# Classement global de classes générationnelles de modèles automobiles

Zaïd Ouni [1,2] & Antoine Chambaz [2] & Cyril Chauvel [1]

[1] *LAB, 132, rue des Suisses, 92000 Nanterre*
*zaid.ouni@lab-france.com*
*cyril.chauvel@lab-france.com*
[2] *Modal'X, Université Paris Ouest Nanterre, 200, av. de la République, 92000 Nanterre*
*achambaz@u-paris10.fr*

**Résumé.** Chaque année, les données BAAC (Bulletin d'Analyse des Accidents Corporels) incluent les accidents de la circulation sur la voie publique française impliquant un ou deux véhicules légers et blessant au moins un des occupants. Chaque véhicule léger se voit associer une "classe générationnelle" (CG), qui en donne une description sommaire. Deux véhicules légers de CGs distinctes n'offrent pas nécessairement la même sécurité à leurs passagers dans différents contextes d'accident. L'objectif de ce travail est d'évaluer dans quelle mesure les nouvelles générations de véhicules légers sont plus sûres que les anciennes à partir des données BAAC.

Dans une étude précédente, nous avons élaboré un méta-algorithme de classement contextuel de CGs de véhicules légers. Dans cette nouvelle étude, notre objectif est de développer une méthode de classement global (par opposition à "contextuel"). Nous procédons par "scoring" : nous cherchons une fonction de score qui associe à toute CG un nombre réel ; plus ce nombre est petit, plus la CG est sûre globalement. Nous utilisons des argumentations causales pour adapter le méta-algorithme en s'affranchissant du contexte.

**Mots-clés.** Analyse causale, classement, ensemble learning, scoring, sécurité automobile

**Abstract.** Each year, the BAAC (Bulletin d'Analyse des Accidents Corporels) data set includes traffic accidents on the French public roads involving one or two ligth 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 safety to their passengers in differents contexts. 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 a previous study, we elaborated a contextuel ranking meta-algorithm of GC of ligth vehicles. In the present study, our objective is to develop global (as opposed to contextual) ranking method across all contexts of accidents. 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. We use a causal argumentation to integrate out the context.

**Keywords.** Causal analysis, car safety, ensemble learning, ranking, scoring

# 1 Introduction

En 2016, les accidents de la route restent une priorité de santé publique aux niveaux mondial, européen et français. Les automobiles sont un des acteurs principaux de l'activité routière. Son amélioration passe donc notamment par une analyse des caractéristiques accidentologiques des automobiles. Les modèles de véhicule sont développés en bureaux d'études et validés en laboratoires. C'est néanmoins la réalité accidentologique qui permet de vraiment cerner les niveaux qu'ils offrent en matière de sécurité active (grâce aux systèmes d'aide à la conduite, qui assurent, par exemple, une meilleure tenue de route et un meilleur freinage) et de sécurité passive (grâce, par exemple, aux ceintures, airbags, structures à déformation programmée). Dans ce cadre, les experts en accidentologie du LAB (Laboratoire d'accidentologie, de biomécanique et du comportement conducteur) souhaitent disposer d'un outil statistique leur permettant, en interne, de suivre l'évolution au cours du temps de la sécurité offerte par des "classes générationnelles" (CG) de véhicules.

Dans notre précédente étude [Ouni et al, 2015], nous avons considéré la sécurité offerte par toute CG dans tout contexte d'accident. Concrètement, nous avons mis au point une procédure statistique permettant, pour toute paire $((x_1, w_1), (x_2, w_2))$ de couples constitués d'une CG ($x_1$ ou $x_2$) et d'un contexte d'accident ($w_1$ ou $w_2$), de déterminer quelle est la combinaison la plus sûre. Nous qualifions un tel classement de "contextuel". Dans la présente étude, nous considérons la sécurité offerte par toute CG *globalement* plutôt que contextuellement. Ainsi, l'objectif est l'élaboration d'une procédure statistique permettant, pour tout paire $(x_1, x_2)$ de CGs, de déterminer laquelle est la plus sûre globalement (c'est-à-dire, à travers une distribution de contextes).

L'étude [Ouni et al, 2015] repose sur le principe de "scoring" [Clémençon et al, 2008 et 2009] : nous cherchons une fonction de score qui associe à tout contexte et toute CG un nombre réel ; plus ce nombre est petit, plus la CG est sûre dans le contexte accidentel donné. La meilleur fonction de score est obtenue à partir de données réelles d'accident par validation croisée, sous forme d'une combinaison convexe optimale des fonctions de score fournies par une librairie d'algorithmes de classement.

Dans cette étude, nous utilisons le même principe : nous cherchons une fonction de score qui associe à toute CG un nombre réel ; plus ce nombre est petit, plus la CG est sûre globalement.

Nous nous appuyons sur les données BAAC (Bulletin d'Analyse des Accidents Corporels). Elles répertorient, chaque année, les accidents de la route ayant eu lieu sur la voie publique française et ayant conduit à au moins un blessé léger. Les bulletins sont

établis par les forces de l'ordre. Ils décrivent les conditions générales de l'accident (date, horaire, localisation géographique, type de choc, ...), le profil de tous les impliqués dans l'accident (âge, sexe, catégorie socio-professionnelle, alcoolémie, ... ) et les conséquences de l'accident pour les impliqués (indemne ou blessé léger ; blessé grave ou tué).

En complément de ces données nationales, des données de flottes automobiles permettent d'associer une CG à chacun des véhicules impliqués. Constituée de sept variables (segment, année de conception et cinq autres variables), la CG d'un véhicule le décrit sommairement. Dans la suite, les données BAAC sont associée avec les données des CGs. Uniquement les accidents à un seul ou deux véhicules en cause sont utilisés dans l'étude. Cette échantillon sera noté BAAC*.

# 2 Modélisation

Les données BAAC* viennent par "clusters", parce qu'un ou deux véhicules sont impliqués, et parce que nous adoptons le point de vue individuel des occupants des véhicules. Une description détaillée de la modélisation et de la distribution de données BAAC* est disponible dans [Ouni et al, 2015]. Chaque accident $\mathbf{O}$ est composé d'un ou deux (selon le nombre de véhicules impliqués dans l'accident) "clusters" $\mathbf{O}_k$ de variables dépendantes $\mathbf{O}_{kj}$ décrivant les données d'accident du point de vue de l'occupant $j$ du véhicule $k$.

Pour simplifier la présentation, nous procéderons comme si nous n'utilisions qu'un seul point de vue $\mathbf{O}_{kj}$ pour chaque accident $\mathbf{O}$. Par ailleurs, dans l'application, nous exploitons toutes les observations en utilisant [Ouni et al (2015), lemme 1, section 4].

Notons $O^1, \ldots, O^n$ les $n$ observations d'un jeu de données BAAC*. Nous les modélisons comme des variables aléatoires indépendantes, identiquement distribuées selon la loi $P$. Soit $P_n$ sa version empirique. Pour tout $1 \leq i \leq n$, la variable $O_i$ est décomposée comme suit : $O_i = (W_i, X_i, Z_i)$ où

- $Z_i$ est la sévérité de blessure de l'occupant du véhicule. C'est une variable binaire valant 1 si l'occupant est tué ou blessé grave (hospitalisé plus de 24h) et 0 s'il est indemne ou blessé léger (hospitalisé moins de 24h).
- $X_i$ est la CG associée au véhicule impliqué dans l'accident. C'est une description sommaire avec 7 variables : le segment, la date de conception et la date de première mise en circulation et 4 autres variables (qualitatives et quantitatives).
- $W_i$ est une description du contexte de l'accident et le profil de l'impliqué. Elle est composée de 29 variables qualitatives et quantitatives.

# 3 Modèle Causal et traitement statistique

L'objectif est d'apprendre à classer une CG en termes de sécurité offerte globalement. Afin de bien résoudre ce problème, nous menons une analyse causale. Tout ce qui suit

vient en complément de l'étude décrite dans [Ouni et al, 2015].

## Modèle causal

Soit $\mathbb{O} = (W, X, (Z_x)_{x\in\mathcal{X}})$ la donnée d'accident contrefactuelle qui décrit toutes les issues contrefactuelles $Z_x (x \in \mathcal{X})$ d'un accident impliquant une CG $x$ dans un contexte $W$, et la CG $X$ qui est effectivement impliquée dans l'accident. L'observation $\mathbf{O}$ est une variable aléatoire qui peut être obtenue à partir de $\mathbb{O}$ en supprimant les issues $Z_x$ pour tout $x \neq X$. La loi $P$ des observations $O$ est une loi marginale de la loi causale $\mathbb{P}$ de $\mathbb{O}$.

Soit $\mathbb{P}^{\otimes 2}$ la loi jointe de $(\mathbb{O}, \mathbb{O}')$ tirée en deux étapes : *(i)* tirer au hasard un contexte d'accident $W_1$ selon la loi marginale de $W$ sous $\mathbb{P}$, puis, *(ii)* tirer independamment $\mathbb{O}$ et $\mathbb{O}'$ selon la distribution obtenue de $\mathbb{P}$ sachant que $W = W' = W_1$.

Si, contrairement aux faits, nous avions accès aux observations contrefactuelles tirées sous $\mathbb{P}^{\otimes 2}$, notre objectif serait exprimé comme suit :

*(i)* apprendre une fonction $\rho : \mathcal{X} \times \mathcal{X} \to \{-1, 0, 1\}$ où $\rho(x, x') = 0$ si et seulement si (ssi) $x = x'$ et tel que la probabilité $\mathbb{P}^{\otimes 2}((Z_x - Z'_{x'})\rho(x, x') > 0)$ soit la plus petit possible pour tout $(x, x') \in \mathcal{X}^2$ ;

*(ii)* déclarer que, pour tout $(x, x') \in \mathcal{X}^2$ tel que $x \neq x'$, la CG $x$ est plus sûre que la CG $x'$ (globalement) ssi $\rho(x, x') = 1$.

Nous montrons que $\mathbb{P}^{\otimes 2}((Z_x - Z'_{x'})\rho(x, x') < 0)$ est minimale *ssi* $\rho = \rho_0$ où

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

avec $\mathbb{Q}(x, W) = E_{\mathbb{P}}(Z_x|W)$ (pour tout $x \in \mathcal{X}$).

Nous interprétons l'expression précédente de $\rho_0$ par : une CG $x$ est plus sûre qu'une CG $x'$ si $E_{\mathbb{P}}(Z_x) < E_{\mathbb{P}}(Z_{x'})$ et une CG $x'$ est plus sûre qu'une CG $x$ si $E_{\mathbb{P}}(Z_x) > E_{\mathbb{P}}(Z_{x'})$.

Nous appelons la règle optimale $\rho_0$ une "règle de scoring" : quand la fonction $x \mapsto E_{\mathbb{P}}(Z_x)$ de $\mathcal{X}$ vers $[0, 1]$ est connue alors $\rho_0$ est connue également. En particulier, l'estimation de $\rho_0$ peut être obtenue par l'estimation de $E_{\mathbb{P}}(Z_x)$ (pour tout $x \in \mathcal{X}$). Les données observées ne nous permettent pas de faire l'estimation de $E_{\mathbb{P}}(Z_x)$ (pour tout $x \in \mathcal{X}$) et $\rho_o$ dans le monde contrefactuel.

## Argumentation causale

Afin d'estimer $\mathbb{Q}(x, W)$ et $E_{\mathbb{P}}(Z_x)$ à partir des données observées $O = (W, X, Z = Z_X)$, nous supposons les hypothèses causales suivantes :

— *hypothèse de randomisation :* $X$ est indépendant de $(Z_x)_{x\in\mathcal{X})}$ sachant $W$ ;

— *hypothèse de consistance :* $Z_x = Z$ quand $x = X$ ;

— *hypothèse de positivité :* pour tout $x \in \mathcal{X}$, $P(X = x|W) > 0$, $P$-presque sûrement.

Sous ces hypothèses, il apparaît que, pour tout $x \in \mathcal{X}$ :

$$\begin{aligned} \mathbb{Q}(x, W) &= E_P(Z|X = x, W), \quad (1) \\ E_{\mathbb{P}}(Z_x) &= E_P[E_P(Z|X = x, W)]. \quad (2) \end{aligned}$$

Les hypothèses causales induisent ainsi un problème statistique qui peut être étudié dans le monde réel à partir des données réellement observées. Ce problème statistique fait par ailleurs sens indépendamment de modèle causal.

## Traitement statistique

Nous allons donc :

1. estimer $Q(x, W) = E_P(Z|X = x, W)$, pour tout $W \in \mathcal{W}$ et $x \in \mathcal{X}$,
2. estimer $s_0(x) = E_P[Q(x, W)]$, pour tout $x \in \mathcal{X}$.

Puis nous décidons que la CG $x$ est plus sûre que la CG $x'$ si $s_0(x) < s_0(x')$.

Soit la fonction de perte $\ell^1_{Q,\mu}$ :

$$\ell^1_{Q,\mu}(f, O) = \int_{\mathcal{X}} \Lambda(Q(x, W), f) d\mu(x)$$

avec $\Lambda(p, q) = p \log(\frac{p}{q}) + (1 - p) \log(\frac{1-p}{1-q})$ la divergence de Kullback-Leibler entre deux distributions de Bernoulli et $\mu$ une mesure de probabilité fournie par l'utilisateur.

La performance statistique d'un estimateur $s_n : \mathcal{X} \mapsto [0, 1]$ de $s_0$ est évaluée en se basant sur le risque :

$$\mathcal{R}_{\widetilde{Q},\widetilde{\mu}}(P)(s_n) = E_P[\ell^1_{\widetilde{Q},\widetilde{\mu}}(s_n, O)] \quad (3)$$

où $\widetilde{Q}$ est un estimateur de $Q$ construit dans [Ouni et al, 2015] par Super Learning [van der Laan, 2007] et $\widetilde{\mu}$ est la mesure empirique sur l'espace $\mathcal{X}$. Nous utilisons un jeu de données indépendantes de celui utilisé pour estimé $s_n$.

En pratique, nous ne pouvons pas explorer tout l'ensemble de fonctions de $\mathcal{X}$ vers $[0, 1]$. Ainsi, nous utilisons des modèles de travail paramétrique $\mathcal{F}_1, \ldots, \mathcal{F}_K$ tel que $\mathcal{F}_k = \{f_{k,\theta}, \theta \in \Theta_k\}$ est un ensemble de fonctions de $\mathcal{X}$ vers $[0, 1]$. En outre, nous supposons que $\theta \mapsto \mathcal{R}_{\widetilde{Q},\widetilde{\mu}}(P)(f_{k,\theta})$ admet un unique minimum $\widehat{\theta}_k(P)$ sur chaque modèle $\mathcal{F}_k$.

Pour chaque $1 \leq k \leq K$, nous supposons qu'il existe un unique minimum $\widehat{\theta}_k(P_n)$ de la version empirique de risque :

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

Nous obtenons un estimateur $f_{k,\widehat{\theta}_k(P_n)}$ de $s_0$ sur chaque modèle de travail $\mathcal{F}_k$, pour tout $1 \leq k \leq K$. Nous procédons à l'identification du meilleur modèle de travail en utilisant le risque cross-validé.

Soit $K_n$ l'indice du modèle de travail qui a le risque cross-validé le plus petit.
Finalement, notre estimateur de $s_0$ est :

$$S_n = f_{K_n, \widehat{\theta}_k(P_n)}.$$

# 4 Application

Les données contextuelles $W_i$ regroupent 29 variables qualitatives et quantitatives décrivant l'accident $i$ du point de vue d'un impliqué(conditions générales d'accident, profil du conducteur, profil de l'impliqué). Rappelons que les CGs $X_i$ regroupent sept variables.

L'estimateur $\widetilde{Q}$ est construit à partir de $K = 49$ algorithmes individuels et des $n =$ 16 877 accidents qui ont eu lieu entre un ou deux véhicules légers en 2011. Le nombre de personnes impliquées s'élève à 37 721. La mesure $\widetilde{\mu}$ est la mesure empirique sur l'espace $\mathcal{X}$ des 1 000 CGs impliquées en 2011. Nous utilisons 5 000 accidents du BAAC* de 2012 pour calculer le risque cross validé et entraîner l'estimateur $S_n$.

La validation industrielle de notre approche repose notamment sur la comparaison de CGs de véhicules de différentes générations au sein du même segment. Il est attendu qu'au sein d'un même segment, un véhicule d'une génération plus récente surclasse un véhicule d'une génération plus ancienne en termes de sécurité passive. Les résultats obtenus sont conformes aux attentes des experts en accidentologie.

Nous évoquerons plus en détails ces résultats, ainsi que d'autres qui confirment que notre procédure est performante.

# 5 Discussion

Les CGs et les données contextuelles ont vocation à être enrichies, par exemple avec les dimensions, silhouettes, systèmes de sécurité embarqués, vitesses d'impact. Cet enrichissement ne remettra en cause ni la théorie ni les algorithmes que nous avons développés.

# Bibliographie

[1] Z. Ouni, C. Denis, C. Chauvel and A. Chambaz (2015), *Contextual ranking by passive safety of generational classes of light vehicles*, https ://hal.archives-ouvertes.fr/hal-01194515.
[2] M. J. van der Laan, E. Polley and A. E. Hubbard (2007), Super learner, Stat. Appl. Genet. Mol. Biol., 6 : Art. 25.
[3] S. Clémençon and N. Vayatis (2009), True-based ranking methods. IEEE Trans. Inform. Theory, 55(9) :4316-4336.
[4] S. Clémençon, G. Lugosi, and N. Vayatis (2008), Ranking and empirical minimization of U-statistics, Ann.statist., 36(2) :844-874, 2008.