver's age: 20/50". Inspecting the columns 1-2 and 5-6 of Table 6 reveals that, in each case, one subscenario dominates the other. Namely, every GC is assessed safer in dark light condition than in daylight, safer in a large urban area than outside urban area, and safer when driven by a 20-year old student than by a 50-year old professional driver, all the other variables describing the context of accident being held fixed. These *contextual* results are somewhat unexpected, but they do not fundamentally contradict the *marginal* results shown in Table 1 (by Simpson's paradox, a trend appearing in different groups of data can disappear or even reverse when these groups are combined). It is possible, however, to explain them a posteriori. For instance, we could argue that one drives faster in daylight than in dark light condition outside urban areas, thus increasing the dangerousness in the event of an accident. In this a posteriori explanation, the light condition and location of accident are used as proxies for speed. Finally, we could argue that, all other things being equal, a younger person better withstands physically an accident than an older one.

In conclusion, it is possible, surprisingly, to rank the seven scenarios by increasing order of dangerousness. It appears that, for each GC, the following scenarios are increasingly less safe: "urban area: yes", "driver's age: 20", "driver's age: 50", "urban area: no" (same as "under influence: no"), "daylight: no", "daylight: yes" and "under influence: yes". It is also possible to rank the seven scenarios across GCs, by comparing all scores. Figure 3 represents the $8 \times 7 = 56$ scores in gray scale. To emphasize that we are eventually interested in ranks, and not the scores that yield them, the gray scale is proportional to the rank. The smaller is the score, the lighter is the color and the safer is the GC in the given context of accident. The pattern that emerges is not as clear as the pattern obtained when ranking the scenarios for each GC separately.

**Ranking thirty-one synthetic GCs in all contexts of the 2012 BAAC* data set.** At the request of one reviewer, to build confidence in our results, we also rank thirty-one synthetic GCs in all contexts of the 2012 BAAC* data set. The synthetic GCs are made up by experts.

| | scenario | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | daylight | | urban area | | driver's age | | under influence | |
| GC code | yes | no | yes | no | 20 | 50 | yes | no |
| S1 | 0.546 | 0.521 | 0.146 | 0.520 | 0.423 | 0.460 | 0.613 | 0.520 |
| S2 | 0.442 | 0.437 | 0.100 | 0.410 | 0.324 | 0.370 | 0.521 | 0.410 |
| S3 | 0.421 | 0.415 | 0.096 | 0.388 | 0.292 | 0.353 | 0.494 | 0.388 |
| L1 | 0.523 | 0.504 | 0.072 | 0.494 | 0.362 | 0.427 | 0.587 | 0.494 |
| L2 | 0.483 | 0.470 | 0.059 | 0.448 | 0.326 | 0.394 | 0.544 | 0.448 |
| L3 | 0.481 | 0.459 | 0.069 | 0.442 | 0.330 | 0.398 | 0.528 | 0.442 |
| M1 | 0.514 | 0.495 | 0.068 | 0.498 | 0.349 | 0.427 | 0.565 | 0.498 |
| M2 | 0.441 | 0.427 | 0.048 | 0.413 | 0.295 | 0.372 | 0.511 | 0.413 |

Table 6: Scores assigned to each GC in every context by the meta-algorithm elaborated by super learning. Rearranging the order of columns reveals an interesting pattern, see Table 7.

| | scenario | | | | | | |
|---|---|---|---|---|---|---|---|
| GC code | ua: yes | da: 20 | da: 50 | ua: no/ui: no | d: no | d: yes | ui: yes |
| S1 | 0.146 | 0.423 | 0.460 | 0.520 | 0.521 | 0.546 | 0.613 |
| S2 | 0.100 | 0.324 | 0.370 | 0.410 | 0.437 | 0.442 | 0.521 |
| S3 | 0.096 | 0.292 | 0.353 | 0.388 | 0.415 | 0.421 | 0.494 |
| L1 | 0.072 | 0.362 | 0.427 | 0.494 | 0.504 | 0.523 | 0.587 |
| L2 | 0.059 | 0.326 | 0.394 | 0.448 | 0.470 | 0.483 | 0.544 |
| L3 | 0.069 | 0.330 | 0.398 | 0.442 | 0.459 | 0.481 | 0.528 |
| M1 | 0.068 | 0.349 | 0.427 | 0.498 | 0.495 | 0.514 | 0.565 |
| M2 | 0.048 | 0.295 | 0.372 | 0.413 | 0.427 | 0.441 | 0.511 |

Table 7: Same table as Table 6, except for the order of columns. With the present ordering, all rows have their entries ranked increasingly. For convenience, we abbreviate "daylight" to "d", "urban area" to "ua", "driver's age" to "da", "under influence" to "ui".

They are arranged in subgroups of two to five comparable GCs. In particular, all GCs in a subgroup have the same size class. Moreover, the subgroups are designed in such a way that the experts share a common view on the ordering of GCs by safety in each subgroup. We refer to Table 9 in the appendix for a partial presentation of the thirty-one GCs.

We compute the scores given to each GC in every context of the 2012 BAAC* data set by the meta-algorithm elaborated by super learning with the library presented in Section 6.2. Once the scores are computed, it is easy to rank the GCs by safety in each subgroup. The last column of Table 9 reports the proportions of rankings which coincide with the ranking expected by the experts in each subgroup.

The proportions of rankings in agreement with the experts' expectations equal 58%, 77%, 89%, 96% and 99% for the five subgroups of two GCs. They equal 66%, 77%, 93% and 97% for the four subgroups of three GCs. They equal 43% and 27% for the subgroups of four and five GCs, respectively.

It came as a surprise that five out of the eleven proportions are larger than or equal to 89%, a very large number. Three of these five proportions are associated with subgroups of two GCs, where only two rankings are possible, obviously. In this light, of the two remaining proportions equal to 58% and 77%, only the smaller is disappointing. In subgroups of three GCs, six rankings are possible. In this light, even the smallest proportion of 66% is rather satisfying (random uniform ranking would yield a proportion smaller than 17%). In subgroups of four and five GCs, 24 and 120 rankings are possible, respectively. In this light, the proportions are quite satisfying too (random uniform ranking would yield proportions smaller than 4% and

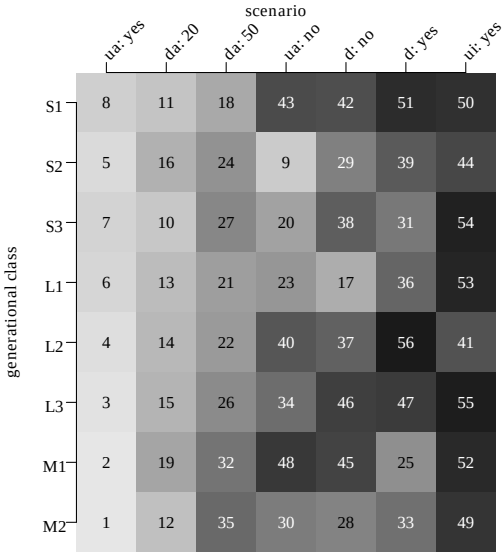


Figure 3: Representing the $8 \times 7 = 56$ scores of Tables 6 and 7. The smaller is the score, the lighter is the color and the safer is the GC in the given context of accident. The gray level is proportional to the score.

1%). Moreover, when focusing on pairwise comparisons within these subgroups, the smallest proportion of rankings in agreement with the experts' expectations equals 58% while the largest equals 98%.

Even though each pair of scores yields a ranking (ties are very unlikely), perhaps the small proportions would be much larger if we only took into account those rankings deemed statistically significant, if such a notion were available. This is not the case yet.

In summary, we believe that these results are quite promising. Our meta-algorithm performs rather well, although the current definition of a GC may not sufficiently capture the essence of the vehicle. Elaborating a notion of statistical confidence, a very delicate problem, will be a priority in future work.

# 7 Discussion

In this article, we address the contextual ranking by passive safety of GCs of light vehicles by elaborating a meta-algorithm. The meta-algorithm is built as a data-adaptive combination of a library of ranking algorithms. An oracle inequality shows the theoretical merit of this ensemble learning approach. To illustrate the use of the meta-algorithm, we rank eight synthetic GCs in seven contexts of accidents derived from a single context by manipulating some elements of its description, and comment on the results. We also rank thirty-one synthetic GCs regrouped in eleven subgroups of comparable GCs in all contexts of the 2012 BAAC* data set, then evaluate the proportions of rankings in agreement with the experts' expectations. The above synthetic GCs are not obtained by averaging a collection of GCs with common date of design, date of entry into service and size class, so none of these synthetic GCs can be interpreted as a typical representant of a class of light vehicles.

The meta-algorithm is contextual (a ranking is conditioned on the occurrence of an accident in a given context) and predictive (it is possible to extrapolate a ranking for any synthetic GC in any context). Based on fleet data and real-life accidents data recorded by the police forces

and gathered in the 2011 and 2012 BAAC* data sets, it is also retrospective.

Our approach is very flexible. If, in the future, the BAAC form included additional relevant information on the accident, such as the violence of impact or a description of the driving assistance systems for active safety embarked in the vehicle, then it would be very easy to use it. Each ranking algorithm in the original library could be modified to account for this new information, yielding a second library. The two libraries could then be merged in a single, richer one. New algorithms could be added as well.

We acknowledge that the meta-algorithm provides ranking from the angle of the law of the BAAC* data sets and not the law of real-life accidents on French public roads in any broader sense. Using capture-recapture methods, the authors of [1, 2, 3, 4] estimate under-reporting correction factors that account for unregistered casualties. The same kind of correction could be implemented in the context of our study, by appropriate weighting.

Inspired by recent advances in causal analysis and epidemiology, we will in future work build upon the present article and go beyond contextual ranking. We will define and address the problem of context-free ranking, treating the contexts of accident like confounding variables. We will also tackle the very delicate problem of the construction of confidence bounds on the scores provided by the meta-algorithm.

**Acknowledgments.** The authors gratefully acknowledge that this research was partially supported by the French National Association for Research and Technology (ANRT) through a CIFRE industrial agreement for training through research. They are very grateful for the reviewers' constructive comments which lead to a much better presentation of our study.

# A Appendix

## A.1 Context of accident from the point of view of one of its actors

Set $1 \leq i \leq n$, $1 \leq k \leq K^i$ and $1 \leq j \leq J^i_k$, where $K^i$ is the number of vehicles involved in the accident described by $\mathbb{O}^i$ and $J^i_k$ is the number of occupants of the vehicle indexed $k$ in that accident. Described in Section 2.1, $\mathbb{O}^i$ includes $W^i_{kj}$, which consists of the following pieces of information, gathered by theme:

*General.*

- Number of vehicles involved in the accident, one or two.
- Number of occupants in the vehicle.

*When and where.*

- Year, month, day of the week, hour when the accident occurred.
- Light condition, either daylight or dark conditions.
- Atmospheric condition, either clear weather, or rain, or other.
- Location of the accident, either outside urban areas (characterized by a number of inhabitants smaller than 5000), or in a small urban area (characterized by a number of inhabitants larger than 5000 and smaller than 300,000), or in a large urban area (either an area with more than 300,000 inhabitants, or the Paris, Hauts-de-Seine, Seine-Saint-Denis, and Val-de-Marne departments).

*What roadway.*

- Intersection, either yes if the accident occurred at an intersection, or no otherwise.
- Infrastructure, either round-about, or other, or unknown.
- Roadway alignment, either straight, or curved, or unknown.
- Roadway profile, either level, or grade, or other, or unknown.

- Roadway surface condition, either dry, or wet, or other, or unknown.
  *What collision.*
- Vehicle responsible of collision, either yes or no.
- Type of collision, either head-on, or rear end, or angle, or other, or no collision.
- Initial contact point, either front, or left-front half, or right-front half, or back, or left-back half, or right-back half, or left, or right, or multiple collisions, or none.
- Fixed obstacle, either building, parapet, wall, or crash barrier, or ditch, embankment slope, rock face, or no obstacle, or parked vehicle, or pole, or tree, or other, or unknown.
- Moving obstacle, either vehicle, or other, or unknown.
  *Which driver.*
- Age, and gender of driver.
- Was the driver's seatbelt fastened, either yes or no.
- Socio-professional category of driver, either artisan, farmer, tradesman, or executive, or professional driver, or retiree, or student, or unemployed, or worker, or other.
- Driver under the influence of alcohol, either yes or no.
- Driver's license status, either valid, or invalid, or learner's permit, or unknown.
- Owner of vehicle, either yes, or no, or unknown.
  *Which occupant.*
- Age, gender of the occupant.
- Socio-professional category of the occupant (same levels as previously presented).
- Role of occupant, either driver or other.
- Seating position of the occupant.
- Was the occupant's seatbelt fastened, either yes or no.

## A.2 ROC curve, AUC, and proofs of results stated in Section 3

**ROC curve, AUC.** The ROC curve of a scoring function $s : \mathcal{Y} \to [0,1]$ is defined by plotting $\mathrm{TPR}_s(t) = P(s(Y) \geq t | Z = 1)$ against $\mathrm{FPR}_s(t) = P(s(Y) \geq t | Z = 0)$. The acronym ROC stands for "receiver operating curve", see [16]. The acronyms TPR and FPR correspond to the expressions "true positive rate" and "false positive rate". They refer to the test of whether $Y$ is drawn from the conditional distribution $P(\cdot|Z=0)$ (null hypothesis) or from the conditional distribution $P(\cdot|Z=1)$ (alternative hypothesis) based on a decision rule of the form "reject the null if $s(Y) \geq t$". In this light, the ROC curve can be seen as the graph of the power of the test as a function of its level $\alpha$, *i.e.*, of the function $\alpha \mapsto \beta_s(\alpha) = \mathrm{TPR}_s(\inf\{t \in (0,1) : \mathrm{FPR}_s(t) \leq \alpha\})$ which maps $[0,1]$ to $[0,1]$.

If $Y$ and $Z$ are independent under $P$, then $\mathrm{TPR}_s = \mathrm{FPR}_s$ and the ROC curve is the diagonal segment $\{(\alpha, \alpha) : \alpha \in [0,1]\}$. The Neyman-Pearson lemma implies that $\beta_{Q_0}$ necessarily dominates $\beta_s$ [11, Proposition B.1]: for all $\alpha \in [0,1]$, $\beta_{Q_0}(\alpha) \geq \beta_s(\alpha)$. Thus, the area under the curve defined as $\mathrm{AUC}_s = \int_0^1 \beta_s(\alpha) d\alpha$ is a measure of how well the above test performs: the larger is $\mathrm{AUC}_s \leq \mathrm{AUC}_{Q_0}$, the better the test statistically performs.

**Proofs of (3), (4), (5), (6).** Following [11, Example 1], note that

$$\begin{aligned} E_{P^{\otimes 2}}\left(L^0(r,O,O')\right) &= E_{P^{\otimes 2}}\big(\mathbf{1}\{r(Y,Y')=1\}\mathbf{1}\{Z > Z'\} \\ &\qquad +\mathbf{1}\{r(Y,Y')=-1\}\mathbf{1}\{Z < Z'\}\big) \qquad (17) \\ &= E_{P^{\otimes 2}}\big(\mathbf{1}\{r(Y,Y')=1\}P^{\otimes 2}(Z > Z'|Y,Y') \\ &\qquad +\mathbf{1}\{r(Y,Y')=-1\}P^{\otimes 2}(Z < Z'|Y,Y')\big) \\ &= E_{P^{\otimes 2}}\big(\mathbf{1}\{r(Y,Y')=1\}Q_0(Y)(1-Q_0)(Y') \\ &\qquad +\mathbf{1}\{r(Y,Y')=-1\}(1-Q_0)(Y)Q_0(Y')\big)\,. \qquad (18) \end{aligned}$$

The above RHS expression is minimized at $r = r_0$, hence (4) and the LHS inequality in (5). Moreover, using that $2\min(Q_0(Y), Q_0(Y'))$ equals $Q_0(Y) + Q_0(Y') - |Q_0(Y) - Q_0(Y')|$, it holds that

$$\begin{aligned} E_{P^{\otimes 2}}\left(L^0(r_0, O, O')\right) &= E_{P^{\otimes 2}}\left(\min(Q_0(Y), Q_0(Y'))\right) - E_P\left(Q_0(Y)\right)^2 \\ &= E_P\left(Q_0(Y)\right) - E_P\left(Q_0(Y)\right)^2 - \frac{1}{2} E_{P^{\otimes 2}}\left(|Q_0(Y) - Q_0(Y')|\right) \\ &= \mathrm{Var}_P(Z) - \frac{1}{2} E_{P^{\otimes 2}}\left(|Q_0(Y) - Q_0(Y')|\right), \end{aligned}$$

as stated in (3). It remains to prove the RHS inequality in (5) for $r = r_s$ with $s : \mathcal{Y} \to [0, 1]$ a scoring function such that $P^{\otimes 2}(s(Y) = s(Y')) = 0$. First, note that (18) implies

$$\begin{aligned} &E_{P^{\otimes 2}}\left(L^0(r_s, O, O')\right) - E_{P^{\otimes 2}}\left(L^0(r_0, O, O')\right) \\ &\quad = E_{P^{\otimes 2}}\big((\mathbf{1}\{r_s(Y, Y') = 1\} - \mathbf{1}\{r_0(Y, Y') = 1\})Q_0(Y)(1 - Q_0)(Y') \\ &\qquad\qquad + (\mathbf{1}\{r_s(Y, Y') = -1\} - \mathbf{1}\{r_0(Y, Y') = -1\})(1 - Q_0)(Y)Q_0(Y')\big) \\ &\quad = E_{P^{\otimes 2}}\big(\mathbf{1}\{(s(Y) - s(Y'))(Q_0(Y) - Q_0(Y')) < 0\} \\ &\qquad\qquad \times (\mathbf{1}\{Q_0(Y) \le Q_0(Y')\}(Q_0(Y') - Q_0(Y)) \\ &\qquad\qquad\quad + \mathbf{1}\{Q_0(Y) \ge Q_0(Y')\}(Q_0(Y) - Q_0(Y'))\big)\big) \\ &\quad = E_{P^{\otimes 2}}\left(\mathbf{1}\{(s(Y) - s(Y'))(Q_0(Y) - Q_0(Y')) < 0\}|Q_0(Y) - Q_0(Y')|\right). \end{aligned} \tag{19}$$

Second, if $Q_0(Y) \le Q_0(Y')$ and $s(Y) \ge s(Y')$, then

$$\begin{aligned} |Q_0(Y) - Q_0(Y')| &= Q_0(Y') - Q_0(Y) = Q_0(Y') - s(Y') + s(Y') - Q_0(Y) \\ &\le Q_0(Y') - s(Y') + s(Y) - Q_0(Y) = |Q_0(Y') - s(Y') + s(Y) - Q_0(Y)| \\ &\le |Q_0(Y') - s(Y')| + |s(Y) - Q_0(Y)|, \end{aligned} \tag{20}$$

and if $Q_0(Y) \ge Q_0(Y')$ and $s(Y) \le s(Y')$, then

$$\begin{aligned} |Q_0(Y) - Q_0(Y')| &= Q_0(Y) - Q_0(Y') = Q_0(Y) - s(Y) + s(Y) - Q_0(Y') \\ &\le Q_0(Y) - s(Y) + s(Y') - Q_0(Y') = |Q_0(Y) - s(Y) + s(Y') - Q_0(Y')| \\ &\le |Q_0(Y) - s(Y)| + |s(Y') - Q_0(Y')|. \end{aligned} \tag{21}$$

Combining (19), (20) and (21) yields

$$\begin{aligned} E_{P^{\otimes 2}}\left(L^0(r, O, O')\right) - E_{P^{\otimes 2}}\left(L^0(r_0, O, O')\right) &\le E_{P^{\otimes 2}}\left(|Q_0(Y) - s(Y)| + |s(Y') - Q_0(Y')|\right) \\ &= 2E_P\left(|Q_0(Y) - s(Y)|\right), \end{aligned}$$

which completes the proof of (5).

We now turn to (6). As already mentioned, the inequality $\mathrm{AUC}_s \le \mathrm{AUC}_{Q_0}$ is a direct by-product of [11, Proposition B.1]. The equality

$$P^{\otimes 2}\left(s(Y) \ge s(Y') | Z = 1, Z' = 0\right) = \mathrm{AUC}_s$$

is guaranteed by [11, Proposition B.2]. Set $p = P(Z = 1)$ hence $p(1 - p) = P^{\otimes 2}(Z = 1, Z' = 0)$. Obviously,

$$\begin{aligned} p(1-p)\left(1 - P^{\otimes 2}\left(s(Y) \ge s(Y') | Z = 1, Z' = 0\right)\right) &= P^{\otimes 2}\left(s(Y) < s(Y'), (Z, Z') = (1, 0)\right) \\ &= P^{\otimes 2}\left(s(Y') < s(Y), (Z', Z) = (1, 0)\right), \end{aligned}$$

where the second equality holds because $Y, Y'$ are exchangeable. Summing up the above equalities and using $P^{\otimes 2}(s(Y) = s(Y')) = 0$ yield

$$2p(1-p)\left(1 - P^{\otimes 2}\left(s(Y) \ge s(Y') \mid Z = 1, Z' = 0\right)\right)$$

$$
\begin{aligned}
&= P^{\otimes 2}\left(s(Y) < s(Y'), (Z, Z') = (1,0)\right) \\
&\quad + P^{\otimes 2}\left(s(Y') < s(Y), (Z', Z) = (1,0)\right) \\
&= P^{\otimes 2}\left((Z - Z')r_s(Y, Y') > 0, (Z, Z') = (1,0)\right) \\
&\quad + P^{\otimes 2}\left((Z - Z')r_s(Y, Y') > 0, (Z, Z') = (0,1)\right) \\
&= E_{P^{\otimes 2}}\left(L(r_s, O, O')\right),
\end{aligned}
$$

hence the equality in (6).

### A.3 Proof of Lemma 1

For $\mathbb{O}$ drawn from $\mathbb{P}$, we denote $K$ the corresponding number of vehicles involved in the accident and "$J_1, \ldots, J_K$" (respectively, "$\Delta_1, \ldots, \Delta_K$") for either $J_1$, the number of occupants of the sole vehicle involved (respectively, $\Delta_1$, the missingness indicator of this vehicle) when $K = 1$ or $(J_1, J_2)$, both numbers of occupants of the two vehicles involved (respectively, $(\Delta_1, \Delta_2)$, both missingness indicators of GCs) otherwise. The proof mainly relies on the tower rule, which justifies the first and fifth equalities below, on assumptions **A1** and **A2**, which justify the third one, and on the fact that the conditional distributions of $J_1$ and $J_2$ given $K = 2$ coincide, which justifies the last but one equality:

$$
\begin{aligned}
E_{\mathbb{P}}(\mathcal{W}(f_1)(\mathbb{O})) &= E_{\mathbb{P}}\left(E_{\mathbb{P}}\left(\mathcal{W}(f_1)(\mathbb{O})|K, J_1, \ldots, J_K, \Delta_1, \ldots, \Delta_K\right)\right) \\
&= E_{\mathbb{P}}\left(\frac{1}{K}\sum_{k=1}^{K}\frac{\mathbf{1}\{\Delta_k = 1\}}{J_k \pi(J_1 \ldots J_K)}\right. \\
&\qquad \left.\times \sum_{j=1}^{J_k} E_{\mathbb{P}}\left(f_1(O_{kj})\middle| K, J_1, \ldots, J_K, \Delta_1, \ldots, \Delta_K\right)\right) \\
&= E_{\mathbb{P}}\left(\frac{1}{K}\sum_{k=1}^{K}\frac{\mathbf{1}\{\Delta_k = 1\}}{J_k \pi(J_1 \ldots J_K)}\sum_{j=1}^{J_k} E_{P_{KJ_k}}(f_1(O))\right) \\
&= E_{\mathbb{P}}\left(\frac{1}{K}\sum_{k=1}^{K}\frac{\mathbf{1}\{\Delta_k = 1\}}{\pi(J_1 \ldots J_K)} E_{P_{KJ_k}}(f_1(O))\right) \\
&= E_{\mathbb{P}}\left(\frac{1}{K}\sum_{k=1}^{K}\frac{E_{\mathbb{P}}(\mathbf{1}\{\Delta_k = 1\}|K, J_1, \ldots, J_K)}{\pi(J_1 \ldots J_K)} E_{P_{KJ_k}}(f_1(O))\right) \\
&= E_{\mathbb{P}}\left(\frac{1}{K}\sum_{k=1}^{K} E_{P_{KJ_k}}(f_1(O))\right) \\
&= E_{\mathbb{P}}\left(\sum_{j_1=1}^{J_{\max}} \mathbf{1}\{K = 1, J_1 = j_1\} E_{P_{1j_1}}(f_1(O))\right. \\
&\qquad \left.+ \sum_{j_1=1}^{J_{\max}}\sum_{j_2=1}^{J_{\max}} \mathbf{1}\{K = 2, J_1 = j_1, J_2 = j_2\}\frac{1}{2}\sum_{k=1}^{2} E_{P_{2j_k}}(f_1(O))\right) \\
&= \sum_{j_1=1}^{J_{\max}} \mathbb{P}(K = 1, J_1 = j_1) E_{P_{1j_1}}(f_1(O)) \\
&\qquad + \sum_{j_1=1}^{J_{\max}} \frac{1}{2} E_{P_{2j_1}}(f_1(O)) \sum_{j_2=1}^{J_{\max}} \mathbb{P}(K = 2, J_1 = j_1, J_2 = j_2)
\end{aligned}
$$

$$
\begin{aligned}
&+\sum_{j_2=1}^{J_{\max}} \frac{1}{2} E_{P_{2j_2}}(f_1(O)) \sum_{j_1=1}^{J_{\max}} \mathbb{P}(K=2, J_1=j_1, J_2=j_2) \\
= &\sum_{j_1=1}^{J_{\max}} \mathbb{P}(K=1, J_1=j_1) E_{P_{1j_1}}(f_1(O)) \\
&+\sum_{j_1=1}^{J_{\max}} \mathbb{P}(K=2, J_1=j_1) E_{P_{2j_1}}(f_1(O)) \\
= &E_P(f_1(O)).
\end{aligned}
$$

This completes the proof.

### A.4 Proofs of Proposition 2 and 3

*Proof of Proposition 2.* We start with the a series of inequalities and equalities. Inequality (22) follows from (8) and (10); it is valid to replace $\widehat{k}_n$ with $\widetilde{k}_n$ as we do in the last RHS term of (23) because $\widehat{\mathcal{R}}_n(\widehat{k}_n) \leq \widehat{\mathcal{R}}_n(k)$ for all $1 \leq k \leq \mathcal{K}_n$; (24) is obtained from (23) by rearranging terms:

$$
\begin{aligned}
0 \leq\ & \widetilde{\mathcal{R}}_n(\widehat{k}_n) - \mathcal{R}(\theta_0) && (22)\\
=\ & E_{B_n}\left(P^{\otimes 2}\left(\ell(\widehat{\Psi}_{\widehat{k}_n}[\mathbb{P}_{n,B_n,0}]) - \ell(\theta_0)\right)\right) \\
& -(1+\delta) E_{B_n}\left(P_{n,B_n,1}^{\otimes 2}\left(\ell(\widehat{\Psi}_{\widehat{k}_n}[\mathbb{P}_{n,B_n,0}]) - \ell(\theta_0)\right)\right) \\
& +(1+\delta) E_{B_n}\left(P_{n,B_n,1}^{\otimes 2}\left(\ell(\widehat{\Psi}_{\widehat{k}_n}[\mathbb{P}_{n,B_n,0}]) - \ell(\theta_0)\right)\right) \\
\leq\ & E_{B_n}\left(P^{\otimes 2}\left(\ell(\widehat{\Psi}_{\widehat{k}_n}[\mathbb{P}_{n,B_n,0}]) - \ell(\theta_0)\right)\right) \\
& -(1+\delta) E_{B_n}\left(P_{n,B_n,1}^{\otimes 2}\left(\ell(\widehat{\Psi}_{\widehat{k}_n}[\mathbb{P}_{n,B_n,0}]) - \ell(\theta_0)\right)\right) \\
& +(1+\delta) E_{B_n}\left(P_{n,B_n,1}^{\otimes 2}\left(\ell(\widehat{\Psi}_{\widetilde{k}_n}[\mathbb{P}_{n,B_n,0}]) - \ell(\theta_0)\right)\right) && (23)\\
=\ & (1+2\delta)\left(\widetilde{\mathcal{R}}_n(\widetilde{k}_n) - \mathcal{R}(\theta_0)\right) + E_{B_n}(U_{\widehat{k}_n} + V_{\widetilde{k}_n}), && (24)
\end{aligned}
$$

where we introduce, for each $1 \leq k \leq \mathcal{K}_n$,

$$
\begin{aligned}
\widehat{H}_k &= P_{n,B_n,1}^{\otimes 2}\left(\ell(\widehat{\Psi}_k[\mathbb{P}_{n,B_n,0}]) - \ell(\theta_0)\right), \\
\widetilde{H}_k &= P^{\otimes 2}\left(\ell(\widehat{\Psi}_k[\mathbb{P}_{n,B_n,0}]) - \ell(\theta_0)\right), \\
U_k &= (1+\delta)(\widetilde{H}_k - \widehat{H}_k) - \delta \widetilde{H}_k, \quad \text{and} \\
V_k &= (1+\delta)(\widehat{H}_k - \widetilde{H}_k) - \delta \widetilde{H}_k.
\end{aligned}
$$

Since the distribution of $B_n$ is discrete, $E_{\mathbb{P}}(E_{B_n}(U_{\widehat{k}_n} + V_{\widetilde{k}_n})) = E_{B_n}(E_{\mathbb{P}}(U_{\widehat{k}_n} + V_{\widetilde{k}_n})) =$. Therefore, it is sufficient to show that the conditional expectations of $\max_{1\leq k\leq \mathcal{K}_n} U_k$ and $\max_{1\leq k\leq \mathcal{K}_n} V_k$ given $B_n$ and $\mathbb{P}_{n,B_n,0}$ are both smaller than half the RHS term in (14) to derive (14) from (24).

We now work conditionally on $B_n$ and $\mathbb{P}_{n,B_n,0}$, and draw inspiration from the proof of Lemma 8.2 in [36]. Let $T_k$ be equal to either $U_k$ or $V_k$ for all $1 \leq k \leq \mathcal{K}_n$. The (conditional) expectation under $\mathbb{P}$ of $\max_{1\leq k\leq \mathcal{K}_n} T_k$ can be written $E_{P^{\otimes 2}}(\max_{1\leq k\leq \mathcal{K}_n} T_k)$. Arbitrarily set $t > 0$, $1 \leq k \leq \mathcal{K}_n$, and introduce

$$
\sigma_k^2 = \mathrm{Var}_P\left(E_{P^{\otimes 2}}\left[(\ell(\widehat{\Psi}_k[\mathbb{P}_{n,B_n,0}])) - \ell(\theta_0))(O, O') \middle| O\right]\right),
$$

$\lambda_k = \delta\sqrt{np}\widetilde{H}_k$, $v_k = 4(1+\delta)^2\sigma_k^2$, $b = 65(1+\delta)c_1/\sqrt{np}$, and $r_k = v_k/b - \lambda_k$. The Bernstein inequality for $U$-processes of [5, Theorem 2] yields that

$$P^{\otimes 2}\left(\sqrt{np}\, T_k \geq t\right) \leq 4\exp\left(-\frac{1}{2}\frac{(t+\lambda_k)^2}{v_k + b(t+\lambda_k)}\right).$$

But

$$\frac{(t+\lambda_k)^2}{v_k + b(t+\lambda_k)} \geq \frac{t^{2-\alpha}\lambda_k^\alpha + \lambda_k^2}{2v_k}\mathbf{1}\{t \leq r_k\} + \frac{t+\lambda_k}{2b}\mathbf{1}\{t > r_k\},$$

hence

$$\begin{aligned} P^{\otimes 2}\left(\sqrt{np}\, T_k \mathbf{1}\{\sqrt{np}\, T_k \leq r_k\} \geq t\right) &\leq 4\exp\left(-\frac{t^{2-\alpha}\lambda_k^\alpha + \lambda_k^2}{4v_k}\right), \\ P^{\otimes 2}\left(\sqrt{np}\, T_k \mathbf{1}\{\sqrt{np}\, T_k < r_k\} \geq t\right) &\leq 4\exp\left(-\frac{t+\lambda_k}{4b}\right). \end{aligned}$$

Consequently, Lemma 8.1 in [36] implies

$$E_{P^{\otimes 2}}\left(\sqrt{np}\max_{1\leq k\leq \mathcal{K}_n} T_k\right) \leq 8\left(\max_{1\leq k\leq \mathcal{K}_n}\left(\frac{v_k}{\lambda_k^\alpha}\right)^{1/(2-\alpha)} + b\right) \times \left(\log\left(1+4\mathcal{K}_n\right)\right)^{1/(2-\alpha)}. \tag{25}$$

By assumption, $\max_{1\leq k\leq \mathcal{K}_n}(v_k/\lambda_k^\alpha) \leq 4(1+\delta)^2 c_2/(\delta\sqrt{np})^\alpha$. Hence, using $2-\alpha \geq 1$, (25) yields the following, slightly looser inequality:

$$E_{P^{\otimes 2}}\left(\max_{1\leq k\leq \mathcal{K}_n} T_k\right) \leq \frac{c_3}{2}\frac{\log(1+4\mathcal{K}_n)}{(np)^{1/(2-\alpha)}}.$$

This completes the proof. □

*Proof of Proposition 3.* Condition (12) holds with $c_1 = 1$. To alleviate notation, introduce $\Delta\ell(\theta)$ given, for each $\theta \in \Theta$, by $\Delta\ell(\theta)(O,O') = \ell(\theta)(O,O') - \ell(\theta_0)(O,O')$. Set $\theta \in \Theta$. A case-by-case analysis reveals that

$$\begin{aligned} |\Delta\ell(\theta)(O,O')| &= \big|\big(\mathbf{1}\{\theta(Y) \leq \theta(Y')\} - \mathbf{1}\{Q_0(Y) \leq Q_0(Y')\}\big) \times \mathbf{1}\{Z=1, Z'=0\} \\ &\quad + \big(\mathbf{1}\{\theta(Y) > \theta(Y')\} - \mathbf{1}\{Q_0(Y) > Q_0(Y')\}\big) \times \mathbf{1}\{Z=0, Z'=1\}\big| \\ &= \big|\mathbf{1}\{\theta(Y) \leq \theta(Y')\} - \mathbf{1}\{Q_0(Y) \leq Q_0(Y')\}\big| \times \mathbf{1}\{Z \neq Z'\} \\ &= \mathbf{1}\{(\theta(Y)-\theta(Y'))(Q_0(Y)-Q_0(Y')) < 0\} \times \mathbf{1}\{Z \neq Z'\} \\ &\leq \mathbf{1}\{(\theta(Y)-\theta(Y'))(Q_0(Y)-Q_0(Y')) < 0\}. \end{aligned} \tag{26}$$

(A similar argument appears in the proof of the RHS of (5).) Moreover, it also holds that $|E_{P^{\otimes 2}}\left(\Delta\ell(\theta)(O,O')\big|O\right)| \leq E_{P^{\otimes 2}}\left(|\Delta\ell(\theta)(O,O')|\big|O\right)$. Therefore,

$$\begin{aligned} &\mathrm{Var}_P\left(E_{P^{\otimes 2}}\left(\Delta\ell(\theta)(O,O')\big|O\right)\right) \\ &\quad \leq E_P\left(E_{P^{\otimes 2}}(\Delta\ell(\theta)(O,O')\big|O)^2\right) \qquad (27) \\ &\quad \leq E_P\left(E_{P^{\otimes 2}}\left(|\Delta\ell(\theta)(O,O')|\big|O\right)^2\right) \\ &\quad \leq E_P\left(\left[E_{P^{\otimes 2}}\left(\mathbf{1}\{(\theta(Y)-\theta(Y'))(Q_0(Y)-Q_0(Y')) < 0\}\big|Y\right)\right]^2\right), \qquad (28) \end{aligned}$$

where the final inequality follows from (26) and the independence of $Y'$ with respect to $O = (Y,Z)$. Using the Cauchy-Schwarz inequality now yields

$$\begin{aligned} &\mathrm{Var}_P\left(E_{P^{\otimes 2}}\left(\Delta\ell(\theta)(O,O')\big|O\right)\right) \\ &\quad \leq E_P\left(\left(E_P\left[\mathbf{1}\{(\theta(Y)-\theta(Y'))(Q_0(Y)-Q_0(Y')) < 0\} \times |Q_0(Y)-Q_0(Y')|^\alpha\big|Y\right]\right) \end{aligned}$$

$$\times \left(E_P\left[|Q_0(Y) - Q_0(Y')|^{-\alpha}\middle|Y\right]\right)$$

which, by (16) and Jensen's inequality, implies in turn

$$\begin{aligned}&\mathrm{Var}_P\left(E_{P^{\otimes 2}}\left(\Delta\ell(\theta)(O,O')\middle|O\right)\right)\\&\quad\leq c_2 E_{P^{\otimes 2}}\left(\mathbf{1}\{(\theta(Y)-\theta(Y'))(Q_0(Y)-Q_0(Y'))<0\}\times|Q_0(Y)-Q_0(Y')|\right)^{\alpha}.\end{aligned}$$

Note that the RHS of the above display can be rewritten $E_{P^{\otimes 2}}\left(\Delta\ell(\theta)(O,O')\right)$ by (19). Therefore, we have shown that

$$\frac{\mathrm{Var}_P\left(E_{P^{\otimes 2}}\left(\Delta\ell(\theta)(O,O')\middle|O\right)\right)}{E_{P^{\otimes 2}}\left(\Delta\ell(\theta)(O,O')\right)}\leq c_2.$$

By taking the supremum over $\theta\in\Theta$, we thus conclude that (16) implies (13) as stated in Proposition 3. □

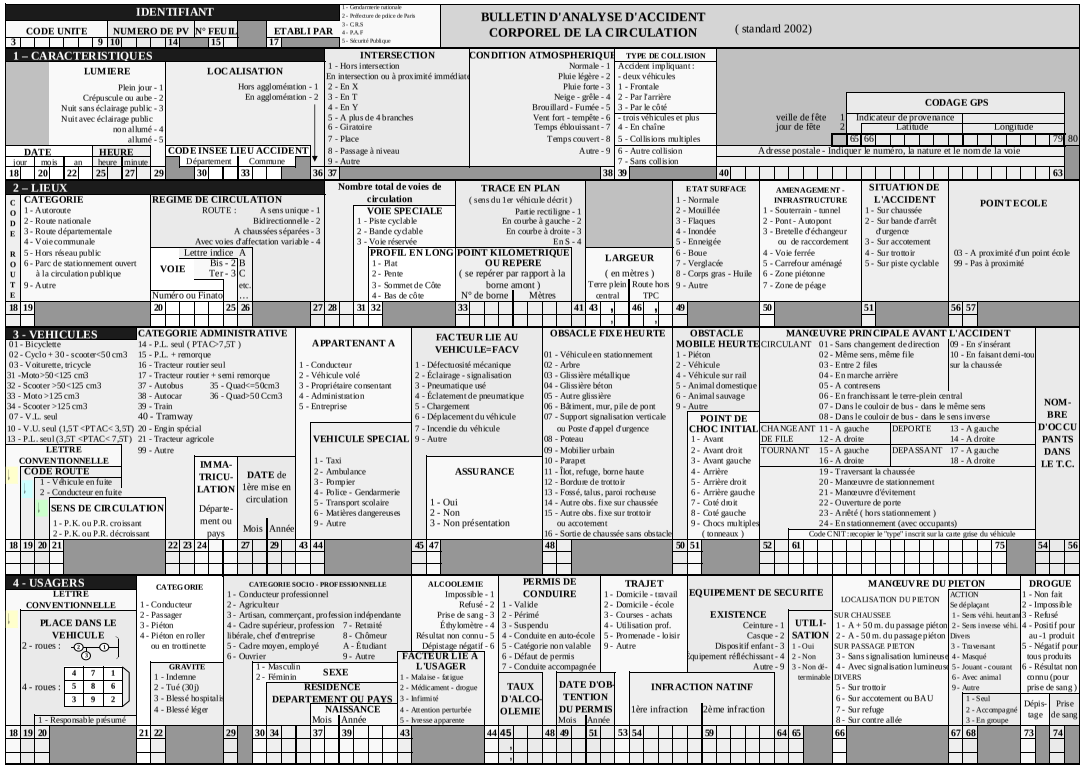


Figure 4: Form for the French national file of personal accidents (*Bulletin d'Analyse d'Accident Corporel de la Circulation*, or BAAC form).

## References


[1] E. Amoros. *Non-fatal road casualties: estimation of frequency and injury severity , France 1996-2006, modelled from a medical registry (Rhône area) and police data (France)*. Phd thesis, Université Claude Bernard–Lyon I, 2007. URL `https://tel.archives-ouvertes.fr/tel-00511718`.

[2] E. Amoros, J-L. Martin, and B. Laumon. Under-reporting of road crash casualties in france. *Accident Analysis & Prevention*, 38(4):627–635, 2006.

[3] E. Amoros, J-L. Martin, and B. Laumon. Estimating non-fatal road casualties in a large french county, using the capture–recapture method. *Accident Analysis & Prevention*, 39 (3):483–490, 2007.

[4] E. Amoros, J-L. Martin, S. Lafont, and B. Laumon. Actual incidences of road casualties, and their injury severity, modelled from police and hospital data, france. *The European Journal of Public Health*, 18(4):360–365, 2008.

[5] M. A. Arcones. A Bernstein-type inequality for $U$-statistics and $U$-processes. *Statist. Probab. Lett.*, 22(3):239–247, 1995.

[6] N. Baskiotis. *TreeRank*, 2010. URL `http://treerank.sourceforge.net/`. R package version 1.0-0.

[7] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[8] L. Breiman. Stacked regressions. *Machine learning*, 24(1):49–64, 1996.

[9] A. Chambaz, D. Choudat, C. Huber, J-C. Pairon, and M. J. van der Laan. Analysis of the effect of occupational exposure to asbestos based on threshold regression modeling of case-control data. *Biostatistics*, 15(2):327–340, 2014.

[10] S. Clémençon and N. Vayatis. Tree-based ranking methods. *IEEE Trans. Inform. Theory*, 55(9):4316–4336, 2009.

[11] S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and empirical minimization of $U$-statistics. *Ann. Statist.*, 36(2):844–874, 2008.

[12] DaCoTa EU project team. Safety ratings. Technical report, European Commission Directorate General for Mobility & Transport, 2013. Deliverable 4.8r of the EC FP7 project DaCoTA.

[13] European Commission. How safe are your roads? Commission road safety statistics show small improvement for 2014. European Commission Press Release IP-15-4656, March 24th, 2015.

[14] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *J. Mach. Learn. Res.*, 4(6):933–969, 2004.

[15] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL `http://www.jstatsoft.org/v33/i01/`.

[16] D. M. Green and J. A. Swets. *Signal detection theory and psychophysics*. Wiley, New-York, 1966.

[17] J. Hackney and C. Kahane. The New Car Assessment Program: Five star rating system and vehicle safety performance characteristics. Technical Report 950888, SAE International, 1995. doi:10.4271/950888.

[18] T. Hastie. *gam: Generalized Additive Models*, 2014. URL `http://CRAN.R-project.org/package=gam`. R package version 1.09.1.

[19] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: A tutorial. *Statist. Sci.*, 14(4):382–417, 1999. ISSN 0883-4237. doi: 10.1214/ss/1009212519. URL `http://dx.doi.org/10.1214/ss/1009212519`. With comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors.

[20] M. Jonker and A. W. van der Vaart. On the correction of the asymptotic distribution of the likelihood ratio statistic if nuisance parameters are estimated based on an external source. *International Journal of Biostatistics*, 10(2):123–142, 2014.

[21] C. Kooperberg. *polspline: Polynomial spline routines*, 2013. URL `http://CRAN.R-project.org/package=polspline`. R package version 1.1.9.

[22] A. Kullgren, A. Lie, and C. Tingvall. Comparison between Euro NCAP test results and real-world crash data. *Traffic Inj. Prev.*, 11(6):587–593, 2010.

[23] E. LeDell, M. L. Petersen, and M. J. van der Laan. Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. Technical Report 304, U.C. Berkeley Division of Biostatistics Working Paper Series, 2012. In review.

[24] E. LeDell, M. J. van der Laan, and M. L. Petersen. AUC-maximizing ensembles through metalearning. *International Journal of Biostatistics*, 12(1):203–218, 2016.

[25] A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3): 18–22, 2002. URL `http://CRAN.R-project.org/doc/Rnews/`.

[26] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, 2014. URL `http://CRAN.R-project.org/package=e1071`. R package version 1.6-4.

[27] A. Peters and T. Hothorn. *ipred: Improved Predictors*, 2013. URL `http://CRAN.R-project.org/package=ipred`. R package version 0.9-3.

[28] E. Polley and M. J. van der Laan. *SuperLearner: Super Learner Prediction*, 2014. URL `http://CRAN.R-project.org/package=SuperLearner`. R package version 2.0-15.

[29] E. C. Polley, S. Rose, and M. J. van der Laan. Super learning. In *Targeted learning*, Springer Ser. Statist., pages 43–66. Springer, New York, 2011.

[30] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL `http://www.R-project.org/`.

[31] G. Ridgeway and and others. *gbm: Generalized Boosted Regression Models*, 2015. URL `http://CRAN.R-project.org/package=gbm`. R package version 2.1.1.

[32] S. Robbiano. Upper bounds and aggregation in bipartite ranking. *Electronic Journal of Statistics*, 7:1249–1271, 2013.

[33] S. Rose and M. J. van der Laan. Simple optimal weighting of cases and controls in case-control studies. *International Journal of Biostatistics*, 4, 2008.

[34] R. E. Schapire. The strength of weak learnability. *Machine learning*, 5(2):197–227, 1990.

[35] M. J. van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. *Stat. Appl. Genet. Mol. Biol.*, 6:Art. 25, 23, 2007.

[36] A. W. van der Vaart, S. Dudoit, and M. J. van der Laan. Oracle inequalities for multi-fold cross validation. *Statist. Decisions*, 24(3):351–371, 2006.

[37] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL `http://www.stats.ox.ac.uk/pub/MASS4`. ISBN 0-387-95457-0.

[38] D. H. Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

| $j_1$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| | 40.58 | 26.75 | 14.60 | 10.29 | 6.41 | 0.96 | 0.40 |
| | $\left(\widehat{\mathbb{P}}(J_1 = j_1 \mid K = 1)\right)$ | | | | | | |

| $j_1$ \ $j_2$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 1 | 53.93 | 27.02 | 7.04 | 2.51 | 0.83 | 0.05 | 0.05 |
| 2 | | 3.55 | 2.40 | 1.07 | 0.35 | 0.08 | 0.05 |
| 3 | | | 0.40 | 0.32 | 0.05 | 0 | 0 |
| 4 | | | | 0.05 | 0.16 | 0 | 0 |
| 5 | | | | | 0.03 | 0.03 | 0.03 |
| 6 | | | | | | 0 | 0 |
| 7 | | $\left(\widehat{\mathbb{P}}(\{J_1, J_2\} = \{j_1, j_2\} \mid K = 2)\right)$ | | | | | 0 |

Table 8: Distributions of numbers of occupants of vehicles when the accident involves only one vehicle (top table) or two vehicles (bottom table), as estimated based on the 2012 BAAC* data set. The estimates are reported in percents to show the smallest values. For every pair $\{j_1, j_2\}$ observed among the 8,827 two-vehicle accidents, $\widehat{\mathbb{P}}(\{J_1, J_2\} = \{j_1, j_2\} | K = 2)$ is the ratio of number of accidents involving $j_1$ and $j_2$ occupants divided by 8,827. Setting $f_1(O) = \mathbf{1}\{\{J_1, J_2\} = \{j_1, j_2\}\}$, $\widehat{\mathbb{P}}(\{J_1, J_2\} = \{j_1, j_2\} | K = 2) = E_{\mathbb{P}_n}(\mathcal{W}(f_1)(\mathbb{O}))/\mathbb{P}_n(K = 2)$, see Lemma 1. In the top table, the sum of the five largest probabilities is close to 99%, showing that the vast majority of one-vehicle accidents involve no more than five occupants. In the bottom table, the numerical values 0.05, 0.03, and 0.00 correspond to three, two, and one accidents. The sum of the 10 largest probabilities among the 28 is larger than 99%.

| | generational class (GC) | | | percentage of |
|---|---|---|---|---|
| GC code | dd | deis | size class | expected ranking |
| SMa1 | 1983 | 1995 | supermini car | 97% |
| SMa2 | 1998 | 2006 | supermini car | |
| SMa3 | 2005 | 2009 | supermini car | |
| SMb1 | 1996 | 2004 | supermini car | 77% |
| SMb2 | 2005 | 2009 | supermini car | |
| SMc1 | 1991 | 2002 | supermini car | 58% |
| SMc2 | 2005 | 2009 | supermini car | |
| Sa1 | 1990 | 1997 | small family car | 93% |
| Sa2 | 1998 | 2006 | small family car | |
| Sa3 | 2005 | 2009 | small family car | |
| Sb1 | 1995 | 2001 | small family car | 66% |
| Sb2 | 2002 | 2007 | small family car | |
| Sb3 | 2008 | 2010 | small family car | |
| Sc1 | 1993 | 2000 | small family car | 77% |
| Sc2 | 2001 | 2007 | small family car | |
| Sc3 | 2007 | 2010 | small family car | |
| La1 | 1991 | 1996 | large family car | 43% |
| La2 | 1997 | 2004 | large family car | |
| La3 | 2004 | 2008 | large family car | |
| La4 | 2010 | 2011 | large family car | |
| Lb1 | 1983 | 1994 | large family car | 27% |
| Lb2 | 1991 | 1995 | large family car | |
| Lb3 | 1997 | 2003 | large family car | |
| Lb4 | 2003 | 2007 | large family car | |
| Lb5 | 2008 | 2010 | large family car | |
| Ma1 | 1994 | 1999 | minivan | 99% |
| Ma2 | 2002 | 2006 | minivan | |
| Mb1 | 1994 | 1998 | minivan | 96% |
| Mb2 | 2002 | 2005 | minivan | |
| Mc1 | 1994 | 2001 | minivan | 89% |
| Mc2 | 2002 | 2004 | minivan | |

Table 9: Thirty-one synthetic GCs. We only report the dates of design (dd), dates of entry into service (deis), size classes, and give each GC a code for reference. The above GCs are not obtained by averaging a collection of GCs with common date of design, date of entry into service and size class, so none of them can be interpreted as a typical representant of a certain class of light vehicles. For a given prefix X, experts expect that X$i$ is safer than X$j$ whenever $i > j$. The last column reports the proportions of contexts (among the 15,852 contexts of the 2012 BAAC* data set) where the rankings of GCs sharing the same prefix are in agreement with the experts' expectations.