From contextual to global rankings by passive safety of generational classes of light vehicles

Zaïd Ouni, Cyril Chauvel, Antoine Chambaz

From contextual to global rankings by passive safety of generational classes of light vehicles

Z. Ouni[1,2], C. Chauvel[2], A. Chambaz[1]
[1] Modal'X, Université Paris Ouest Nanterre
[2] Laboratoire d'Accidentologie et de Biomécanique

July 25, 2016

Abstract

Each year, the BAAC (Bulletin d'Analyse des Accidents Corporels) data set includes traffic accidents on French public roads involving one or two light vehicles and injuring at least one of the passengers. Each light vehicle is associated with its "generational class" (GC), which gives a raw description of the vehicle. Two light vehicles with two different GCs do not necessarily offer the same level of passive safety to their passengers in different contexts of traffic accident. 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.

In [8], we elaborated an algorithm for the contextual ranking of GCs. In the present study, our objective is to develop an algorithm for the global (as opposed to contextual) ranking of GCs. Like in [8], we rely on "scoring": we look for a score function that associates any GC with a real number; the smaller is this number, the safer is the GC across all contexts of accident. Causal arguments help to formalize our objective in statistical terms. We rely on cross-validation to select the best score function among a collection of candidate score functions which are built based on the algorithm for the contextual ranking of GCs and a collection of working models. We implement the resulting algorithm, apply it, and show some results.

Keywords: car safety, causal analysis, cross-validation, scoring.

1 Introduction

The title is a reference to that of a first article that we have devoted to the ranking by passive safety of generational classes (GCs) of light vehicles in any context of traffic accident [8]. Our objective here is to integrate out the context of traffic accident from the ranking, therefore yielding a global (as opposed to local/contextual) ranking by passive safety of GCs of light vehicles.

Our previous study relied on "scoring": we looked for a score function that associates any context of traffic 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 was learned from real-life traffic accidents data by cross-validation, under the form of an optimal convex combination of score functions produced by a library of ranking algorithms by scoring. In this light, we now look for a score function that associates any GC with a real number in such a way that the smaller is this number, the safer is the GC across all contexts (or rather, across a distribution of contexts).

1.1 Background

In 2016, preventing *traffic accidents* (we will simply write *accident*s in the rest of the article) and limiting their often tragic aftermaths is a worldwide priority for all the actors involved in road safety. Enhancing road safety notably requires to apprehend vehicles from the angle of accidentology, the study and analysis of the causes and effects of accidents. Considerable efforts are made when designing new models of vehicles based, notably, on the analysis of real-life accidents to evaluate the extent to which new systems provide better safety.

We focus on the passive safety, as opposed to the 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) whereas active safety refers to the prevention of accidents (by means of driving assistance systems). When not stated otherwise, *safety* will now stand for *passive safety*.

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 [5]. Typically, safety ratings are either predictive or retrospective. Predictive safety ratings assess the safety of vehicles based on crash tests [6]. Retrospective safety ratings assess the safety of vehicles based on real-life accidents from police and insurance claim data. In Europe, the two major predictive and retrospective safety ratings are, respectively, the European New Car Assessment Programme and Folksam Car Safety Rating System. It has been shown that there is a strong correlation between the two [7, and references therein].

The safety rating for GC of light vehicles (we will simply write *vehicles* in the rest of the article) that we elaborated in [8] is both retrospective and predictive. Retrospective because its construction exploits real-life accidents data. Predictive, in the usual statistical sense: it is possible to extrapolate a safety ranking for a synthetic GC of vehicles even in the absence of data relative to it. Moreover, it is also contextual: the safety ranking is conditioned on the occurrence of an accident in any given context. As we explained, our objective is to average out the context from the the safety ranking in order to provide a global ranking by (passive) safety.

1.2 BAAC* data set

As in [8], we use the French national file of personal accidents called BAAC data set. BAAC is an acronym for a French expression translating 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. 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. An example of blank BAAC form is given in [8, Figure 4].

It is suggested in the previous paragraph 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]. See [8, Section 1.2] and these references for details. 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.

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

1.3 Methodology

We use three main ingredients to build the algorithm for the global ranking of GCs by safety from the algorithm for the local/contextual ranking of GCs elaborated in [8]. First, a causal model helps to formalize our statistical objective. Second, working models are used to infer candidate score functions. Third, the best among the candidate score functions is identified by cross-validation.

1.4 Organization of the article

Section 2 briefly presents the data and their distribution. Section 3 describes the statistical objective of this study. It lists four main challenges that we face and how we take them up. Section 4 summarizes the specifics of the implementation, illustrates the resulting algorithm to rank GCs globally by passive safety, and validates its use. Section 5 concludes the article with a discussion.

2 Data and their distribution

2.1 Simplification

We refer the reader to [8, Section 2] for a detailed modelling of the BAAC* data and their distribution. The modelling is not trivial because a generic accident contributes a complex data-structure $\boldsymbol{O}$ consisting of one or two (depending on the number of vehicles involved in the accident) clusters $\boldsymbol{O}_k$ of dependent, individual, smaller data-structures O_{kj} describing the accident from the point of view of each occupant $1 \leq j \leq J_k$ of each vehicle k. Moreover, we have to deal with the potential missingness of the components of $\boldsymbol{O}_k$ describing the GC of vehicle k.

In these sections, we state, comment on and justify four assumptions that allow us to make inference. Lemma 1 in [8, Section 4] shows how to carry out estimation as if we observed the individual, smaller data-structures drawn, independently, from the distribution of interest (that of the accident from the point of view of any of its actors). To alleviate the present exposition, we will proceed as if we randomly selected one single individual, smaller data-structure O_{kj} from every complex data-structure $\boldsymbol{O}$. However, in our application, we will exploit [8, Lemma 1, Section 4] to use *all* observations.

2.2 Modelling

We observe a data set of n data-structures $O_1, \ldots, O_n$ independently drawn from the distribution of interest P. We denote P_n the corresponding empirical measure. Set $1 \leq i \leq n$. The data-structure O_i decomposes as $O_i = (W_i, X_i, Z_i)$ where Z_i indicates the severity of injuries incurred

by the corresponding occupant of the vehicle, X_i is a raw description of the vehicle, and W_i summarizes the context of accident.

Specifically, Z_i 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 Z_i equals zero if the injury is light (occupant hospitalized for less than 24 hours) or the occupant is unharmed. The GC X_i 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". The context W_i consists of 27 variables. We list them in [8, Section A.1], regrouped in six themes: general, when and where, what roadway, what collision, which driver, which occupant.

3 Statistical challenge

Our main objective is to learn to rank GCs by safety across all contexts of accident. This statement is better explained by resorting to causal arguments.

3.1 Causal argumentation

Expressing the objective in a counterfactual world. Let $\mathbb{O} = (W, X, (Z_x)_{x\in\mathcal{X}})$ be a full, counterfactual data-structure describing all the counterfactual outcomes Z_x ($x \in \mathcal{X}$) of an accident involving GC x in context W, and the GC X which is actually involved in the accident. The observed (as opposed to counterfactual) data-structure $O = (W, X, Z = Z_X)$ is the summary measure derived from $\mathbb{O}$ by removing the counterfactual outcomes Z_x for all $x \neq X$. The distribution P of O is a marginal joint distribution of the counterfactual distribution $\mathbb{P}$ of $\mathbb{O}$.

Let $\mathbb{P}^{\otimes 2}$ be the joint distribution of $(\mathbb{O}, \mathbb{O}')$ drawn in two steps by *(i)* sampling a context of accident W_1 from the marginal distribution of W under $\mathbb{P}$ then *(ii)* sampling independently $\mathbb{O}$ and $\mathbb{O}' = (W', X', (Z'_x)_{x\in\mathcal{X}})$ from the distribution derived from $\mathbb{P}$ by conditioning on $W = W' = W_1$.

If, contrary to facts, we had access to counterfactual observations drawn from $\mathbb{P}^{\otimes 2}$, then our objective could be expressed as follows:

(i) learn a mapping $\rho : \mathcal{X}^2 \to \{-1, 0, 1\}$ with $\rho(x, x') = 0$ if and only if (iff) $x = x'$ and such that the probabilities $\mathbb{P}^{\otimes 2}((Z_x - Z'_{x'})\rho(x, x') > 0)$ be as small as possible for all $(x, x') \in \mathcal{X}^2$;
(ii) declare that, for any two $x, x' \in \mathcal{X}$ with $x \neq x'$, GC x is safer than GC x' (across all contexts of accident) iff $\rho(x, x') = 1$.

It is how we intend to use ρ (see *(ii)*) that justifies our wish to minimize the probabilities $\mathbb{P}^{\otimes 2}((Z_x - Z'_{x'})\rho(x, x') > 0)$ (see *(i)*). Indeed, for any $(x, x') \in \mathcal{X}^2$ with $x \neq x'$, $Z_x, Z'_{x'} \in \{0, 1\}$ implies that

$$\begin{aligned}\mathbb{P}^{\otimes 2}((Z_x - Z'_{x'})\rho(x, x') > 0) \\ = E_{\mathbb{P}^{\otimes 2}}\left[\mathbb{P}^{\otimes 2}(Z_x = 1, Z'_{x'} = 0, \text{“}x \text{ declared safer than } x'\text{”}\middle|W)\right] \\ + E_{\mathbb{P}^{\otimes 2}}\left[\mathbb{P}^{\otimes 2}(Z_x = 0, Z'_{x'} = 1, \text{“}x' \text{ declared safer than } x\text{”}\middle|W)\right]. \quad (1)\end{aligned}$$

The above RHS expression allows to interpret the LHS one as a ranking error (obtained by averaging out the context, see the outer expectations) because, whichever is the context of accident W, "$Z_x = 1$ and $Z'_{x'} = 0$" means that GC x' proved safer than GC x in context W.

Reaching the objective in the counterfactual world. It is easy to derive the optimal ρ_0 from (1). Let us introduce the conditional expectation $\mathbb{Q}$ characterized by

$$\mathbb{Q}(x, W) = E_{\mathbb{P}}(Z_x|W) \quad (\text{all } x \in \mathcal{X}). \tag{2}$$

Note that the tower rule yields

$$E_{\mathbb{P}}[\mathbb{Q}(x, W)] = E_{\mathbb{P}}(Z_x) \quad (\text{all } x \in \mathcal{X}).$$

By conditional independence of Z_x and $Z'_{x'}$ given W, (1) yields

$$\begin{aligned}
&\mathbb{P}^{\otimes 2}((Z_x - Z'_{x'})\rho(x, x') > 0) \\
&\qquad = \mathbf{1}\{\rho(x, x') = 1\} E_{\mathbb{P}}[\mathbb{Q}(x, W)] \left(1 - E_{\mathbb{P}}[\mathbb{Q}(x', W)]\right) \\
&\qquad\quad + \mathbf{1}\{\rho(x, x') = -1\} \left(1 - E_{\mathbb{P}}[\mathbb{Q}(x, W)]\right) E_{\mathbb{P}}[\mathbb{Q}(x', W)] \\
&\qquad = \mathbf{1}\{\rho(x, x') = 1\} E_{\mathbb{P}}(Z_x) \left(1 - E_{\mathbb{P}}(Z_{x'})\right) \\
&\qquad\quad + \mathbf{1}\{\rho(x, x') = -1\} \left(1 - E_{\mathbb{P}}(Z_x)\right) E_{\mathbb{P}}(Z_{x'})
\end{aligned}$$

Therefore, $\mathbb{P}^{\otimes 2}((Z_x - Z'_{x'})\rho(x, x') > 0)$ is minimized iff $\rho(x, x') = \rho_0(x, x')$ with

$$\rho_0(x, x') = 2\mathbf{1}\{E_{\mathbb{P}}(Z_x) < E_{\mathbb{P}}(Z_{x'})\} - 1.$$

In words, declare that GC x is safer than GC x' if $E_{\mathbb{P}}(Z_x) < E_{\mathbb{P}}(Z_{x'})$ and that GC x' is safer than GC x if $E_{\mathbb{P}}(Z_x) \geq E_{\mathbb{P}}(Z_{x'})$ (safer across all contexts of accident).

The optimal ρ_0 is a "scoring ranking rule" in the sense that ρ_0 is fully known when the mapping $x \mapsto E_{\mathbb{P}}(Z_x)$ from $\mathcal{X}$ to $[0, 1]$ is known. In particular, the definition of ρ_0 depends on $\mathbb{P}$ and not on $\mathbb{P}^{\otimes 2}$. Consequently, the estimation of ρ_0 could be addressed through the estimation of $E_{\mathbb{P}}(Z_x)$ (every $x \in \mathcal{X}$) which could be carried out based, for instance, on the loss function $\mathbb{L}^1_x$ given by

$$-\mathbb{L}^1_x(f, \mathbb{O}) = Z_x \log\left(f(x)\right) + (1 - Z_x) \log\left(1 - f(x)\right)$$

(where $\mathbb{O}$ is drawn from $\mathbb{P}$ and f ranges over a class of functions mapping $\mathcal{X}$ to $]0, 1[$). The performance of the resulting estimator of ρ_0 could be expressed in terms of a cross-validated empirical aggregated risk based on $\mathbb{L}^1_x$ (all $x \in \mathcal{X}$). However, one could argue that a tailored measure of performance should take the form of a cross-validated empirical aggregated risk based on $\mathbb{L}^2_{x,x'}$ (all $x, x' \in \mathcal{X}$, $x \neq x'$) given by

$$\mathbb{L}^2_{x,x'}(\rho, \mathbb{O}, \mathbb{O}') = \mathbf{1}\{(Z_x - Z'_{x'})\rho(x, x') > 0\})$$

(where $(\mathbb{O}, \mathbb{O}')$ is drawn from $\mathbb{P}^{\otimes 2}$ and ρ ranges over a class of functions mapping $\mathcal{X}^2$ to $\{-1, 0, 1\}$).

Reaching the objective in the real world under causal assumptions. Under so called causal assumptions, it is possible to estimate $\mathbb{Q}(x, W)$ and $E_{\mathbb{P}}[\mathbb{Q}(x, W)] = E_{\mathbb{P}}(Z_x)$ (each $x \in \mathcal{X}$) from "real world observations" such as $O = (W, X, Z = Z_X)$ (as opposed to counterfactual data-structures $\mathbb{O}$) drawn from the "real world distribution" P (as opposed to the counterfactual distribution $\mathbb{P}$). Namely, let the randomization assumption postulate that X is conditionally independent from $(Z_x)_{x \in \mathcal{X}}$ given W ($\mathbb{P}$-almost surely) and let the positivity assumption postulate that the conditional distribution of X given W puts positive mass almost everywhere ($\mathbb{P}$ or P-almost surely). Under these causal assumptions, for each $x \in \mathcal{X}$,

$$\mathbb{Q}(x, W) = E_{\mathbb{P}}(Z_x|W) \overset{(a)}{=} E_{\mathbb{P}}(Z_x|X = x, W) \overset{(b)}{=} E_P(Z|X = x, W) \tag{3}$$

where *(a)* follows from the randomization and positivity assumptions, and *(b)* follows from the equality $Z = Z_X$ (sometimes called the consistency assumption) and the definition of P as the

marginal joint distribution of the summary measure O derived from $\mathbb{O}$ drawn from $\mathbb{P}$. Moreover, (3) straightforwardly implies that

$$E_{\mathbb{P}}[\mathbb{Q}(x,W)] = E_{\mathbb{P}}(Z_x) = E_P\left[E_P(Z|X=x,W)\right]. \tag{4}$$

Thus, the estimation of ρ_0 can be addressed in two steps through the estimation of

$$Q(X,W) = E_P(Z|X,W) \tag{5}$$

and

$$s_0(x) = E_P[Q(x,W)] \tag{6}$$

for all $x \in \mathcal{X}$. The estimation of Q can be carried out based, for instance, on the loss function L^1 given by

$$-L^1(f,O) = Z\log(f(X,W)) + (1-Z)\log(1-f(X,W)) \tag{7}$$

(where O is drawn from P and f ranges over a class of functions mapping $\mathcal{X}\times\mathcal{W}$ to $]0,1[$). Given an estimator $\widetilde{Q}$ of Q and an empirical distribution $\widetilde{P}_W = \widetilde{n}^{-1}\sum_{i=1}^{\widetilde{n}} \mathrm{Dirac}(\widetilde{W}_i)$,

$$\widetilde{s}(x) = E_{\widetilde{P}_W}[\widetilde{Q}(x,W)] = \frac{1}{\widetilde{n}}\sum_{i=1}^{\widetilde{n}} \widetilde{Q}(x,\widetilde{W}_i) \tag{8}$$

estimates $s_0(x)$ for every $x \in \mathcal{X}$. The estimator $\widetilde{s}$ yields the empirical scoring ranking rule $\widetilde{\rho}$ given by

$$\widetilde{\rho}(x,x') = 2\mathbf{1}\{\widetilde{s}(x) < \widetilde{s}(x')\} - 1 \quad (\text{all } x,x' \in \mathcal{X}, x \neq x').$$

In the counterfactual world, the performance of $\widetilde{\rho}$ could be expressed in terms of a cross-validated empirical aggregated risk based either on $\mathbb{L}^1_x$ (all $x \in \mathcal{X}$) or on $\mathbb{L}^2_{x,x'}$ (all $x,x' \in \mathcal{X}$, $x \neq x'$). In the real world, however, only one of these options can be considered. Indeed, for every $f : \mathcal{X} \to]0,1[$ and $x \in \mathcal{X}$, reasoning as in (3) implies

$$E_P\left[E_P(L^1(f,O)|X=x,W)\right] = E_{\mathbb{P}}\left[\mathbb{L}^1_x(f,\mathbb{O})\right].$$

On the contrary, there is no such equality relating $E_{\mathbb{P}^{\otimes 2}}\left[\mathbb{L}^2_{x,x'}(f,\mathbb{O},\mathbb{O}')\right]$ to an expectation involving P. This is because for all observed contexts of accident we never observe two (conditionally) independent accidents taking place in this context, just one.

3.2 Statistical roadmap

Statistical objective (in the real world). The causal argumentation developed in Section 3.1 has given rise to a sound statistical problem. The problem is freed from the causal modeling. By this, we mean that it makes fully sense and can be addressed in the real world without further reference to the counterfactual world described in the causal model as an extension to the real world. If, eventually, one wished to give a causal interpretation to the solution of the statistical problem, then one could rely on the causal assumptions (some of them untestable from real data) proposed in Section 3.1.

Let us summarize what is the statistical problem. As stated in Section 2.2, we observe $O_1,\ldots,O_i=(W_i,X_i,Z_i),\ldots,O_n$ independently drawn from P. We wish to estimate $s_0 : \mathcal{X} \to [0,1]$ given by $s_0(x) = E_P[Q(x,W)]$ (6) where $Q : \mathcal{X}\times\mathcal{W} \to [0,1]$ is characterized by $Q(X,W) = E_P(Z|X,W)$ (5). The statistical performance of an estimator $s_n : \mathcal{X} \to [0,1]$ of s_0 will be evaluated based on (but not limited to) the aggregated risk

$$\int_{\mathcal{X}} E_P[E_P(L^1(s_n,O)|X=x,W)]d\mu(x) \tag{9}$$

(for a user-supplied measure μ on $\mathcal{X}$; we omit the measurability issues) where the loss function L^1 is given in (7). Such an evaluation is not tailored to the fact that we are secondarily interested in s_0 and primarily interested in the scoring ranking rule $(x, x') \mapsto 2\mathbf{1}\{s_0(x) < s_0(x')\} - 1$ yielded by s_0, for the sake of ranking GCs of vehicles. However, the nature of our data sets does not allow a tailored evaluation, unfortunately.

Aggregated loss and risk. We now elaborate the aggregated loss $\ell^1_{Q,\mu}(s_n, W)$ and related risk $\mathcal{R}_{Q,\mu}(P)(s_n) = E_P[\ell^1_{Q,\mu}(s_n, W)]$ that we will use to evaluate the statistical performance of s_n. To do this, let us analyze the integrand in (9). For every $x \in \mathcal{X}$, it holds that

$$E_P[E_P(L^1(s_n, O)|X = x, W)] = E_P[L^1_{Q,x}(s_n, W)]$$

where, for any $f : \mathcal{X} \to]0, 1[$,

$$-L^1_{Q,x}(f, W) = Q(x, W) \log(f(x)) + (1 - Q(x, W)) \log(1 - f(x)).$$

To evaluate the performance of f across $\mathcal{X}$ (as an estimator of s_0), we aggregate the loss functions $L^1_{Q,x}$ $(x \in \mathcal{X})$.

Denote

$$\Lambda(p, q) = p \log(p/q) + (1 - p) \log((1 - p)/(1 - q))$$

the Kullback-Leibler divergence between the Bernoulli laws with parameters $p, q \in]0, 1[^2$ and let μ be a probability measure on $\mathcal{X}$. We propose the aggregated loss function $\ell^1_{Q,\mu}$ given (omitting the measurability issues) by

$$\begin{aligned} \ell^1_{Q,\mu}(f, W) &= \int_{\mathcal{X}} \Big[L^1_{Q,x}(f, W) + Q(x, W) \log(Q(x, W)) \\ &\qquad + (1 - Q(x, W)) \log(1 - Q(x, W)) \Big] d\mu(x) \\ &= \int_{\mathcal{X}} \Lambda(Q(x, W), f(x)) d\mu(x). \end{aligned} \tag{10}$$

Note that we actually aggregate translated versions of the loss functions $L^1_{Q,x}$ to ensure non-negativeness of the integrand in (10). In particular, Fubini's theorem thus yields that the resulting aggregated risk of s_n:

$$\mathcal{R}_{Q,\mu}(P)(s_n) = E_P[\ell^1_{Q,\mu}(s_n, W)] \tag{11}$$

equals (9) up to the term

$$\int_{\mathcal{X}} E_P\Big[Q(x, W) \log(Q(x, W)) + (1 - Q(x, W)) \log(1 - Q(x, W)) \Big] d\mu(x)$$

which does not depend on s_n. This additional term justifies why we wrote "based on (but not limited to)" before (9).

Implementation. The implementation poses *four* challenges. The three first challenges are that:

1. we do not know Q;
2. we must provide a probability measure μ on $\mathcal{X}$;
3. we must find a practical way to explore the set of functions from $\mathcal{X}$ to $[0, 1]$.

The fourth challenge will arise once we have solved the three first ones. We propose the following practical solutions:

1. We estimate Q with $\widetilde{Q}$ based on an independent data set. Specifically, $\widetilde{Q}$ is the estimator that we constructed in [8, Section 6.2] by super learning [10, 9] with 49 different algorithms. In addition, denoting $\widetilde{P}_W$ the empirical distribution of W in the data set used to build $\widetilde{Q}$, we also define $\widetilde{s}$ as in (8) for future use.
2. The probability measure μ on $\mathcal{X}$ that we provide is the empirical distribution of X in the data set used to construct $\widetilde{Q}$. This simple choice guarantees that μ puts weight on meaningful GCs, whereas the construction "by hand" of a synthetic μ would be prone to putting weight on unrealistic GCs. The empirical distribution of X yields other distributions of interest by conditioning on the values of one of the seven components of X. For instance, the empirical distributions of X conditional on size-class are five other meaningful probability measures on $\mathcal{X}$.

From now on, $\widetilde{Q}$, $\widetilde{s}$ and $\widetilde{\mu}$ are treated as fixed. We acknowledge that this may result in slightly over-optimistic statements regarding our statistical performances.

3. We also provide low-dimensional, parametric working models $\mathcal{F}_1, \ldots, \mathcal{F}_K$ where each $\mathcal{F}_k = \{f_{k,\theta} : \theta \in \Theta_k\}$ is a set of functions from $\mathcal{X}$ to $[0,1]$. We make sure that each $\mathcal{F}_k$ is identifiable: $f_{k,\theta} = f_{k,\theta'}$ implies $\theta = \theta'$. Moreover, we assume that

$$\theta \mapsto \mathcal{R}_{\widetilde{Q},\widetilde{\mu}}(P)(f_{k,\theta}) = E_P[\ell^1_{\widetilde{Q},\widetilde{\mu}}(f_{k,\theta}, W)] \tag{12}$$

admits a unique minimizer $\widehat{\theta}_k(P)$ over each $\mathcal{F}_k$.

For each $1 \leq k \leq K$, let us assume that there exists a unique minimizer $\widehat{\theta}_k(P_n)$ over $\mathcal{F}_k$ of the empirical counterpart to the aggregated risk (12):

$$\theta \mapsto \mathcal{R}_{\widetilde{Q},\widetilde{\mu}}(P_n)(f_{k,\theta}) = E_{P_n}[\ell^1_{\widetilde{Q},\widetilde{\mu}}(f_{k,\theta}, W)] = \frac{1}{n}\sum_{i=1}^{n} \ell^1_{\widetilde{Q},\widetilde{\mu}}(f_{k,\theta}, W_i).$$

The corresponding element of $\mathcal{F}_k$, $f_{k,\widehat{\theta}_k(P_n)}$, estimates s_0 and yields the empirical scoring ranking rule $(x,x') \mapsto 2\mathbf{1}\{f_{k,\widehat{\theta}_k(P_n)}(x) < f_{k,\widehat{\theta}_k(P_n)}(x')\} - 1$. We can now state the fourth challenge:

4. we must identify and select the best working model of the collection introduced to solve challenge 3 above.

The identification and selection must use the aggregated risk $\mathcal{R}_{\widetilde{Q},\widetilde{\mu}}$ but cannot be based on comparisons of $\mathcal{R}_{\widetilde{Q},\widetilde{\mu}}(P_n)(f_{k,\widehat{\theta}_k(P_n)})$, $1 \leq k \leq K$, because they do not account for the fact that bigger working models will often yield smaller, minimal aggregated risks at the cost of more variability. We propose to rely on cross-validation.

4. Let $B_n \in \{0,1\}^n$ be a random vector indicating splits into a training sample, $\{O_i : 1 \leq i \leq n, B_n(i) = 0\}$, and a validation sample $\{O_i : 1 \leq i \leq n, B_n(i) = 1\}$. The vector B_n is drawn independently of $O_1, \ldots, O_n$ from a distribution such that $n^{-1}\sum_{i=1}^n B_n(i) = p$, for $p \in]0,1[$ a deterministic proportion. For notational simplicity, we choose p so that np be an integer. Then, given B_n, $P_{n,B_n,0} = (n(1-p))^{-1}\sum_{i=1}^n \mathbf{1}\{B_n(i) = 0\}\text{Dirac}(O_i)$ and $P_{n,B_n,1} = (np)^{-1}\sum_{i=1}^n \mathbf{1}\{B_n(i) = 1\}\text{Dirac}(O_i)$ are, respectively, the training and validation empirical measures.
   For each $1 \leq k \leq K$, the risk of $\widehat{\theta}_k(P_{n,B_n,0})$ is assessed through

$$\begin{aligned}\mathcal{R}_{\widetilde{Q},\widetilde{\mu}}(P_{n,B_n,1})\left(f_{k,\widehat{\theta}_k(P_{n,B_n,0})}\right) &= \frac{1}{np}\sum_{1\leq i\leq n} \mathbf{1}\{B_n(i)=1\}\ell^1_{\widetilde{Q},\widetilde{\mu}}\left(f_{k,\widehat{\theta}_k(P_{n,B_n,0})}, W_i\right) \\ &= P_{n,B_n,1}\ell^1_{\widetilde{Q},\widetilde{\mu}}\left(f_{k,\widehat{\theta}_k(P_{n,B_n,0})}, \cdot\right).\end{aligned}$$

   This results in a cross-validated aggregated risk of working model $\mathcal{F}_k$ defined as

$$E_{B_n}\left[P_{n,B_n,1}\ell^1_{\widetilde{Q},\widetilde{\mu}}\left(f_{k,\widehat{\theta}_k(P_{n,B_n,0})}, \cdot\right)\right]. \tag{13}$$

The best working model among $\mathcal{F}_1, \ldots, \mathcal{F}_K$ is the one indexed by the minimizer of these criteria,

$$K_n = \underset{1 \leq k \leq K}{\arg\min} E_{B_n} \left[P_{n,B_n,1} \ell^1_{\widetilde{Q},\widetilde{\mu}} \left(f_{k,\widehat{\theta}_k(P_{n,B_n,0})}, \cdot \right) \right].$$

It is because we resort to cross-validation that we must treat $\widetilde{Q}$ as fixed. Indeed, the computational burden of the estimation of Q_0 with $\widetilde{Q}$ as we carried it out in [8, Section 6.2] is so considerable that it cannot be iterated across the successive folds.

Finally, we estimate s_0 with the score function

$$S_n = f_{K_n,\widehat{\theta}_{K_n}(P_n)} \tag{14}$$

which is obtained by training the best working model on the whole data set.

4 Application

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. 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. The 2013 BAAC* data set consists of 15,004 reports of accidents. There are 6,718 one-vehicle and 8,286 two-vehicle accidents reported in it. The 2014 BAAC* data set consists of 15,323 reports of accidents. There are 6,771 one-vehicle and 8,552 two-vehicle accidents reported in it.

We exploit the 2011 BAAC* data set for two purposes. First, we build $\widetilde{Q}$ by super learning [8, Section 6.2] using all observations. Second, we arbitrarily select 1,000 different GCs among the GCs that appear in the data set and define $\widetilde{\mu}$ as the probability measure putting mass 10^{-3} on each of the selected GC.

Moreover, we arbitrarily decompose the 2012 BAAC* data set in two disjoint subsets, each consisting of 5,000 reports of accidents. One is used to build $\widetilde{s}$ from $\widetilde{Q}$ as in (8). The other one yields the empirical measure P_n which is referred to in Section 3.2. It is thus used to identify the best working model and to train it as in (14).

Finally, the 2013 and 2014 BAAC* data sets are used in Section 4.3 to evaluate the global ranking yielded by S_n.

4.1 Identifying by cross-validation the best among 25 working models

We elaborate $K = 25$ different working models.

The first one is the singleton $\{\widetilde{s}\}$ (see solution 1 to challenge 1 at the end of Section 3.2). The second one is a logistic model using only the categorical components of x. The third one is a logistic model using only the numerical components of x. The fourth one is a logistic model using all the components of x. The fifth one is a logistic model using all the components of x and the squares of the numerical components of x. The sixth one is a logistic model using all the components of x and the squares and cubes of the numerical components of x. The next seven working models are logistic models using all but one of the components of x. The twelve remaining working models are obtained by using $\widetilde{s}(x)$ as an additional predictive variable in the twelve previous working models.

We identify the best among the $K = 25$ working models as described in challenge 4 in Section 3.2. The distribution of B_n is uniform on the set $\{b_1, \ldots, b_{10}\}$ where $b_j \in \mathbb{R}^n$ is given by

$b_j(i) = 1$ iff $n(j-1)/10 + 1 \leq i \leq nj/10$ for $j = 1, \ldots, 10$. We compute the values of the cross-validated risks (13) for all working models. The working model with the largest cross-validated risk is the singleton $\{\widetilde{s}\}$. Each model using $\widetilde{s}$ as a predictor has a smaller cross-validated risk than its counterpart which does not use $\widetilde{s}$ as a predictor. The best working model, *i.e.*, the working model whose cross-validated risk is the smallest, is the the logistic model that uses all components of x and the squares of the numerical components of x in addition to $\widetilde{s}(x)$. So we select and train it on the whole data set, yielding the estimator S_n of s_0, see (14).

4.2 Illustration

We arbitrarily characterize eight GCs to rank by global passive safety. The GCs are partially presented in columns 2-4 in Table 1.

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.

We observe that, within each size class, the scores decrease as the date of design increases: S1 $\prec$ S2 $\prec$ S3, L1 $\prec$ L2 $\prec$ L3, M1 $\prec$ M2. In words, within each size class, the global passive safety is improved from one generation to the next (the word "generation" refers to the date of design). The same holds for the date of entry into service: within each size class, the scores decrease as the date of entry into service increases. This is in agreement with the expert assessment.

Comparisons can also be made across size classes, by ranking the scores from the largest to the smallest. This yields the following global ranking by increasing passive safety: M1 $\prec$ S1 $\prec$ L1 $\prec$ S2 $\prec$ M2 $\prec$ S3 $\prec$ L2 $\prec$ L3. Commenting on this global ranking is uneasy. Actually, two experts may very well expect diverging global rankings since it is difficult to compare GCs of different size class, notably because they are not used similarly.

However, it is easy to give one explanation to one feature of this ranking. If we associate the date of its design with every GC (between parentheses), then the global ranking writes: M1 (1994) $\prec$ S1 (1998) $\prec$ L1 (2001) $\prec$ S2 (2005) $\prec$ M2 (2002) $\prec$ S3 (2011) $\prec$ L2 (2007) $\prec$ L3 (2013). In particular, the sequence of dates of design is not increasing. One may wonder naively why would M2 designed in 2002 be globally safer than S2 designed in 2005, and why would L2 designed in 2007 be globally safer than S3 designed in 2011? For experts, this does not come as a surprise. An undisputable element of explanation is that, in general, GCs of the smallest size class (small family car) are not as well equipped as GCs of larger size classes.

Finally, one should interpret this rankings cautiously. In particular, it is not possible to disentangle completely the effects of better industrial design, wear due to time into service, and more stringent safety regulations. Consider for instance a GC x designed *and* put into service in 1994 like our synthetic GC M1 in Table 1. Its score $S_n(x)$ quantifies its global safety with respect to a distribution of context derived from the 2012 BAAC* data set. The large value of $S_n(x)$ can be due to the facts that *(i)* x is an old-designed GC in 2012, *(ii)* x is a worn GC in 2012, and *(iii)* the distribution of context derived from the 2012 BAAC* data set differs drastically from the distribution of context we would have derived from, say, a 1995 BAAC* data set. How *(i)*, *(ii)* and *(iii)* contribute to the making of $S_n(x)$ cannot be determined, if at all, without a more systematic analysis.

In a preliminary effort, we select 28 emblematic GCs. For each GC x among them, and for any date of entry into service δ, we denote $x(\delta)$ the GC obtained by substituting δ for the original date of entry into service of x. We observe how $\delta \mapsto S_n(x(\delta))$ behaves, where δ ranges over a set of meaningful dates of entry into service. Systematically, the above mapping is decreasing.

	generational class (GC)			
GC code, x	date of design	date of entry into service	size class	score, $S_n(x)$
S1	1998	2001	small family car	0.327
S2	2005	2007	small family car	0.304
S3	2011	2011	small family car	0.298
L1	2001	2003	large family car	0.311
L2	2007	2008	large family car	0.294
L3	2013	2014	large family car	0.288
M1	1994	1994	minivan	0.339
M2	2002	2002	minivan	0.302

Table 1: 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. In the last column, we report the scores $S_n(x)$ of each of these GCs x.

This finding is in agreement with the experts' expectations: all other things being fixed, a new GC is safer than a used GC, where "new" and "used" refer to the use and wear of the GC to which the date of entry into service is a proxy.

4.3 Evaluation

In this section, we evaluate the global ranking yielded by S_n. For this, we first correlate the scores $S_n(x_j)$ derived from the BAAC* data set with scores $S_{\mathrm{NCAP}}(x_j)$ derived from consumerist studies for a collection $\{x_1, \ldots, x_J\} \subset \mathcal{X}$ of $J = 155$ GCs. Second, we compare the empirical distributions of $\{S_n(X_i) : i \in \mathcal{S}_1\}$ and $\{S_n(X_i) : i \in \mathcal{S}_2\}$ with $(\mathcal{S}_1, \mathcal{S}_2)$ ranging over a collection of couples of disjoint subsets of $\{1, \ldots, n\}$. See below for details.

Correlation with European New Car Assessment Programme consumerist ratings. The European New Car Assessment Programme (Euro NCAP) consumerist association rates vehicles in terms of a five-star safety rating to help consumers identify the safest choice for their needs. The safety rating is determined from a series of vehicle tests, designed and carried out by Euro NCAP. They represent, in a simplified way, important real-life accident scenarios that could result in injured or killed car occupants or other road users.

The Euro NCAP rating methodology has been evolving through the years, and we refer the interested reader to the association's website for a detailed description (`http://www.euroncap.com/en/for-engineers/protocols/`). We focus on scores derived from frontal-impact and side-impact crash tests that quantify the protection of the driver and front passenger. We identify three major periods during which the corresponding methodology did not change significantly: 1996–2000, 2001–2008, 2009–2014. During the first period, only one side-impact test (side-impact with a mobile deformable barrier) was conducted. It yielded a side-impact grade lying in $[0, 16]$. During the second period, an additional side-impact test (side-impact with a pole) was optionally conducted. Either way, a single grade summarized the test(s), with values in $[0, 16]$ if one test was conducted and in $[0, 18]$ otherwise. During the third period, both side-impact tests were systematically conducted and yielded two grades lying in $[0, 8]$. One single frontal-impact test was conducted during all periods, yielding a grade lying in $[0, 16]$. Larger grades mean better protection.

Based on these grades, we elaborate a score by adding all (two or three) grades and substracting the result to 100, so that smaller scores mean better protection. We analyze the Euro

NCAP data set and manage to compute a collection $\{S_{\mathrm{NCAP}}(x_j) : 1 \leq j \leq J\}$ of so called Euro NCAP scores for $J = 155$ different GCs $x_1, \ldots, x_J \in \mathcal{X}$. The analysis is tedious because it cannot be automated. Each vehicle in the Euro NCAP data set for which we were able to determine its GC is included.

Comparisons between Euro NCAP test results and real-world crash data have already been done [7, and references therein]. Here, we evaluate how correlated are our scores $S_n(x_j)$ with $S_{\mathrm{NCAP}}(x_j)$ for $1 \leq j \leq J = 155$, see Figure 1 for a visual representation. The three plots correspond to the three major periods 1996–2000 (38 GCs indexed by $j \in \mathcal{J}_1$), 2001–2008 (70 GCs indexed by $j \in \mathcal{J}_2$) and 2009–2014 (47 GCs indexed by $j \in \mathcal{J}_3$). Visually, it seems that the cloud of points $\{(S_n(x_j), S_{\mathrm{NCAP}}(x_j)) : j \in \mathcal{J}_k\}$ shifts down and to the left as k goes from 1 to 3. Moreover, it seems that the y-range of the cloud tends to decrease.

Kruskall-Wallis and one-sided Wilcoxon non-parametric tests confirm all but one of the visual findings regarding how the clouds of points shift. Indeed, the one-sided Wilcoxon test comparing the distributions of $\{(S_n(x_j), S_{\mathrm{NCAP}}(x_j)) : j \in \mathcal{J}_k\}$ for $k = 2, 3$ does not support the fact that the former is stochastically smaller than the latter.

For each period $k = 1, 2, 3$, we compute the ratio of the standard deviation of $\{S_n(x_j) : j \in \mathcal{J}_k\}$ to its mean and the ratio of the standard deviation of $\{S_{\mathrm{NCAP}}(x_j) : j \in \mathcal{J}_k\}$ to its mean. We obtain: 4.17% and 6.91% (1996–2000), 4.24% and 5.44% (2001–2008), 4.02% and 3.18% (2009–2014). We note that the ratios based on S_n do not vary much across periods whereas the ratios based on S_{NCAP} decrease. This second fact shows that the variability of $\{S_{\mathrm{NCAP}}(x_j) : j \in \mathcal{J}_k\}$, contrary to that of $\{S_n(x_j) : j \in \mathcal{J}_k\}$, tends to narrow (relative to their mean) as k goes from 1 to 3. The same result holds when considering the difference of the maximum and minimum values instead of the standard deviation.

For each period $k = 1, 2, 3$, we also compute Spearman's correlation and the p-value of the test of "no correlation" against "positive correlation". Spearman's correlation is meant to assess how well the relationship between two variables can be described using a monotonic function. Therefore, it is a particularly convenient measure of association since we consider S_n and S_{NCAP} as score functions to rank GCs by safety. Thus, we interpret a large estimate of Spearman's correlation as a guarantee that, for any $x, x' \in \mathcal{X}$, if we observe $S_n(x) \leq S_n(x')$, then it is likely that we also observe $S_{\mathrm{NCAP}}(x) \leq S_{\mathrm{NCAP}}(x')$, hence x is declared safer than x' both by S_n and by S_{NCAP}. We respectively obtain: 29% and 0.0409 (1996–2000), 55% and 4×10^{-7} (2001–2008), 44% and 0.00877 (2009–2014). If the first p-value is not small enough to yield a significant result, the two others are very small and show that, during both periods 2001–2008 and 2009–2014, the S_n and S_{NCAP} scores are strongly positively correlated.

In summary, despite the fact that the definitions and derivations of S_n and S_{NCAP} hinge on very different methodologies and data, it thus appears that the two score functions are very similar for the sake of ranking by safety. We had not anticipated this result.

For years, the design teams of the major French car makers have been encouraged to evaluate *in terms of the Euro NCAP ratings* what was the impact of the evolution of the designs. Now that we have shown the strong positive correlation between a component of the Euro NCAP rating (what we call S_{NCAP}) and the score function that we have built based on real-life accidents data (what we call S_n), the design teams will be reassured that such an evaluation is meaningful in real life.

Evidence-based validation. What we call evidence-based validation consists in a three-step procedure. First, we make groups of observations relative to accidents that occurred in similar contexts. We develop what we mean by "similar contexts" in the next paragraph. If an accident involves two GCs, then one of them is arbitrarily selected and the other discarded. Second,

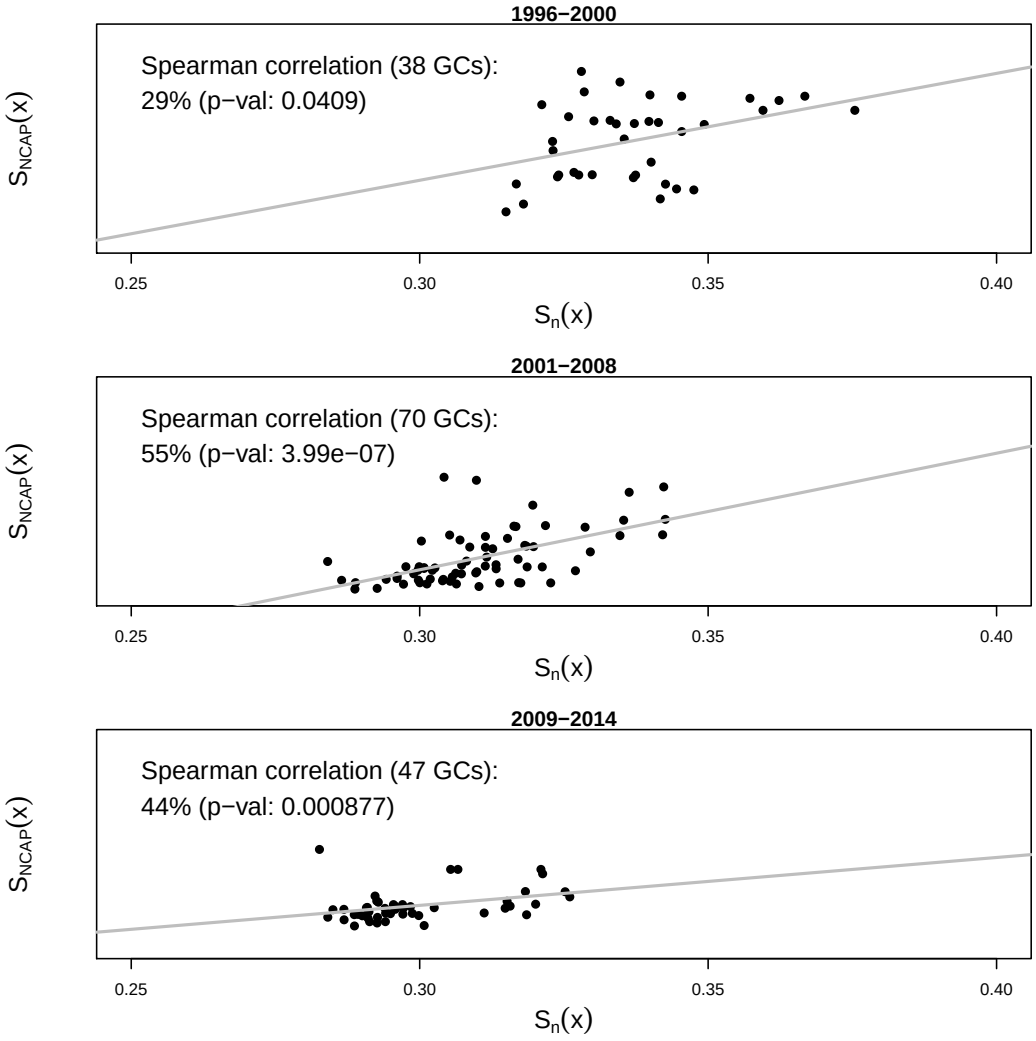

Figure 1: Comparing the scores $S_n(x)$ yielded by our method with scores $S_{\mathrm{NCAP}}(x)$ derived from Euro NCAP consumerist ratings for 155 different GCs x. Each plot corresponds to a period during which the Euro NCAP methodology did not change significantly for our purpose. We also report the estimates of the Spearman correlations and p-values of the tests of no correlation against positive correlation. The grey lines are fitted by least squares. The three plots share the same x- and y-scales.

we compute the scores of all the GCs selected during the first step (there are 1,550 of them). Third, within each group of observations, we test if the conditional distribution of score given that the accident resulted in a fatal or severe injury for the driver is stochastically smaller than the conditional distribution of score given that the accident did not result in a fatal or severe injury for the driver. In other words, within each group of observations, we test if the former distribution's CDF (cumulative distribution function) lies above the latter distribution's CDF.

We regroup the observations by *similar* contexts and not identical contexts because each context is unique in the 2013 and 2014 BAAC* data sets. In a given group $\{O_i : i \in \mathcal{I}\}$ of accidents with similar contexts (similar $W_i, i \in \mathcal{I}$), it is meaningful to regroup the accidents according to the severities of injuries. Specifically, we write $\mathcal{I} = \mathcal{I}_0 \cup \mathcal{I}_1$ with $i \in \mathcal{I}_1$ if and only if the driver of the vehicle involved (and selected, in case of a two-vehicle accident) indexed by i was fatally or severely injured.

The test described in the third step compares the CDF F_1 of $\{S_n(X_i) : i \in \mathcal{I}_1\}$ to the CDF F_0 of $\{S_n(X_i) : i \in \mathcal{I}_0\}$; we expect that the GCs which better protected their occupants (indexed by $i \in \mathcal{I}_0$) have smaller scores than the other GCs (indexed by $i \in \mathcal{I}_1$). Specifically, we carry out a Wilcoxon test of the null hypothesis "$F_0 \geq F_1$" against its alternative "$F_0 < F_1$".

We make 32 groups of interest and study them as presented above. To make the groups, we create 480 coarse contexts of reference by considering all combinations of "number of vehicles involved" (1 or 2), "season" (October to March or April to September), "weekend" (yes or no), "light condition" (daylight or dark), "urban area" (outside, small, or large), "intersection" (yes or no), "type of collision" (head-on, rear end, angle, no collision, other). For each combination, we look for the observed accidents in the 2013 and 2014 BAAC* data sets whose contexts correspond with the combination. We only keep the accidents such that the driver was aged under 20 and 60 years old, had her/his seatbelt fastened and was not driving under the influence of alcohol. If the number of such accidents such that the driver was fatally or severely injured and if the number of such accidents such that the driver was not fatally or severely injured are both larger than 10, then the set of these observed accidents qualifies as a group of interest.

We report the corresponding p-values and cardinalities of the subgroups in Table 2.[1] The smallest p-value equals 34%, so we never reject the null for its alternative. Thus, we find no evidence in the data supporting the hypothesis that, in at least one of the groups, the conditional distribution of score given that the accident resulted in a fatal or severe injury for the driver is not stochastically larger than the conditional distribution of score given that the accident did not result in a fatal or severe injury for the driver.

Even if this conclusion is not definite, we find it reassuring. Again, we had not anticipated that none of these comparisons would invalidate the stochastic domination of the conditional distribution of score given that the accident did not result in a fatal or severe injury for the driver over the conditional distribution of score given that the accident resulted in a fatal or severe injury for the driver in a coarse context.

5 Discussion

In this article, we address the global ranking of GCs by passive safety: for any two GCs $x, x' \in \mathcal{X}$, x is declared globally safer than x' if $S_n(x) \leq S_n(x')$. The score function $S_n : \mathcal{X} \to [0, 1]$ is essentially built in two steps: following [8] we first build a score function $\widetilde{Q} : \mathcal{X} \times \mathcal{W} \to [0, 1]$ for the contextual ranking of GCs (for any two couples $(x, w), (x', w') \in \mathcal{X} \times \mathcal{W}$ of GCs x, x' and

[1] When card($\mathcal{I}_1$) and card($\mathcal{I}_0$) are larger than 50, then we also carry out a one-sided Kolmogorov-Smirnov test of "$F_0 \geq F_1$" against "$F_0 < F_1$" (the computation of its p-value is based on asymptotic arguments). All these additional tests confirm the decisions of the Wilcoxon tests.

contexts of accident w, w', the combination (x, w) is declared safer than (x', w') if $\widetilde{Q}(x, w) \leq \widetilde{Q}(x', w'))$ by combining data-adaptively a library of ranking algorithms; second, using causal arguments, we derive S_n from $\widetilde{Q}$ and a collection of working models by relying on cross-validation.

We illustrate the use of S_n by comparing eight different GCs. These 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. We also observe how S_n behaves as a function of date of entry into service only for 28 emblematic GCs.

To validate the use of S_n, we propose a consumerist validation and an evidence-based validation. The consumerist validation consists in evaluating how correlated are the rankings yielded by S_n and the Euro NCAP method, which relies on frontal- and side-impact crash tests. For the evidence-based validation, we define 32 coarse contexts, or patterns, of traffic accident. For each pattern, we retrieve all accidents that occurred in contexts featuring that pattern, we compute the scores of all the involved GCs, and we test if the conditional distribution of score given that the accident resulted in a fatal or severe injury for the driver is stochastically smaller than the conditional distribution of score given that the accident did not result in a fatal or severe injury for the driver. Both validation procedures yield satisfying results.

Our approach is very flexible and calls for improvement. If, in the future, the BAAC form included additional relevant piece of 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. In this spirit, we are currently trying to enrich the definition of a generational class by relying on auxiliary data sets. The main challenge that we wish to take up next is that of the elaboration of a confidence region around S_n. Once solved, this delicate theoretical problem will have a considerable practical impact.

We acknowledge that S_n 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.

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 (2013/0333).

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] 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.

[6] 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.

[7] 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.

[8] Z. Ouni, C. Denis, C. Chauvel, and A. Chambaz. Contextual ranking by passive safety of generational classes of light vehicles. Technical report, 2015. URL `https://hal.archives-ouvertes.fr/hal-01194515`.

[9] 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.

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

card($\mathcal{I}_0$)	42	72	102	72	44	31	156	10
card($\mathcal{I}_1$)	16	20	80	12	22	21	176	10
p-value	0.34	0.41	0.45	0.46	0.63	0.64	0.67	0.69
card($\mathcal{I}_0$)	110	130	48	64	54	52	90	19
card($\mathcal{I}_1$)	150	10	26	34	20	13	66	12
p-value	0.69	0.73	0.76	0.77	0.78	0.79	0.81	0.82
card($\mathcal{I}_0$)	124	48	50	126	83	20	32	88
card($\mathcal{I}_1$)	74	44	12	102	14	24	10	74
p-value	0.84	0.85	0.89	0.91	0.94	0.94	0.96	0.96
card($\mathcal{I}_0$)	42	40	148	76	38	40	120	62
card($\mathcal{I}_1$)	14	50	138	106	10	12	56	58
p-value	0.97	0.98	0.99	0.99	1.00	1.00	1.00	1.00

Table 2: Evidence-based validation. We *(i)* collect from the 2013 and 2014 BAAC* data sets 32 groups of observations relative to accidents that occurred in similar contexts, *(ii)* compute, for each accident, the score of the involved GC (or one of the involved GCs), and *(iii)* test, within each group, if the conditional distribution of score given that the driver was fatally or severely injured ($\mathcal{I}_1$) is stochastically larger than the conditional distribution of score given that the driver was not fatally or severely injured ($\mathcal{I}_0$). We carry out Wilcoxon tests and report the p-values ranked by increasing order along with the cardinalities of the two samples used for each test.