ED 139 : Ecole doctorale
Connaissance, langage, modélisation
EA 3454 - Modélisation aléatoire de Paris-X

Zaïd OUNI

Statistique pour l'anticipation des niveaux de sécurité secondaire des générations de véhicules

Thèse présentée et soutenue publiquement le 19 juillet 2016
en vue de l'obtention du doctorat de
Mathématiques appliquées et applications des mathématiques
de l'Université Paris Ouest Nanterre La Défense

sous la direction de M. Antoine Chambaz

Jury :

Rapporteur	M. Avner Bar-Hen	CNAM
Rapporteur	M. Vivian Viallon	IFSTTAR et Université Lyon 1
Membre du jury	M. Antoine Chambaz	Co-encadrant, Université Paris Ouest Nanterre La Défense
Membre du jury	M. Cyril Chauvel	Co-encadrant, LAB
Membre du jury	Mme. Cécile Durot	Examinatrice, Université Paris Ouest Nanterre La Défense
Membre du jury	Mme. Anne Guillaume	Examinatrice, LAB
Membre du jury	M. Jean-Louis Martin	Examinateur, IFSTTAR
Membre du jury	M. Jean-Christophe Thalabard	Examinateur, Université Paris Descartes
Membre du jury	M. Vivian Viallon	Rapporteur, IFSTTAR et Université Lyon 1

Résumé

La sécurité routière est une priorité mondiale, européenne et française. Parce que les véhicules légers (ou simplement "les véhicules") sont évidemment l'un des acteurs principaux de l'activité routière, l'amélioration de la sécurité routière passe nécessairement par l'analyse de leurs caractéristiques accidentologiques. Si les nouveaux véhicules sont développés en bureau d'étude et validés en laboratoire, c'est la réalité accidentologique qui permet de vraiment cerner comment ils se comportent en matière de sécurité secondaire, c'est-à-dire quelle sécurité ils offrent à leurs occupants lors d'un accident. C'est pourquoi les constructeurs souhaitent procéder au classement des générations de véhicules en fonction de leurs niveaux de sécurité secondaire réelle.

Nous abordons cette thématique en exploitant les données nationales d'accidents corporels de la route appelées BAAC (Bulletin d'Analyse d'Accident Corporel de la Circulation). En complement de celles-ci, les données de parc automobile permettent d'associer une classe générationelle (CG) à chaque véhicule. Nous élaborons deux méthodes de classement de CGs en termes de sécurité secondaire. La première produit des classements contextuels, c'est-à-dire des classements de CGs plongées dans des contextes d'accident. La seconde produit des classements globaux, c'est-à-dire des classements de CGs déterminés par rapport à une distribution de contextes d'accident.

Pour le classement contextuel, nous procédons par "scoring" : nous cherchons une fonction de score qui associe un nombre réel à toute combinaison de CG et de contexte d'accident ; plus ce nombre est petit, plus la CG est sûre dans le contexte d'accident donné. La fonction de score optimale est estimée par "ensemble learning", sous la forme d'une combinaison convexe optimale de fonctions de score produites par une librairie d'algorithmes de classement par scoring. Une inégalité oracle illustre les performances du méta-algorithme ainsi obtenu. Le classement global est également basé sur le principe de "scoring" : nous cherchons une fonction de score qui associe à toute CG un nombre réel ; plus ce nombre est petit, plus la CG est jugée sûre globalement. Des arguments causaux permettent d'adapter le méta-algorithme évoqué ci-dessus en s'affranchissant du contexte d'accident. Les résultats des deux méthodes de classement sont conformes aux attentes des experts.

Mots clés : agrégation, analyse causale, classement, inégalité oracle, sécurité routière, statistique.

Statistics for anticipating the levels of secondary safety for generations of vehicles

Abstract

Road safety is a world, European and French priority. Because light vehicles (or simply "vehicles") are obviously one of the main actors of road activity, the improvement of road safety necessarily requires analyzing their characteristics in terms of traffic road accident (or simply "accident"). If the new vehicles are developed in engineering department and validated in laboratory, it is the reality of real-life accidents that ultimately characterizes them in terms of secondary safety, ie, that demonstrates which level of security they offer to their occupants in case of an accident. This is why car makers want to rank generations of vehicles according to their real-life levels of safety.

We address this problem by exploiting a French data set of accidents called BAAC (Bulletin d'Analyse d'Accident Corporel de la Circulation). In addition, fleet data are used to associate a generational class (GC) to each vehicle.

We elaborate two methods of ranking of GCs in terms of secondary safety. The first one yields contextual rankings, ie, rankings of GCs in specified contexts of accident. The second one yields global rankings, ie, rankings of GCs determined relative to a distribution of contexts of accident.

For the contextual ranking, we proceed by "scoring": we look for a score function that associates a real number to any combination of GC and a context of accident; the smaller is this number, the safer is the GC in the given context. The optimal score function is estimated by "ensemble learning", under the form of an optimal convex combination of scoring functions produced by a library of ranking algorithms by scoring. An oracle inequality illustrates the performance of the obtained meta-algorithm. The global ranking is also based on "scoring": we look for a scoring function that associates any GC with a real number; the smaller is this number, the safer is the GC. Causal arguments are used to adapt the above meta-algorithm by averaging out the context. The results of the two ranking procedures are in line with the experts' expectations.

Keyswords: ensemble learning, causal analysis, oracle ineguality, ranking, road safety, statistics.

Cette thèse est une thèse CIFRE préparée aux:

- Laboratoire de modélisation aléatoire (Modal-X), Université Paris Ouest Nanterre La Défense
  20 avenue de la république, 92001 Nanterre
- Laboratoire d'accidentologie, de biomécanique et du comportement des conducteurs
  132, rue des suisses, 92000 Nanterre

Table des matières

Chapitre I

Introduction

I.1 Introduction

Contexte

Chaque année, près de 1,25 million de personnes décèdent dans un accident de la route et entre 20 et 50 millions sont blessées dans le monde. Selon l'organisation mondiale de la santé (OMS), en 2015, ces traumatismes représentent la huitième cause de décès au niveau mondial et la première cause de décès chez les jeunes âgés de 15 à 29 ans. 90% des décès se produisent dans les pays à revenue faible ou intermédiaire, qui possèdent 54% du parc de véhicules motorisés. Selon les prévisions de l'OMS, dans l'état actuel des choses, les accidents de la route deviendront la septième cause de mortalité d'ici 2030. En septembre 2015, l'Assemblée Générale des Nations Unies a fixé un objectif ambitieux pour la sécurité routière : diviser par deux le nombre total des morts et de blessés dûs aux accidents de la route à l'horizon 2020.

Au niveau Européen, le nombre de décès est en diminution continue depuis quinze ans. En 2014, 25 700 personnes sont mortes et plus de 200 000 ont été blessés gravement sur les routes européennes. Le nombre de véhicules motorisés a atteint 370 millions, avec une progression de +15% depuis 2005. Le taux de mortalité moyen dans tous les états européens est de 51 décès par million d'habitants en 2014. Les routes maltaises, hollandaises, anglaises et suédoises sont les plus sûres en Europe. Elles ont un taux de mortalité au-dessous de 30 décès par million d'habitants. L'objectif de la Commission Européenne, en matière de sécurité routière, pour la décennie en cours (2011-2020) est de réduire de moitié le nombre de morts dûes aux accidents de la route. Si cet objectif est atteint, 90 000 vies pourront être sauvées.

En France, selon le rapport définitif de l'accidentalité routière 2015 de l'Observatoire National Interministérielle de la Sécurité Routière (ONISR) [1], 56 603 accidents corporels ont été enregistrés sur les routes, représentant 3 461 morts. La mortalité automobiliste en 2015 représente 52% de la mortalité totale. D'après la même source [1], le nombre de blessés est estimé à 70 802 en 2015, dont 26 595 blessés hospitalisés plus de 24 heures.

Selon les chiffres de la Comité des Constructeurs Français d'Automobile (CCFA), en 2015, le parc automobile français atteint 38 408 000 véhicules au 1er Janvier 2015. L'objectif du gouvernement est de réduire la mortalité routière au-dessous de 2000 morts en 2020.

Malgré l'augmentation du nombre de véhicules présents dans le parc, la baisse du nombre de morts sur la route est bien présente. Elle est le résultat d'actions mises en œuvres par l'ensemble des acteurs de la sécurité routière et d'objectifs ambitieux du gouvernement français demandant le renforcement des efforts de tous les responsables de l'activité routière (pouvoirs publics, constructeurs et équipementiers automobiles, usagers de la route).

Que'est ce que un accident de la route ?

Un accident de la route est un dysfonctionnement du système homme-véhicule-infrastructure. Les leviers d'amélioration de la sécurité routière concerne donc les trois parties de ce système. L'application stricte des lois et des campagnes de sensibilisation, contre les principaux facteurs de risque (la vitesse, l'alcoolémie, le non port de la ceinture, le non port du casque, la non utilisation des dispositifs de retenue pour enfants (DRE)), sont nécessaires pour améliorer les comportements et la protection des usagers de la route.

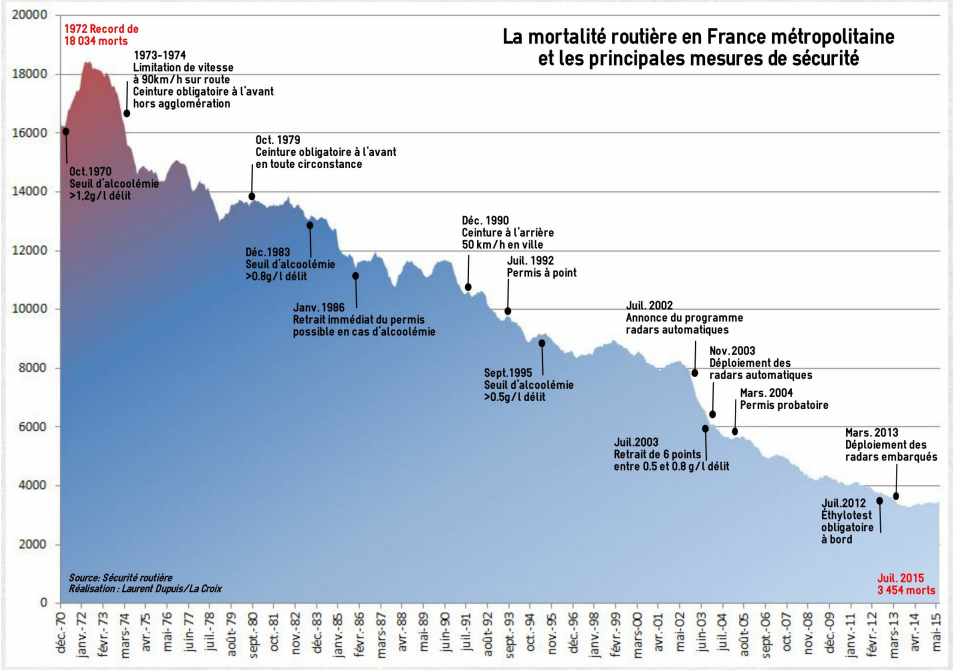

FIGURE I.1 – Évolution de la mortalité routière en France métropolitaine et les principales mesures de sécurité

La figure I.1 illustre les effets des mesures appliquées par le gouvernement français depuis le pic de mortalité routière de 1972. En complément des législations, l'aménagement de l'infrastructure routière (construction d'autoroutes, entretien de chaussées, protection des obstacles fixes latéraux, ...) est indispensable pour réduire le risque d'accident et la gravité des blessures. Parallèlement, les constructeurs automobiles conçoivent des voitures

de plus en plus sûres. Celles-ci sont équipées de multiples systèmes de protection pour ses occupants en cas d'accident (ceinture, airbags, structure, ...) et de systèmes d'assistance à la conduite (frein, ABS, ESP, ...), afin d'éviter l'accident ou en réduire la violence du choc.

Les études accidentologiques

L'ensemble des constructeurs mondiaux mènent des études accidentologique aidant à concevoir leurs nouveaux véhicules tant au niveau des protection interne offerte qu'en externe (partner protection, VRU,...). Ils le font de manière soit indépendante soit conjointe. Depuis 1969, les deux constructeurs automobiles français s'investissent en sécurité via le Laboratoire d'Accidentologie, de Biomécanique et du comportement des conducteurs (LAB).

I.2 Le LAB et ses activités

Le LAB

Le LAB est un laboratoire de recherche commun à PSA-Peugeot/Citroën et Renault. Il a été créé en 1969 et il a pour mission : *(i)* d'acquérir et de transmettre de la connaissance sur les mécanismes accidentels et lésionnels des accidents de la route et *(ii)* de spécifier et d'évaluer l'efficacité des contremesures en s'appuyant sur la réalité des accidents.

Le LAB intervient auprès de nombreux acteurs de la sécurité routière. Il traite principalement les demandes provenant des deux constructeurs, PSA et Renault. De plus, il existe de nombreuses collaborations avec des instituts, des laboratoires et des institutions liées au monde d'automobile, à la sécurité routière ou au monde universitaire.

Ses activités

La biomécanique des chocs, les études du comportement de conducteur et l'accidentologie forment ses 3 activités :

La biomécanique des chocs est l'étude de la réponse du corps humain au choc. Cette étude permet d'identifier les liens entre les lésions du corps humain et les efforts et/ou décélérations. Ce travail permet de caractériser des limites en dessous desquelles les blessures peuvent être évitées. Ces limites varient en fonction des caractéristiques humaines (l'âge, la taille, la morphologie, ...).

L'étude du comportement des conducteurs a pour objectif de comprendre et d'évaluer le comportement du conducteur. Ce dernier est le principal acteur de l'accident, il est donc important de pouvoir décrire son comportement durant toutes les phases qui précèdent l'accident. Cela consiste à étudier ses perceptions (du véhicule adverse, du piéton qui traverse, ...), ses interprétations de la situation (évaluation de

la situation, du danger, ...), ses décisions face à la situation (tenter un évitement, accélérer, freiner, ...) et ses actions effectives (vitesse et amplitude des coups de volant, blocage de roues, ...)

L'accidentologie est l'étude scientifique des accidents de la route. Elle a pour objectif la compréhension des mécanismes accidentels et des conséquences corporelles sur les usagers. C'est une discipline à la croisée de plusieurs disciplines (dynamique du véhicule, épidémiologie, médecine, statistique).

Cette activité est articulée au LAB autour de 3 volets :

- les études d'accidents et la construction de différentes bases de données sur les accidents et les lésions ;
- le diagnostic de sécurité routière (qui dresse un panorama des problèmes accidentels et lésionnels qui restent à résoudre) ;
- l'évaluation de l'efficacité attendue ou de l'efficacité observée des mesures de prévention de l'accident ou de protection des usagers (notamment les occupants de voitures).

C'est dans le cadre de ce troisième axe que s'inscrit les travaux de cette thèse.

I.3 L'accidentologie

L'accidentologie routière est une science qui a pour objectif de comprendre l'étiologie de l'accident et des lésions, afin de proposer des contre-mesures pour diminuer le nombre des accidents de la route et des lésions corporelles conséquentes. La compréhension de ces mécanismes nécessite, entre autres, des études approfondies sur les accidents réels et la construction de modèles théoriques et de grilles de lecture et d'analyse de l'accident permettant à l'accidentologiste de structurer le déroulement de l'accident et de démêler les interactions entre les facteurs et les causes l'ayant provoqué. Un des modèles les plus connus en France est le modèle séquentiel qui décompose l'accident en 5 phases :

1. **la situation précédant le déplacement** dans laquelle va notamment s'opérer le choix modal ;
2. **la situation de conduite** qui est explicative de la nature et des conditions du déplacement, et des stratégies adoptées en abord du site de l'accident ;
3. **la situation de rupture ou d'accident** qui est créée généralement par un élément nouveau ou imprévu ;
4. **la situation d'urgence** qui est celle où le conducteur doit impérativement exécuter une manœuvre de sauvegarde sous contraintes de temps et d'espace pour éviter le choc ;
5. **la situation de choc** qui comprend les conditions du choc lui-même et les évènements consécutifs au choc.

Alors que l'accidentologie primaire traite pour l'essentiel des phases 1 à 4, l'accidentologie secondaire traite de la situation des occupants et de leur véhicule dans la phase 5 : ce sont les conséquences des accidents qui sont alors analysées en détail pour fournir aux

constructeurs automobiles une expertise qui doit leur permettre d'anticiper les priorités à venir en termes de protection des occupants, et de développer la capacité de vérifier l'efficacité des contre-mesures mises en œuvre avec la plus grande réactivité possible.

Les experts en accidentologies distinguent trois niveaux de sécurité et d'études :

L'accidentologie Primaire

L'accidentologie Primaire ou Active a pour but de mieux connaître les mécanismes des accidents pour tenter de les éviter ou de réduire la violence du choc, d'estimer les gains potentiels de victimes par tel ou tel nouveau système et de préciser les cibles prioritaires en termes de victimes évitables ;

La sécurité Primaire désigne toute action amenant à prévenir l'accident avant que ce dernier ne se produise. Le travail du constructeur automobile dans ce domaine porte donc principalement sur la tenue de route, le freinage, la visibilité, la perception de l'environnement extérieur, le confort,...

L'accidentologie Secondaire

L'accidentologie Secondaire ou passive, a pour but de mieux connaître le monde réel des accidents.

Ses objectifs sont les suivants :

- décrire les mécanismes des lésions par types de choc, par niveau de sévérité ou par fréquence.
- vérifier sur le terrain l'efficacité des systèmes de sécurité.
- estimer les gains potentiels de victimes par tel ou tel nouveau système.
- préciser les cibles prioritaires en termes de victimes évitables.

La sécurité Secondaire permet la réduction du risque de blessures en améliorant la protection des usagers lorsque l'accident n'a pas pu être évité. Des modifications de la structure du véhicule (habitable rigide et avant déformable, meilleure conception du volant et de la colonne de direction), une ceinture de sécurité, un sac gonflable (airbag), un casque, sont des éléments qui contribuent à la sécurité secondaire.

L'accidentologie Tertiaire

L'accidentologie Tertiaire, qui est l'activité la plus récente au LAB, consiste à mieux connaître les mécanismes de prise en charge des impliqués suite à un accident de la route.

La sécurité Tertiaire vise à réduire le risque de mortalité, l'aggravation des lésions avant l'intervention des secours ou au cours du transport médicalisé ainsi que le risque de

séquelles post-traumatiques par une meilleure prise en charge de l'usager accidenté par les services de secours : rapidité d'intervention par une meilleure localisation de l'accident, qualité de l'intervention par une connaissance anticipée de l'accident et du niveau de gravité des impliqués, qualité de la désincarcération des occupants,...

Dans la suite de ce manuscrit, nous nous interesserons exclusivement à l'accidentologie et à la sécurité secondaire. Nous l'appellerons simplement **la sécurité réelle**.

I.4 Enjeux et objectifs

Les modèles automobiles sont développés en bureaux d'études et validés en laboratoire de crash-test. C'est néanmoins la réalité accidentologique qui permet de cerner parfaitement les niveaux qu'ils offrent en matière de sécurité. Dans ce cadre, les constructeurs automobiles français, via le LAB, souhaitent disposer d'un outil statistique leur permettant, en interne, de suivre l'évolution au cours du temps de la sécurité réelle des différentes Classes Générationnelles (CG) de leurs voitures.

La sécurité réelle offerte par CG est souvent évaluée à partir des conséquences de l'accident sur les occupants. Ces conséquences dépendent du CG (sa forme, systèmes de sécurité embarqués, sa date de conception, ...), de l'environnement de l'accident (conditions atmosphériques, luminosité, localisation, type d'obstacle, ...), du profil du conducteur et des passagers. Ainsi, l'évaluation de la sécurité réelle offerte par les technologies introduites dans la CG, en tenant en compte de l'effet des autres facteurs de risque (les conditions atmosphériques, la localisation, l'âge du conducteur, ...), est délicate. Cette évaluation est rendue encore plus difficile, par la moindre implications en accidents des CGs les plus récentes.

La sécurité "potentielle" offerte par une nouvelle CG est notée par des organisations consuméristes, dont Euro NCAP en Europe, en utilisant des données de crash tests. L'objectif de cette notation est d'informer les acheteurs des voitures nouvellement produites des niveaux de sécurité offerts. L'Euro NCAP note la sécurité potentielle par un score qui varie d'une à cinq étoiles : plus la voiture est sûre plus celle-ci obtient d'étoiles. Afin de garantir la note maximale, les constructeurs automobiles ont besoin d'évaluer en amont de la mise en circulation, le niveau de sécurité potentielle offert par leurs voitures.

Dans ce cadre, cette thèse consiste à élaborer une méthode statistique pour :

1. classer les CGs existantes en terme de sécurité réelle ;
2. positionner, dans ce classement, une nouvelle CG, juste après ou même avant sa sortie sur le marché.

I.5 Bases de données

Lorsque nous parlons d'études et de modèles statistiques, il est fondamental d'y associer un certains nombre de bases de données pour faire tourner ces modèles. Dans le

domaine de l'accidentologie, le LAB dispose de nombreuses bases de données d'accidents de la route avec des volumes et des niveaux de granulométrie différents, ainsi que d'une base de données du parc automobile français. Ces différentes bases de données sont classées en 3 catégories différents :

- les bases de données macro d'accidents : elles regroupent un volume important des données accidents avec un niveau limité de détails accidentologiques (sévérité du choc, les systèmes embarqués dans les véhicules impliqués, bilans lésionnels ...). Elles peuvent contenir jusqu'à une centaine de variables différentes.
- les bases de données micro d'accidents : elles regroupent un volume limité des données accidents avec un niveau important de détails accidentologiques. Elles peuvent contenir jusqu'à plus de mille variables différentes.
- les bases de données micro d'accidents : elles regroupent un volume limité des données accidents avec un niveau important de détails accidentologiques. Elles peuvent contenir jusqu'à plus de mille variables différentes.
- les bases de données d'exposition : ces données d'exposition sont souvent utilisée pour caractériser le risque d'une population et pour comparer plusieurs populations entre elles. La base de données d'exposition utilisée au LAB, aujourd'hui, est la base nationale d'immatriculation des voitures.

Toutes ces bases sont décrites plus précisément en B de ce manuscrit

Dans cette thèse, seules 2 bases ont été utilisées. Tout da'bord la base de données nationales d'accidents corporels de la route (BAAC). Ces données sont souvent nommés par données BAAC (Bulletin d'Analyse des Accidents Corporels). Elles répertorient, chaque année, les accidents de la route ayant ont lieu sur les voies publiques françaises et ayant conduit au moins à un blessé léger. Les bulletins sont établis par les forces de l'ordre. Ils décrivent les conditions générales de l'accident. Cette base de données est la source officielle utilisée par les pouvoirs publics pour communiquer sur les chiffres de la réalité accidentologique en France. Cependant dans [2], [25], les auteurs montrent qu'il existe un écart entre les chiffres de blessés recensés par les forces de l'ordre et la réalité accidentologique. Cet écart a été observé suite à un rapprochement de la base de données BAAC et le registre médicale, qui recense les blessées dans les accidents corporels de la route dans le département de Rhône. La méthode capture-recapture [2] permet d'estimer le nombre de blessés n'étant enregistrés par aucune des deux sources de collecte.

La deuxième base utilisée est, comme mentionné plus haut, une base de données d'exposition : données du Parc automobile français. C'est la base de données des voitures immatriculées en France, achetée chaque année par le LAB à l'AAA. Cette extraction contient 22 variables pour décrire les véhicules immatriculés. Ces données sont utilisées en complément des données BAAC pour traiter notre problématique.

Choix de la base de données

Nous avons choisi de travailler avec les données BAAC pour deux raisons : *(i)* cette base de données est continue dans le temps. *(ii)* cette base de données est la base la plus complète (en termes de volume) parmi toutes les bases disponibles au LAB (Annexe B). Ainsi, la continuité de la base dans le temps garantie la continuité d'utilisation de la méthode de classement. En outre, le volume de la base donne plus de performance à notre approche statistique.

Nous avons utilisé les accidents constitués d'un seul ou de deux véhicules légers en cause, disponibles dans la base nationale BAAC. En effet, Nous nous intéressons aux classements des CGs des véhicules en termes de protection offerte vis-à-vis de leurs occupants. Afin d'éviter une sous-évaluation ou une sur-évaluation de la protection offerte, nous écartons les accidents impliquant des véhicules lourds (véhicule utilitaire, poids lourds, ...) ou des usagers vulnérables (piéton, cycliste, motocycliste). En complément de ces données d'accidents, les données du parc sont utilisées pour associer une CG à chacun des véhicules impliqués. La base de données obtenue sera noté BAAC*.

Parce qu'un ou deux véhicules sont impliqués, et parce que nous adaptons le point de vue individuel des occupants des véhicules, les données viennent par "clusters". Chaque composante du cluster (la description de l'accident du point de vue d'un occupant) est une combinaison de la description du contexte de l'accident et la description de la CG du véhicule. En effet, le contexte de l'accident est décrit par 30 variables : les circonstances de l'accident (date, infrastructure, type du choc, ...), le profil du conducteur (âge, sexe, alcoolémie, ...) et le profil du l'occupant (âge, sexe, ...) auxquels nous nous intéressons. La CG du véhicule est décrit par 7 variables (segment, date de mise en circulation, date de conception et 4 autres variables).

Les données BAAC* des années 2011 à 2014 sont utilisées dans ces travaux. Dans la première partie de l'étude (classement contextuel), l'entrainement de l'algorithme de classement est effectué avec les données BAAC* 2011 et la validation est faite avec les données BAAC* 2012. Dans la seconde partie de l'étude (Classement global), les données BAAC* 2011 et 2012 sont utilisées conjointement pour l'entrainement de l'algorithme de classement. En outre, les données BAAC* 2013 et 2014 sont exploitées pour la validation.

I.6 Etat de l'art de l'évolution de la sécurité automobile

La sécurité automobile est un enjeu important pour les constructeurs automobiles. Depuis la naissance de la première voiture, de nombreux équipements de sécurité ont été conçus et ont été installés sur les voitures afin de les rendre plus sûres.

Divers systèmes de freinage ont été développés bien avant l'arrivée d'un freinage global sur les 4 roues. Avec l'évolution de vitesse de la circulation, de nouveaux équipements sont apparus, essentiellement pour améliorer la vision : rétroviseur, essuie-glaces, feux de croisement et de brouillard, puis les clignotants.

Les système de sécurité passive

Ensuite, les développements en matière de sécurité ont été axés sur la protection des occupants en cas d'accident (sécurité secondaire). Les ceintures de sécurité à deux points de fixation sont apparues dans les années 50, puis celles à trois points fixation, dans le but d'éviter l'impact du conducteur ou du passager avant avec le volant ou le tableau de bord et de mieux repartir sur l'ensemble du corps les efforts créés par la décélération. Par la suite, des enrouleurs automatiques ont été rajoutés à la ceinture pour que les passagers soient plus libres de leurs mouvements. A partir de 1984, la ceinture est équipée de prétentionneur pyrotechnique qui se retend sur l'occupant en cas de choc et limites les jeux. Par ailleurs, l'effort le plus important de la ceinture, pouvant provoquer des blessures aux niveaux du thorax et du bassin, est contrôlé par des limiteurs d'effort qui modulent l'effort de la ceinture et délèguent ensuite la fin du choc aux coussins gonflables (cf ci-dessous). Aujourd'hui, les ceintures de sécurité de certains modèles d'automobiles sont équipés d'un enrouleur électrique réversible, qui optimise la position de la ceinture dans des situations de conduites sportives ou bien avant des situations critiques d'accidents. Il peut être activé grâce à des informations provenant de systèmes de sécurité active : ESC, assistance au freinage (cf ci-dessous). Il peut également se déclencher grâce aux données des capteurs d'environnement tel que le radar avant.

Comme évoqué dans le paragraphe précédent, le second équipement de sécurité secondaire est le coussin gonflable de sécurité ou plus communement appelé airbag. Il a été introduit en 1973 afin de protèger la tête du conducteur d'un éventuel contact avec le volant, ainsi que pour protéger les occupants en plus de la ceinture. En 1986, un dispositif, le Proconten, est apparu. C'était un dispositif mécanique, constitué de câbles fixés à l'avant de la voiture qui rétractent le volant en cas d'enfoncement de l'avant. Ce système n'a pas eu de suite et a été dépassé par le coussin gonflable. Après les airbags frontaux installés dans le volant et au-dessus de la boite à gants, d'autres airbags sont apparus pour des protections plus ciblées : l'aibag latéral pour protéger le bassin et le thorax ; l'aibag rideau pour protéger la tête en choc latéral et l'airbag de genoux en choc frontal. La figure I.2 montre les emplacements de certains de ces airbags.

La plupart des voitures d'aujourd'hui ont une structure à déformation programmée. C'est une structure avec un avant déformable de manière maitrisée et une cellule habitable rigidifiée, pour réduire le risque des blessures dûes aux intrusions dans l'habitacle en cas d'accident. En effet, cette déformation permet d'absorber l'énergie cinétique en cas de choc frontal et d'éviter un niveau de décélération inacceptable pour le corps humain. Dans le cas de choc latéral, la distance de déformation est réduite. Ainsi, les portes sont rigidifiées par des barres de protection latérale. Les zones rouges dans la figure I.3 sont les zones de renforcement de l'habitacle et de dissipation de l'énergie. La figure I.4 illustre la différence de déformation que subissent trois véhicules de trois générations différentes, impliqués dans des accidents à sévérité d'impact proche. La structure du premier véhicule (I.4(a)) est une structure sans déformation programmée. Les véhicules (I.4(b), I.4(c)) sont constitués de structure à déformation programmée. Ainsi, la déformation du premier véhicule (I.4(a)) est plus importante que la déformation de celles des deux autres (I.4(b), I.4(c)). En outre, le conducteur du premier véhicule a été tué, le conducteur du deuxième véhicule et les occupants du troisième véhicule ont été blessés légers.

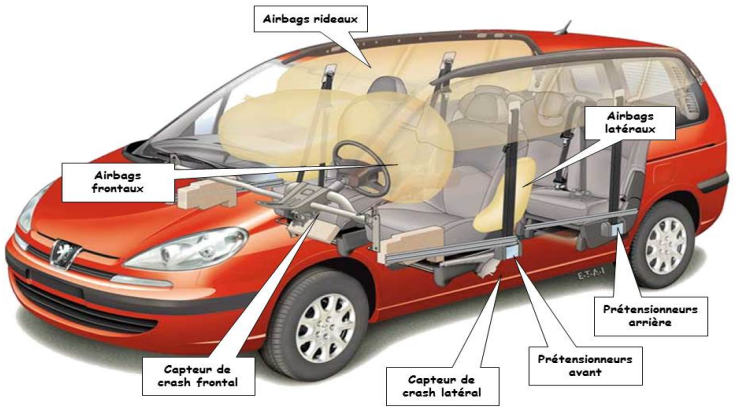

FIGURE I.2 – Emplacements d'airbags

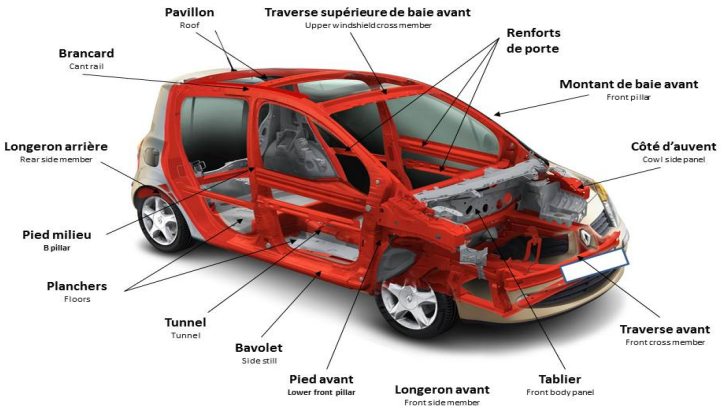

FIGURE I.3 – Zones de renforcements de l'habitacle et de dissipation de l'énergie

Divers autres systèmes de protection sont développés pour améliorer la protection des occupants : la colonne de direction collapsable pour que le volant ne vienne pas augmenter les efforts sur la cage thoracique du conducteur et pour laisser plus de temps de décélération et d'absorption d'énergie ; le pare-choc à l'avant et à l'arrière du véhicule pour absorber l'énergie cinétique du véhicule en cas de choc faible dit "choc réparabilité" et la bosse en avant de l'assise pour limiter le risque de rotation du bassin sous la ceinture que l'on nomme sous-marinage.

Les systèmes de sécurité active

En plus des systèmes de protection au moment de l'accident, les constructeurs automobiles ont développé d'autres systèmes d'assistance à la conduite mise en œuvre en situation d'urgence, afin d'éviter l'accident ou de limiter la violence du choc. Considérons

(a) Le conducteur est tué, ESS = 55km/h

(b) Le conducteur est blessé léger, ESS = 55km/h

(c) Quatre occupants sont blessés légers, ESS = 65km/h

FIGURE I.4 – Effet de la structure à déformation programmée

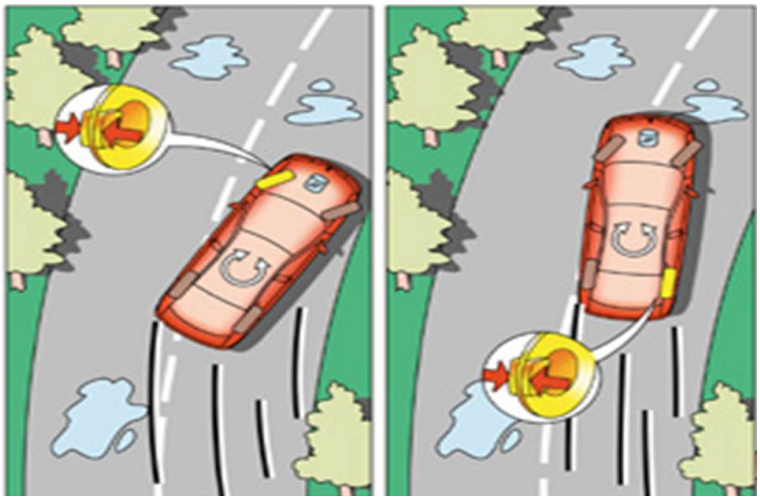

FIGURE I.5 – Fonctionement de Electronic Stability Control (ESC)

par exemple, un conducteur qui, dans une situation d'urgence, appuie brusquement sur la pédale de frein. Alors, le système d'aide au freinage d'urgence (AFU) rajoute une force supplémentaire de freinage et permet de réduire la distance de freinage. Ensuite, le système d'anti-blocage des roues (ABS) se met en route pour éviter le blocage en régulant la force d'assistance et maintenir le contrôle du véhicule. Ces deux premiers systèmes d'assistance au freinage fonctionnent quand le conducteur a appuyé sur la pédale de frein. Enfin, le correcteur électronique de trajectoire Electronic Stability Control (ESC) ajuste le freinage de manière indépendante sur chaque en cas de perte de contrôle transversale. En cas de sous virage (le véhicule allant tout droit), le système (ESC) freine la roue arrière intérieur au virage de façon à ce que le véhicule reprenne sa trajectoire et suive la consigne donnée au volant. Dans le cas contraire (sur-virage : le véhicule a tendance à prendre un virage plus serré), l'ESC freine la roue avant externe au virage de façon de redresser le véhicule pour qu'il suive la consigne donnée au volant. Une illustration de fonction de ESC est dans la figure I.5.

Les experts en accidentologie au LAB ont réalisé plusieurs études pour évaluer l'efficacité des systèmes de sécurité présentés ci-dessus. Les résultats de ces études sont résumés dans la figure I.6.

Ces dernières années, les nouveaux véhicules sont équipés par un système de freinage d'urgence automatique Autonomous Emergency Braking (AEB). Quand un obstacle (voiture, piéton, deux roues) est détecté sur la trajectoire, par la camera et le radar installés à l'avant, l'AEB prévient le conducteur. Si le conducteur ne fait aucune action, le système applique une pression automatique sur le frein pour ralentir le véhicule et tenter de minimiser la gravité du choc lorsque la collision n'est pas évitable. Dans le cadre de contrôle électronique de trajectoire, un régulateur adaptatif de vitesse (ACC Adaptative Cruise Control) a été développé pour réguler la vitesse du véhicule en fonction des conditions de trafic.

Dans l'objectif d'amélioration de connaissance de l'environnement de conduite, un système d'alerte de franchissement involontaire de ligne Line Kepping Warning (LKW)

FIGURE I.6 – Évaluation d'efficacité

alerte le conducteur du véhicule par des vibrations au niveau de l'assise du siège ou au niveau de la jante de volant. Dans un autre système, chaque coté du véhicule est équipé par une caméra numérique surveillant l'entrée d'un objet mobile (voiture, camion, moto) dans l'angle mort du véhicule et non visible dans le rétroviseur. Lorsque le système de suppression de l'angle mort "Blind Spot Information System" (BLIS) détecte une présence, une lumière clignote sur le rétroviseur de l'angle associé.

Cette liste de systèmes d'assistance à la conduite n'est pas exhaustive. Il existe bon nombre d'autres systèmes qui ne sont pas cités dans ce rapport et qui sont soit déjà présents sur des véhicules soit encore en cours de développement.

I.7 Etat de l'art de classements en sécurité des modèles automobiles

I.7.1 Classements prospectif en sécurité des modèles automobiles (Euro NCAP)

Les nouveaux modèles automobiles doivent respecter la réglementation du pays de production en matière de sécurité. Dans le but de comparer et promouvoir leurs sécurités, plusieurs organisations consuméristes dans le monde offrent une évaluation prospective, plus poussée, de cette sécurité en effectuant des crash-tests spécifiques et différents des essais réglementaires.

Le programme d'évaluation de sécurité de nouvelles voitures "New Car Assement Program" (NCAP) est un programme d'évaluation de la sécurité qui fournit aux automobilistes une information sur la sécurité potentielle offerte pour une nouvelle voiture. Il s'appuie sur des crash-tests. Le premier NCAP, US-NCAP, est créé en 1979 au United States (US) par "Nationale Highway Trafic Safety Administration (NHTSA)". Le

programme européen, Euro NCAP, est créé en 1996 par le laboratoire de recherche en Transport en Royaume-Uni (UK). Il a été ensuite soutenu par plusieurs gouvernements de l'Union Européenne. D'autres programmes similaires existent dans le monde : le programme australien ANCAP, le programme chinois C-NCAP, le programme japonais JNCAP, le programme de l'Amérique latine Latin NCAP. Chaque programme repose sur ses propres protocoles, qui sont adaptés à la réalité accidentologique du pays.

Le programme européen regroupe 12 membres de gouvernements et associations de consommateurs. Les tests sont effectués dans 8 laboratoires. Les résultats sont présentés sous forme d'un classement sous forme d'étoiles : plus le véhicule obtient d'étoiles, plus celui-ci est déclaré comme sûr. Aujourd'hui, la note maximale est de cinq étoiles. Alors que les voitures deviennent de plus en plus sûres, l'Euro NCAP contraint ses protocoles au fil des années. Nous pouvons distinguer deux phases dans les notations Euro NCAP.

La première phase s'étend de 1997 à 2008. Trois notations sont attribuées individuellement :

- la notation de la protection des adultes a été la seule notation publiée (4 étoiles) jusqu'à fin 2000. Elle a été basée sur deux chocs : le premier est un choc frontal à 64 km/h avec recouvrement de 40% de la face avant du véhicule contre une barrière fixe déformable ; le deuxième est un choc latéral gauche avec un impacteur barrière mobile déformable lancé à 50 km/h. A partir de 2001, la cinquième étoile est introduite, avec l'ajout facultatif du choc latéral poteau à 29 km/h et le témoin de blocage de ceinture dans les essais. La gravité de blessures est mesurée sur deux mannequins placés aux places avant.
- la notation de la protection des piétons est apparue en 2002. Elle est noté sur quatre étoiles. Quatre projections d'impacteurs tête adulte, tête enfant, fémur et jambe à 40 km/h contre l'avant du véhicule (bouclier, capot, pare-brise) sont réalisées pour mesurer la gravité de blessures.
- la notation de la protection des piétons est apparue en 2003. La gravité des blessures est mesurée sur deux mannequins enfants placés aux places arrières en choc frontal.

La deuxième phase a commencé en 2009 avec l'apparition de la notation globale (l'overall rating) : c'est une seule notation obtenue à partir de quatre notations, adultes, piétons, enfants et systèmes d'assistance en utilisant la pondération suivante : 50% pour la protection des adultes, 20% pour la protection des piétons, 20% pour la protection des enfants et 10% pour les systèmes d'assistance. Une nouvelle pondération est appliquée à partir de 2014, la notation adultes a perdu 10% de son poids dans la notation globale en faveur de la notation des systèmes d'assistance. Un véhicule testé peut obtenir cinq étoiles (la note maximale) s'il a eu cinq étoiles dans les quatre notations. L'obligation du choc poteau, le coup de lapin (whiplash) en chocs frontal et arrière, l'AEB city (freinage automatique d'urgence en collision fronto-arrière à une vitesse inférieure à 50 km/h), le choc frontal à 50 km/h avec recouvrement 100% contre un mur rigide sont les principales évolutions dans les protocoles de la catégorie adultes. Les systèmes d'assistance sont valorisés à partir du moment où le montage en série dans le véhicule atteint un niveau minimum. De nombreux systèmes d'assistance sont introduits dans la notation depuis 2009.

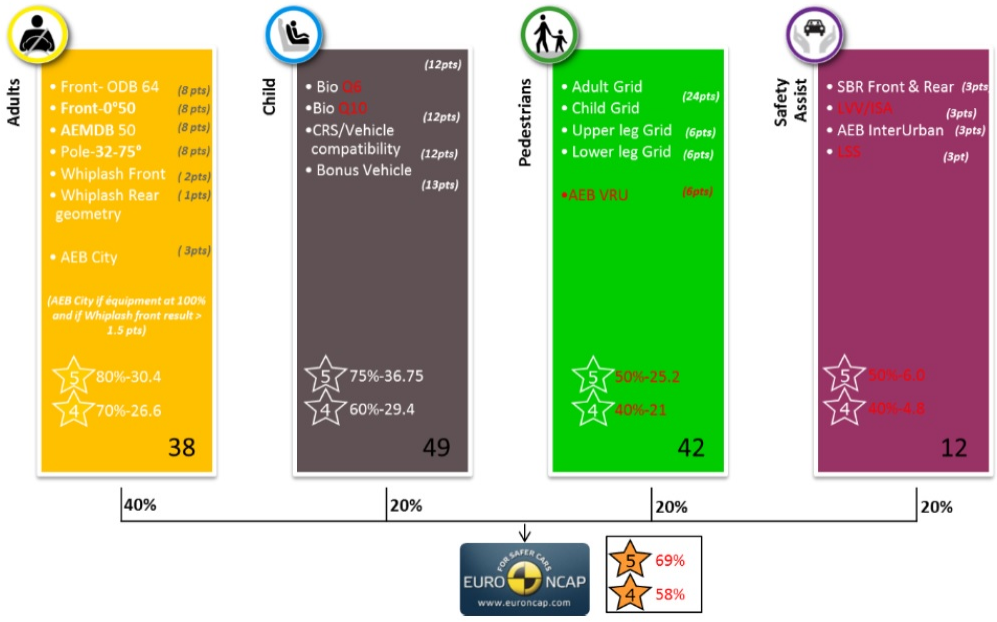

FIGURE I.7 – Répartition des points dans les différentes catégories en 2016

Les répartitions des points dans chaque catégorie et les limites d'étoiles (le nombre de points requis pour décider le nombre d'étoiles) ont été changées à plusieurs reprises. La figure I.7 illustre les répartitions des points et les limites d'étoiles utilisées aujourd'hui.

Les mannequins et les segments corporels mesurés ont également subis des changements afin de rendre la notation plus pertinente.

En 2016, un véhicule testé obtient cinq étoiles s'il a au moins 80% du maximum de points possibles dans la protection des adultes, au moins 75% des points possibles dans la protection des enfants, au moins 50% des points possibles dans la protection des piétons, au moins 50% des points possibles dans l'évaluation des systèmes d'assistance et au moins 79% des points possibles dans l'évaluation globale. A partir de cette année, les constructeurs automobiles peuvent avoir une double notation de leurs véhicules : une première notation sur le véhicule de série et une deuxième notation sur le véhicule de série avec le pack de sécurité en option. Ce dernier doit être disponible dans les 28 pays de l'Union Européenne et doit être composé de 4 systèmes d'assistance (AEB City, AEB VRU, AEB interurban et Speed Assist System).

I.7.2 Etat de l'art de classement rétrospectif en sécurité des modèles automobiles (projet SARAC)

Plusieurs Classements Rétrospectifs (CRs) en sécurité automobile utilisant des données réelles d'accidents existent dans divers pays européens, Australie et États-Unis. En 1994, l'association allemande d'assurances German Insurance Association (GDV) a lancé un groupe de recherche, formé par des experts en accidentologie, venant des associations

gouvernementales, des universités et des constructeurs automobiles, afin de piloter la recherche sur les CRs en sécurité automobile. Cinq workshops ont été organisés par ce groupe entre 1995 et 1998 pour discuter avec tous les experts à travers le monde du meilleur système d'évaluation, de l'influence de la méthode statistique sur l'évaluation du risque, de la meilleure base de données et de la meilleure présentation des résultats.

Cette initiative a incité la Commission Européenne (CE) à lancer, en 1999, le projet de recherche Quality Creteria for the Safety Assessment of Cars Based on real-World Crashes. Ce projet est commun avec le Comité Européen des Assurances (CEA). La Commission consultative pour l'évaluation de la sécurité SAfety Rating Advisory Committee (SARAC) a été fondée par les membres du projet, venant de 10 pays dont l'Europe, les Etats-Unis, l'Australie et le Japon. L'ensemble des organisations membres est présenté dans la figure 1 de [24].

Ce projet de recherche a pour objectifs : *(i)* la description complète de tous les CRs en sécurité automobile qui existent à travers le monde (base de données, critères de classement, facteurs d'ajustement) ; *(ii)* l'analyse comparative de leurs résultats, afin de comprendre les différences et les points communs.

Ces CR en sécurité automobile, étudiés dans ce projet, sont développés par les organisations internationales suivantes (membres du projet SARAC) :

- Road and Transport Laboratory, University d'Oulu, Finlande ;
- Departement of the Environement, Transport and Regions (DETR), Royaume-Uni ;
- Folksam Insurance, Suède ;
- Volkswagen AG (VW), Allemagne
- Monash University, Accident Research Centre (MUARC) Australie ;
- Insurance Institute for Highway Safety (IIHS), Etats-Unis ;
- Highway Loss Data Institute (HLDI), Etats-Unis.

Les différents CRs sont décrits dans [24]. Les auteurs se focalisent sur les différences majeures entre les différents classements.

La première différence majeure entre ces CRs en sécurité automobile est la source de la base de données analysée. La plupart des organisations utilisent les données d'accidents de la route récoltées par la police, soit au niveau national, soit au niveau régional. Seuls Oulu et IIHS analysent les données d'accidents provenant de compagnies d'assurances. L'assureur Folksam et l'université Monash utilisent les données provenant de deux sources d'information, police et compagnies d'assurances.

La deuxième différence majeure est l'échantillon des accidents analysés. Deux organisations (Oulu, MURAC) analysent tous les accidents matériels et corporels. Les autres organisations utilisent uniquement les accidents corporels (au moins un blessé léger parmi les impliqués dans l'accident). Les évaluations d'Oulu, de DETR et de Folksam sont basées sur les accidents à deux véhicules où le véhicule à évaluer est impliqué. Par ailleurs, les évaluations Monash, IIHS, HLDI et VW reposent sur tous types d'accidents.

Tous les classements décrits dans SARAC sont des classements en sécurité secondaire. Cette dernière est souvent évaluée par deux quantités :(i) la résistance aux chocs, c'est-à-dire la protection offerte par un véhicule donné à ses propres occupants ; (ii) l'agressivité, c'est-à-dire l'agression subie par les occupants d'un véhicule opposé au véhicule à évaluer. Seuls MUARC et DETR ont évalué l'agressivité. Les évaluations de IIHS et HLDI couvrent certains aspects de la sécurité primaire.

Souvent, la résistance aux chocs est évaluée par le risque de blessures ou de blessures graves du conducteur du véhicule à noter, quand il est impliqué en accident. Oulu et DETR se limitent à l'évaluation du risque d'être blessé pour le conducteur. Par ailleurs, les autres organisations utilisent le risque d'être blessé grave pour le conducteur. Dans les cas de Folksam et MUARC, le risque des blessures graves est calculé en trois étapes : *(i)* calculer le risque d'être blessé dans un accident ; *(ii)* calculer le risque d'être blessé grave sachant que le conducteur est blessé ; *(iii)* multiplier les deux risque obtenues dans les étapes précédentes. Diverses méthodes ont été utilisées pour calculer ces risques (calcul de ratio, régression logistique, en particulier).

Plusieurs facteurs (âge et sexe du conducteur, type de choc, masse du véhicule, type du véhicule opposé, vitesse d'impact, ...) ont des effets sur la protection offerte. Il est donc nécessaire d'ajuster les évaluations (les risques de blessures) afin de mieux quantifier la protection offerte par le véhicule considéré (le critère de classement). Chaque organisation ajuste son critère de classement selon la disponibilité des facteurs dans la base de données. L'ensemble des facteurs d'ajustement est présenté dans le tableau 3 de [24].

Dans [10], l'étude consiste à comparer les résultats de six CRs en résistances aux chocs, afin de comprendre leurs différences. Cette comparaison nécessite une base de données commune, qui permet d'appliquer les six méthodes d'évaluation de résistance aux chocs. Deux bases de données (Oulu et IIHS) sont choisies pour réaliser cette étude car elles sont les plus adéquates.

L'effet d'ajustement dans chacun des CRs est étudié. Pour chaque critère de classement, le coefficient de corrélation entre le critère de classement non ajusté et le critère de classement ajusté est calculé en utilisant la base de données IIHS. Ainsi, les coefficients de corrélations obtenus (figures 1 à 4 de [10]) montrent une forte corrélation positive entre les deux critères pour chaque système d'évaluation.

Pour comparer les différents résultats de CR, un critère de classement de référence pour la résistance aux chocs est développé. Il est calculé selon la méthode MURAC (le risque d'être blessé ou blessé grave) en s'appuyant sur tous les accidents corporels et matériels. Tous les facteurs d'ajustement disponibles sont ajustés en utilisant une régression logistique.

La comparaison des résultats est effectuée de trois façons différentes :

- Pour chaque critère de classement, sa corrélation avec le critère de référence est calculé. Les résultats obtenus (tableaux 4 et 5 dans [10]) montrent une forte corrélation entre les différents CRs et le CR de référence. Par ailleurs, ces corrélations deviennent moins importantes quand l'évaluation de référence est ajustée à la masse du véhicule (tableaux 6 et 7 dans [10]).

- Les différentes méthodes de classement sont comparées selon leurs capacités à classer les 20 véhicules les plus impliqués en accidents à deux véhicules aux Etats-Unis. La figure 20 dans [10] présente les ordres des différents véhicules en utilisant le risque des blessures graves comme critère de classement. En général, les rangs obtenus par différentes évaluations sont similaires. Certains véhicules sont classés différemment par les différentes méthodes à cause de la nature des accidents utilisés dans l'analyse.
- Les véhicules présents dans chaque base de données analysée par les différents CRs sont classés en trois groupes : "inférieur", "non défini" et "supérieur". Cette classification est basée sur les bornes de l'intervalle de confiance à 95% de l'estimation de la résistance aux chocs, et ces bornes sont comparées aux bornes respectives de la moyenne des résistances aux chocs de tous les véhicules. Quatre évaluations (Folksam, DETR, MUARC et Neswtead) sont utilisées pour faire la classification des véhicules présents dans les bases de données américaines et finlandaises. Les classements obtenus sont comparés au classement de référence. Les pourcentages des véhicules qui sont classés de la même façon par chaque CR et le classement de référence sont présentés dans les tableaux 8 et 9 de [10]. Le classement Folksam est le moins en adéquation avec la classifications de référence dans les deux bases de données. Par ailleurs, les classements Newstead et DETR sont les plus en accord avec le classement de référence respectivement dans les bases de données américaine et finlandaise.

Ensuite, SARAC a étudié la corrélation entre quelques CRs en sécurité automobile en utilisant les données réelles avec le CR basé sur les crash-tests. Trois études de comparaison entre les deux classements en sécurité automobile sont décrites dans [29]. Ces études sont réalisées dans trois pays différents : Australie, Etats-Unis, Royaume-Uni. La comparaison est effectuée entre les deux classements (rétrospectif et prospectif) développés par les organisations suivantes : MURAC et ANCAP en Australie, IIHS et NHTSA aux Etats-Unis et DETR et Euro NCAP au Royaume-Unis. Les études australiennes et américaines analysent la corrélation entre la note de sécurité réelle (évaluation basée sur les accidents réels) et les mesures enregistrées sur les mannequins, ainsi que la corrélation entre la note de sécurité réelle et le score à étoiles. Par ailleurs, l'étude anglaise repose sur le calcul de l'évaluation moyenne de la sécurité des véhicules qui ont le même nombre d'étoiles Euro Ncap. Les résultats des trois études montrent une forte corrélation entre les CRs en sécurité automobile basés sur les données réelles d'accidents et les CPs en sécurité automobile basés sur les crash-tests.

En résumé, ces méthodes des CRs sont orientées vers les classements des véhicules déjà impliqués en accidents. Parce que un nouveau véhicule qui vient de sortir sur le marché n'a pas assez des données accidentologiques, son positionnement dans le classement est difficile.

I.8 Organisation de la suite du manuscrit

La suite de ce manuscrit est composé de trois chapitres et deux annexes. Le premier chapitre est un chapitre d'introduction. Le premier et le deuxième chapitres sont les deux prépublications issues des mes travaux de thèse. Le dernier chapitre est un chapitre de conclusion et perspectives. La première annexe complète le premier chapitre. Les différents bases de données disponibles au LAB sont présentées dans le deuxième annexe.Elles peuvent en effet être intéressantes pour la suite de mes travaux.

I.8.1 Résumé du chapitre II

Le chapitre II est notre première prépublication disponible sur le site d'archive HAL [30]. Elle est soumise depuis Septembre 2015. Une révision est en cours.

Nous commençons le chapitre par une présentation générale du contexte de l'étude, de l'état de l'art des méthodes de classement et d'une présentation de données BAAC*. Parce qu'un ou deux véhicules sont impliqués, et parce que nous adopteons le point de vue individuel des occupants des véhicules, les données viennent par "clusters".

Notons $\mathbb{O}^1, \dots, \mathbb{O}^n$ les n observations d'un jeu de données BAAC*. nous les modélisons comme des variables aléatoires indépendantes, identiquement distribuées selon la loi $\mathbb{P}$. Considérons la $i$ème observation $\mathbb{O}^i$ correspondant au $i$ème accident.

- Si un seul véhicule léger est impliqué dans l'accident, alors $\mathbb{O}^i = \mathbb{O}^i_1$.
- Si deux véhicules sont impliqués dans l'accident, alors $\mathbb{O}^i = (\mathbb{O}^i_1, \mathbb{O}^i_2)$.
- Pour $k = 1, 2$, $\mathbb{O}^i_k$ se décompose en $\mathbb{O}^i_k = (O^i_{k1}, \dots, O^i_{kJ_k})$, où J_k est le nombre d'occupants du $k$ème véhicule.
- Pour $k = 1, 2$ et $j = 1, \dots, J_k$ fixés, la variable $O^i_{kj} = (Y^i_{kj}, Z^i_{kj})$ décrit l'accident du point de vue du $j$ème occupant du $k$ème véhicule. La variable $Y^i_{kj} = (W^i_{kj}, X^i_{kj})$ contient les données contextuelles W^i_{kj} et la CG du véhicule X^i_{kj}. La variable $Z^i_{kj} \in \{0, 1\}$ décrit quant à elle le degré de sévérité des conséquences de l'accident sur l'occupant (indemne ou blessé léger, blessé grave ou tué).

Nous avons déjà dit que nous nous intéressons tout particulièrement au point de vue des occupants des véhicules. Nous pouvons formaliser cette affirmation comme suit : Nous nous intéressons tout particulièrement à la loi commune des O^i_{kj} qui, sous des hypothèses raisonnables, s'écrit

$$P = \sum_{J=1}^{J_{\max}} \mathbb{P}(K = 1, J_1 = J)\mathbb{P}_{K=1,J_1=J} + \sum_{J=1}^{J_{\max}} \mathbb{P}(K = 2, J_1 = J)\mathbb{P}_{K=2,J_1=J},$$

où nous notons $\mathbb{P}_{K=k,J_1=J}$ la loi conditionnelle commune des J composantes de $\mathbb{O}_1 = (O_{11}, \dots, O_{1J})$, l'accident décrit du point de vue du premier véhicule, sachant $K = k$ (k véhicules sont impliqués dans l'accident) et $J_1 = J$ (le premier véhicule transporte J occupants). Nous démontrons le lemme 1 qui me permet, quel que soit le paramètre d'intérêt Ψ, d'apprendre $\Psi(P)$ à partir de l'échantillon $\mathbb{O}^1, \dots, \mathbb{O}^n$ qui est tiré sous $\mathbb{P}$ et non sous P.

Nous élaborons un algorithme de classement fondé sur le principe du "scoring". Il existe déjà dans la littérature de telles procédures de classement, voir par exemple [15], [12],[11]. Cet algorithme de classement se démarque de l'état de l'art à deux titres : d'une part, elle est capable de s'adapter à la dépendance existant entre les composantes de chacune des observations indépendantes et identiquement distribuées ; d'autre part, elle s'appuie sur le principe de la validation croisée pour élaborer un méta-algorithme de classement à partir d'une librairie d'algorithmes de classement dont on ne sait pas à l'avance lequel va s'avérer être le plus performant.

Le principe de "scoring" consiste à :

1. construire une fonction de "scoring" $s : \mathcal{Y} \to [0, 1]$,
2. attribuer un score $s(w, x)$ à toute combinaison (w, x) d'un contexte d'accident w associé à une CG x,
3. décider que la combinaison (w, x) est plus sûre que la combinaison (w', x') si $s(w, x) \leq s(w', x')$.

Notons r_s une telle procédure fondée sur la fonction de score s. La performance statistique de r_s est évaluée en termes de risque de "ranking" $R(r_s) = E_{P^{\otimes 2}}(L(r_s, O, O'))$, pour la perte $L(r_s, O, O')$ caractérisée par

$$L(r_s, O, O') = \mathbf{1}\{(Z - Z')(s(Y) - s(Y')) < 0\})$$

avec $O = (Y, Z)$ et $O' = (Y', Z')$. Il est bien connu [15] que la règle de classement r_π associée à $y \mapsto \pi(y) = P(Z = 1|Y = y)$ est optimale. Précisément, il apparaît que, quelle que soit la règle de "scoring" r_s, on a

$$0 \leq R(r_s) - R(r_\pi) \leq 2E_P(|\pi(Y) - s(Y)|).$$

L'optimalité est en fait même valable sur la classe de toutes les règles de classement, et pas seulement sur celle de classement par "scoring".

Il y a autant de règles de "scoring" que d'estimateurs de la fonction π. Soit $\widehat{\Psi}^1, \ldots, \widehat{\Psi}^K$ K algorithmes d'estimation de π.

Au lieu d'identifier et de sélectionner un unique, meilleur algorithme, nous construisons une combinaison convexe des algorithmes que nous l'appelons un méta-algorithme. Il prend la forme

$$\widehat{\Psi}_n^* = \sum_{k=1}^{K} \alpha_n^k \widehat{\Psi}^k$$

pour un optimal $\alpha_n \in [0, 1]^K$ tel que $\sum_{k=1}^K \alpha_n^k = 1$ déterminé par validation croisée. Cette façon de procéder est connue sous le nom de Super Learning dans la littérature. Elle a été introduite par [38].

Nous démontrons grâce à une inégalité oracle que la règle de classement fondée sur $\widehat{\Psi}_n^*$ est presqu'aussi performante pour établir des classements que la meilleure des règles de classement fondées sur $\widehat{\Psi}^1, \ldots, \widehat{\Psi}^K$. La mesure de performance s'appuie sur le risque R introduit ci-dessus.

Le méta-algorithme de classement r_s, $s = \widehat{\Psi}_n^*(\mathbb{P}_n)$, 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 $\sum_{i=1}^{n} \sum_{k=1}^{K^i} J_k^i = 37\ 721$. L'échantillon de validation correspond aux 15 852 accidents qui ont eu lieu en 2012 entre un ou deux véhicules légers, pour un total de 35 636 impliqués. Les données contextuelles W_{kj}^i regroupent 30 variables. Rappelons que les CGs X_k^i regroupent sept variables.

Dans l'application, nous évaluons la performance de notre méta-algorithme à l'aide de l'aire sous la courbe ROC (AUC). Sa valeur est estimée à 82.8% avec $[82.4\%, 83.2\%]$ comme intervalle de confiance à 95%.

Pour l'illustration, nous choisissons aléatoirement un contexte d'accident du BAAC* 2012 (voir Tableau II.3). Ensuite, nous construisons huit CGs synthétiques à classer dans sept contextes synthétiques d'accidents.

Les sept contextes synthétiques d'accidents (Tableau II.5) sont construits à partir du contexte d'accident choisi (Tableau II.3). Ils sont regroupés dans huit scénarios obtenu en imposant tour à tour les contraintes suivantes : "daylight : yes", daylight : no","urban area : yes", "urban area : no", "driver's age : 20", "driver's age : 50", "under influence of alcohol : yes" et "under influence of alcohol : no". A chaque combinaison de CG et scénario, nous calculons le score en utilisant le méta-algorithme élaboré par Super Learning.

Le résultats obtenus sont reportés dans le Tableau II.6. Dans 18 combinaisons parmi les 21 différentes combinaisons, le score décroit avec la date de conception. Ces résultats sont conformes avec les attentes des experts. Il est également possible de classer les sept contextes par ordre décroissant de danger. Pour chaque CG, les scénario suivant sont classés de plus sûr à moins sûr : "urban area : yes", "driver's age : 20", "driver's age : 50", "urban area : no" (pareil que "under influence : no"), "daylight : no", "daylight : yes" and "under influence : yes".

I.8.2 Résumé du chapitre III

Le chapitre III est ma seconde prépublication, biontôt disponible sur le site d'archive HAL. Il s'agit d'un prolongement des travaux décrits dans le chapitre II, afin de passer d'un classement contextuel à un classement global.

Dans la section III.1, nous commençons par un rappel du contexte de ma thèse et un résumé du chapitre II. La modélisation des données utilisées et leur distribution sont présentées dans la section III.2. La section III.3 décrit l'objectif du chapitre et les différentes étapes de la méthodologie. La procedure d'apprentissage, les illustrations de la méthode et son évaluation sont données dans la section III.4.

Dans ce chapitre, nous utilisons encore le principe "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 nous appuyons sur les données BAAC*. Nous rappelons que les données BAAC* viennent par "clusters". Une description détaillée de la modélisation et de la distribution de données BAAC* est disponible dans le chapitre II

Pour simplifier la présentation, nous procédons comme si nous n'utilisons 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 le lemme 1 dans le section II.4.

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.

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 avons accès aux observations contrefactuelles tirées sous $\mathbb{P}^{\otimes 2}$, notre objectif sera 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 ρ_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 $E_{\mathbb{P}}(Z_x) > E_{\mathbb{P}}(Z_{x'})$.

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 $x \in \mathcal{X}$:

$$\mathbb{Q}(x, W) = E_P(Z|X = x, W), \tag{I.1}$$
$$E_{\mathbb{P}}(Z_x) = E_P[E_P(Z|X = x, W)]. \tag{I.2}$$

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.

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 $\ell^1_{Q,\mu}$:

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

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 μ 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 :

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

où $\widetilde{Q}$ est un estimateur de Q construit dans le chapitre II par Super Learning [38] 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 $[0,1]$.

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, l'estimateur de s_0 est :

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

Nous élaborons 25 modèles de travail. Ils sont décrits en détailles dans la section III.4.

Le meilleur modèle de travail (le modèle de travail qui a le plus petit risque cross-validé) est le modèle logistique qui utilise toutes les composantes de x et les carrées des composantes numériques de x en plus de $\widetilde{s}(x)$.

Pour évaluer la qualité de classement donné par la fonction de scoring S_n, Nous étudions sa corrélation avec une fonction de score déduite de l'évaluation Euro NCAP. Nous distinguons trois protocoles différents dans les notations Euro NCAP en chocs frontal

et latéral. Ils s'étendent sur les trois périodes suivantes : 1996–2001, 2002–2008 et 2009–2014. Pour chacune de période, Nous calculons un coefficient de corrélation de Spearman et le p-value de test de " non corrélation" contre " corrélation positive". Nous obtenons respectivement les résultats suivants : 29% et 0.0409 (1996–2000), 55% et 4×10^{-7} (2001–2008), 44% et 0.00877 (2009–2014). Si le premier p-value n'est pas suffisamment petit pour donner de résultat significatif, les deux autres sont assez petits et montrent que, pour les deux périodes 2002–2008 et 2009–2014, les fonctions des scores sont fortement corrélées positivement.

Afin de valider notre approche statistiquement, Nous définissons une procédure statistique composée de trois étapes : premièrement, nous regroupons les observations d'accidents par contextes similaires ; deuxièmement, nous calculons les scores des CGs sélectionnées dans la première étape ; troisièmement, pour chaque groupe d'observations, nous testons si la distribution conditionnelle de score sachant que l'accident a conduit à des blessures graves est dominée stochastiquement par la distribution conditionnelle de score sachant que l'accident a conduit à des blessures non graves. Les observations du BAAC* 2013 et 2014 sont regroupées dans 32 groupes de contextes similaires. Aucun test de 32 tests effectués ne rejette l'hypothèse nulle. nous jugeons ces résultats statifisantes.

Chapter II

Contextual ranking by passive safety of generational classes of light vehicles

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.

II.1 Introduction

II.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 [14], 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.

II.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 [13]. 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) [18] 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 [13]. 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 [23, 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.

II.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 A.1. 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 alphanumerical 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 [2, 3, 4, 5] 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 II.7 on this matter).

II.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 15, 12, 11, 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 [38, 33].

II.1.5 Organization of the article

Section III.2 presents the BAAC* data set and a model for its distribution. Section III.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 II.4 shows how to infer features of P from observations drawn from $\mathbb{P}$ by weighting. Section II.5 describes the construction of a meta-algorithm for ranking by super learning and provides theoretical background to motivate its use. Section II.6 summarizes the specifics of the application, illustrates properties of the inferred meta-algorithm and how it can be used. Section II.7 is a closing discussion. Finally, an appendix gathers some technical material, including proofs of our main results.

II.2 Data and their distribution

II.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 $1 \leq i \leq n$. If one single vehicle is involved in the accident described by $\mathbb{O}^i$, then $\mathbb{O}^i$ decomposes as $\mathbb{O}^i = (O^i_{11}, \ldots, O^i_{1J^i_1})$, where J^i_1 is the number of occupants of the vehicle. For convenience, we will also use the alternative notation $\mathbb{O}^i = \mathbb{O}^i_1$ when $J^i_1 = 1$. If two vehicles are involved, then $\mathbb{O}^i$ decomposes as $\mathbb{O}^i = (\mathbb{O}^i_1, \mathbb{O}^i_2)$ with $\mathbb{O}^i_1 = (O^i_{11}, \ldots, O^i_{1J^i_1})$ and $\mathbb{O}^i_2 = (O^i_{21}, \ldots, O^i_{2J^i_2})$. Here, J^i_1 and J^i_2 are 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. We formalize this statement in the second next paragraph.

Set $k = 1$ if one single vehicle is involved and $1 \leq k \leq 2$ otherwise, then $1 \leq j \leq J^i_k$. The data-structure O^i_{kj} describes the accident from the point of view of the jth occupant of the vehicle labelled as k. The choice of what we call the first to J^i_kth occupants is also made in such a way that it is uninformative. The content of O^i_{kj} is specified in Section II.2.3. Common to $O^i_{k1}, \ldots, O^i_{kJ^i_k}$ are the number of vehicles involved, K^i, the number of occupants of the vehicle, J^i_k, a missingness indicator $\Delta^i_k \in \{0, 1\}$, and the product $\Delta^i_k X^i_k$, where X^i_k is the GC of the vehicle labelled as k. If $\Delta^i_k = 1$, then $O^i_{k1}, \ldots, O^i_{kJ^i_k}$ all include X^i_k, otherwise X^i_k is missing.

Recall that, by definition, m random variables $U_1, \ldots, U_m$ are exchangeable if the m-tuples $(U_{\sigma(1)}, \ldots, U_{\sigma(m)})$, where $\sigma \in \mathfrak{S}_m$ ranges over the set of permutations of $\{1, \ldots, m\}$, all share the same distribution. This notably implies that $U_1, \ldots, U_m$ follow the same marginal distribution. The uninformativeness of the labelling of the two vehicles, when two vehicles are involved in the accident, and that of the occupant(s) of the vehicle(s) can be formalized as follows: conditionally on K^i,

— if $K^i = 1$ then, conditionally on J^i_1, $O^i_{11}, \ldots, O^i_{1J^i_1}$ are exchangeable;
— if $K^i = 2$, then $\mathbb{O}^i_1$ and $\mathbb{O}^i_2$ are exchangeable; moreover, both conditionally on J^i_k and on (J^i_1, J^i_2), $O^i_{k1}, \ldots, O^i_{kJ^i_k}$ are exchangeable (for both $k = 1, 2$).

Consequently, there exists a finite collection of distributions $\{\widetilde{P}^i_{kj} : 1 \leq k \leq 2, 1 \leq j \leq J_{\max}\}$ such that, conditionally on K^i,

— if $K^i = 1$ then, conditionally on J^i_1, $O^i_{11}, \ldots, O^i_{1J^i_1}$ are identically distributed and drawn from $\widetilde{P}^i_{1,J^i_1}$;
— if $K^i = 2$, then $\mathbb{O}^i_1$ and $\mathbb{O}^i_2$ follow the same distribution; moreover, conditionally on J^i_k, $O^i_{k1}, \ldots, O^i_{kJ^i_k}$ are identically distributed and drawn from $\widetilde{P}^i_{2,J^i_k}$ (for both $k = 1, 2$).

The integer $J_{\max}$ is the maximal number of occupants of a vehicle. If $K^i = 2$, then it also holds that, conditionally on (J^i_1, J^i_2), $O^i_{k1}, \ldots, O^i_{kJ^i_k}$ are identically distributed (for both $k = 1, 2$).

II.2.2 Assumptions

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 Δ_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 O_{k1} given Δ_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 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 $1 \leq j_1, j_2 \leq J_{\max}$.

With **A1**, we neglect the information that (J^i_2, Δ^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, Δ^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 tell 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 $(O_{kj} \setminus (\Delta_k, \Delta_k X_k), X_k)$, the data-structure O_{kj} deprived of Δ_k with X_k substituted for $\Delta_k X_k$, is independent from Δ_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

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

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 II.4. We actually estimate the conditional probabilities in **A4** based on a validation data set, see Section II.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 [21] 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.

II.2.3 Context, generational class, severity

Set $1 \leq i \leq n$, $1 \leq k \leq K^i$ and $1 \leq j \leq J_k^i$. The data-structure O_{kj}^i decomposes as $O_{kj}^i = (Y_{kj}^i, Z_{kj}^i)$, where $Z_{kj}^i \in \{0, 1\}$ indicates the severity of injuries incurred by the jth occupant of vehicle k in accident i and $Y_{kj}^i = (W_{kj}^i, \Delta_k^i, \Delta_k^i X_k^i)$ gathers all the remaining information.

The component Z_{kj}^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_{kj}^i equals zero if the injury is light (occupant hospitalized for less than 24 hours) or the occupant is unharmed.

The 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".

The context W_{kj}^i 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 influence of alcohol (we will simply write under influence in the rest of the article), 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.

II.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

What follows is inspired from [15, 12, 11]. For self-containedness, we give the easy proofs in Section A.1.

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

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

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

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

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

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

This inequality still holds when $r_0(Y, Y')$ is chosen arbitrarily in (II.2) for couples (Y, Y') such that $Q_0(Y) = Q_0(Y')$. Equality (II.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.1. 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 (II.4) can be easily bounded to yield the following stronger version of (II.4):

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

Moreover, the ranking risk of r_s satisfies

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

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

II.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. The easy proof is presented in Section A.2.

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*

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

where $\pi(J_1 \dots J_K)$ 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 $\mathbb{P}_n = n^{-1}\sum_{i=1}^{n} \text{Dirac}(\mathbb{O}^i)$, it appears that

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

is an estimator of $E_P(f_1(O))$ based on the observations $\mathbb{O}^1, \dots, \mathbb{O}^n$ which are independently drawn from $\mathbb{P}$. The rationale of the definition of $\mathcal{W}(f_1)$ is easy to explain in light of (II.7): it is possible to use every O_{kj}^i with an observed GC for the estimation of $E_P(f_1(O))$ based on $\mathbb{P}_n$, provided that we properly balance the contributions of each $\mathbb{O}^i$ depending *(a)* on how many actors contribute their own description of the single accident summarized by $\mathbb{O}^i$, and *(b)* on how likely it is to observe a GC. Note that our unique assumption on how the components O_{kj}^i of $\mathbb{O}^i$ depend on each other is **A1** (**A2** specifies how the components of O_{kj}^i 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 $\mathbb{O}^i = (O_{11}^i, \dots, O_{1J_1^i}^i)$.

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

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

II.5 Building a meta-algorithm for ranking by super learning

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

In Section II.5.1 (and in Section A.3 as well), given a measure μ and a μ-integrable function f, we use the shorthand notation $\mu f = \int f d\mu$ for clarity of exposition.

II.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 37, 41, 8, 9, 20, to cite only a few seminal works, with an emphasis on methods using the cross-validation principle]. Specifically, we choose to rely on the super learning methodology [38, 33].

Let $\mathcal{M}$ be a set of probability distributions on $\mathcal{O}$ such that $P \in \mathcal{M}$, where P is defined in (II.1). Let $\Psi : \mathcal{M} \to \Theta$ be a mapping/parameter from $\mathcal{M}$ to a parameter set Θ. 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 ℓ 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

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

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 ℓ be *symmetric*: for all $\theta \in \Theta$, $(o, o') \in \mathcal{O}^2$, $\ell(\theta)(o, o') = \ell(\theta)(o', o)$.

Let $\widehat{\Psi}_1, \ldots, \widehat{\Psi}_{K_n}$ be K_n algorithms for the estimation of θ_0. For each $1 \leq k \leq K_n$, for each subset $\{\mathbb{O}^i : i \in S\}$ of the complete data set and related empirical measure $\mathbb{P}_n^S = (\mathrm{card}(S))^{-1} \sum_{i\in S} \mathrm{Dirac}(\mathbb{O}^i)$, $\widehat{\Psi}_k[\mathbb{P}_n^S] \in \Theta$ is an estimator of θ_0. We want to determine which of the algorithms better estimates θ_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. For notational simplicity, we choose p so that np be an integer. Then, given B_n, $\mathbb{P}_{n,B_n,0} = (n(1-p))^{-1}\sum_{i=1}^n \mathbf{1}\{B_n(i) = 0\}\mathrm{Dirac}(\mathbb{O}^i)$ and $\mathbb{P}_{n,B_n,1} = (np)^{-1}\sum_{i=1}^n \mathbf{1}\{B_n(i) = 1\}\mathrm{Dirac}(\mathbb{O}^i)$ are, respectively, the training and validation empirical measures.

For each $1 \leq k \leq K_n$, 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

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

The performances of $\widehat{\Psi}_{\widehat{k}_n}$ relative to θ_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 (II.8) is the optimal risk and, for each $1 \leq k \leq K_n$,

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

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

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

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

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

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

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

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

The proof of Proposition 2 essentially relies on [39].

II.5.2 Ranking by super learning

We now turn to the elaboration of a meta-algorithm for ranking. Earlier results can be found for instance in [12] (see the oracle inequality in Corollary 8 for a ranking rule obtained by minimizing an empirical risk over a class of rules) and in [36] (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 Θ as the set of functions mapping $\mathcal{Y}$ to $[0,1]$. The parameter Ψ 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 ℓ characterized over $\Theta \times \mathcal{O}^2$ by

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

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 (II.4), condition (II.8) is met and ℓ, which is symmetric, does

identify θ_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.

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

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

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

It is easy to verify that (II.12) holds with $c_1 = 1$ when ℓ is given by (II.15). Proposition 7 in [12] guarantees that (II.16) implies (II.13).

As underlined in [12], (II.16) is rather weak. When $\alpha = 0$, it actually poses no restriction at all, but the rightmost term in (II.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 $0 < \alpha < 1$, (II.16) holds with $c_2 = 2c_4/(1-\alpha)$ by Corollary 8 in [12]. As α gets closer to one, the rightmost term in (II.14) decreases faster, at the cost of a larger constant c_3.

II.6 Application

II.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 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 A.1. 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.

	daylight		driver's age		under influence	
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 II.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 influence, or not (columns 5-6 of top table).

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 influence, 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 II.1.

II.6.2 Library of algorithms and resulting meta-algorithm

The meta-algorithm built by super learning relies on $K=49$ 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 II.2 lists the different methodologies and how we tune them. The coding is performed in the language `R` [34]. It greatly benefits from packages contributed by the community, first and foremost the `SuperLearner` package [32].

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 III.3, we give the smooth approximation L^β to it characterized by $L^\beta(r_s,O,O')=1-\mathrm{expit}((Z-Z')(s(Y)-s(Y'))/\beta)$, where $\mathrm{expit}(x)=1/(1+\exp(-x))$ (all $x\in\mathbb{R}$) and β is a fine-tune parameter (we set $\beta=1/30$). Moreover, we specify that we want to identify the best convex combination of the $K=49$ base algorithms provided as inputs, not the best single one. This statement is easily clarified using the terms of Section II.5. Denote $\widehat{\psi}_1,\ldots,\widehat{\psi}_K$ the K base algorithms and $\mathcal{A}_{n,K}$ a net over the simplex $\Sigma^K=\{a=(a_1,\ldots,a_K)\in\mathbb{R}_+^K:\sum_{k=1}^K a_k=1\}$ with cardinality $K_n=\mathcal{O}(n^K)$ and such that, for all $a\in\Sigma^K$, there exists $a'\in\mathcal{A}_{n,K}$ with $\|a-a'\|\leq 1/n$.

The K base algorithms give rise to K_n algorithms $\widehat{\Psi}_a = \sum_{k=1}^{K} a_k \widehat{\psi}_k$ ($a \in \mathcal{A}_{n,K}$). Identifying the best convex combination of $\widehat{\psi}_1, \ldots, \widehat{\psi}_K$ amounts to inferring which element of $\{\widehat{\Psi}_a : a \in \mathcal{A}_{n,K}\}$ better estimates Q_0 for the sake of ranking. In practice, there is no need to specify $\mathcal{A}_{n,K}$, the numerical optimization being carried out over Σ^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 $a_n \in \mathcal{A}_{n,K}$ be the vector of weights that characterizes the meta-algorithm resulting from the numerical optimization. Only six of its $K = 49$ 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 II.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 II.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 II.2 presents the empirical ROC curve of our meta-algorithm. Formula (II.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 26, 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).

II.6.3 Illustration

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

methodology (`R` package)	tuning
bagging classification trees (`ipred` [31])	applied to all variables, or only numeric variables, or only factors; with or without random forest variable importance screening
generalized additive models (`gam` [19])	`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` [35])	`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` [16])	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 (`knn` [40])	`k` set to $5, 7, \ldots, 23, 25$; applied to numeric variables only, with stratification by size class
multivariate adaptive polynomial spline regression (`polspline` [22])	applied to all variables, or only numeric variables, or only factors; with or without random forest variable importance screening `screen.randomForest`
neural network (`nnet` [40])	applied to numeric variables only
random forest (`randomForest` [27])	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` [28])	`nu` set to 0.05, 0.01, 0.1, 0.2
tree based ranking (`treeRank` [7])	applied to numeric variables only

Table II.2 – Library of algorithms combined by super learning.

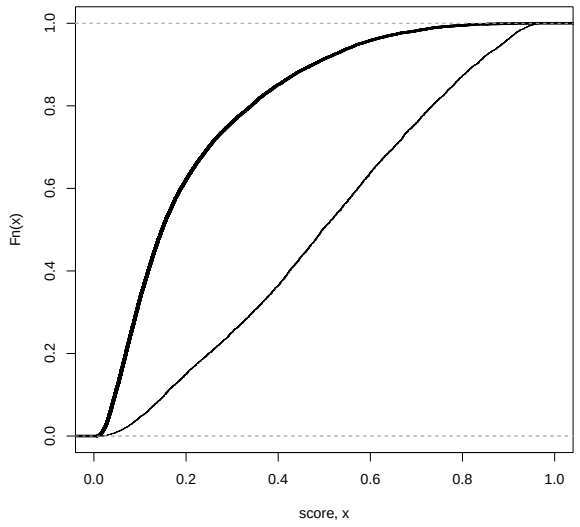

Figure II.1 – Empirical cumulative distribution functions of scores assigned by the meta-algorithm to GCs in some contexts derived from the 2012 BAAC* data set. See the last but one paragraph of Section II.6.2 for details. The top curve corresponds to scores of GCs in contexts of accidents seen from 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.

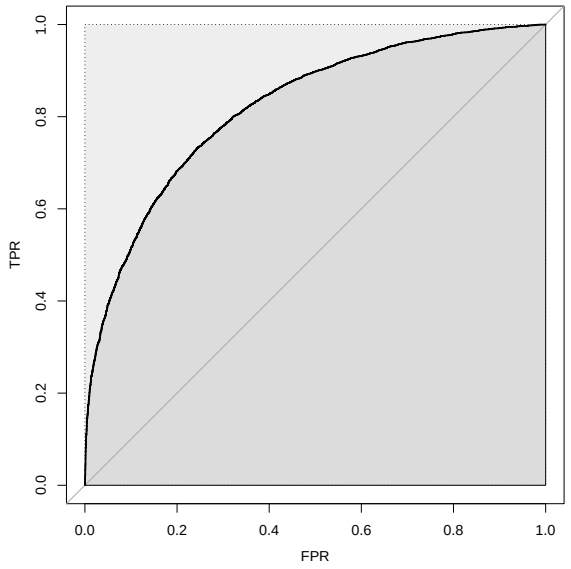

Figure II.2 – Empirical ROC curve of our meta-algorithm. The estimated value of the cross-validated AUC (with $V = 5$ folds) equals 82.8%, with $[82.4\%, 83.2\%]$ as 95%-confidence interval.

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 influence. He owned the vehicle he was driving, and his driving license was valid.
Which occupant	The occupant of interest is the driver himself.

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

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 II.3, see Table II.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 "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 influence, 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 II.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 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).

	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 II.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.

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 influence set to either yes (under influence: yes) or no (under influence: no)

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

It is also known by experts that driving under influence is far more dangerous than driving sober. Inspecting the last two columns of Table II.6 reveals that, in the context described in Table II.3, every GC is assessed safer when driven sober relative to under influence. 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 II.6 reveals that, in the context described in Table II.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 II.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. Although these *contextual* results do not fundamentally contradict the *marginal* results shown in Table II.1, they are somewhat unexpected. 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 II.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.

II.7 Discussion

In this article, we address the contextual ranking by passive safety of GCs of light vehicles by elaborating a meta-algorithm. It is built by combining data-adaptively 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. 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.

The meta-algorithm is contextual (a ranking is conditioned on the occurrence of an

	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 II.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 II.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 II.7 – Same table as Table II.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".

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 [2, 3, 4, 5] estimate under-reporting correction factors that account for unregistered casualties. The same kind

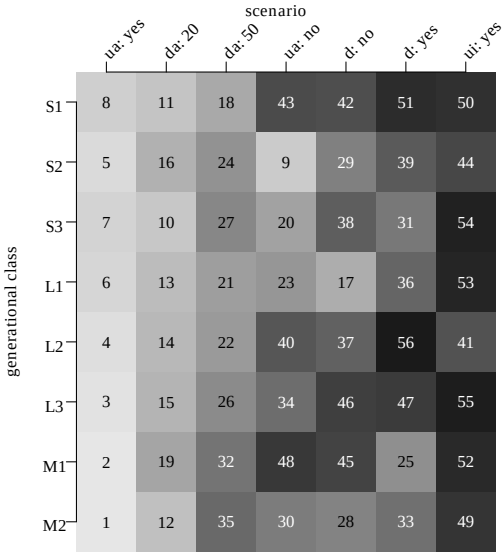

Figure II.3 – Representing the $8 \times 7 = 56$ scores of Tables II.6 and II.7. The smaller is the score, the lighter is the color and the safer is the GC in the given context of accident. The gray level is proportional to the score.

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.

Chapter III

"Contextualized out" ranking by passive safety of generational classes of light vehicles

Abstract

Each year, the BAAC (Bulletin d'Analyse des Accidents Corporels) data set includes traffic accidents on 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 passive safety to their passengers in differents 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 [30], 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 [30], 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 built based on the score function of 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.

III.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 [30]. Our objective here is to integrate out the context of traffic accident from the ranking (or, to paraphrase our title, to "contextualize out" 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).

III.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 [13]. Typically, safety ratings are either predictive or retrospective. Predictive safety ratings assess the safety of vehicles based on crash tests [18]. 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 [23, 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 [30] 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 "contextualise out" the safety ranking in order to provide a global ranking

by (passive) safety.

III.1.2 BAAC* data set

As in [30], 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 [30, 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 [2, 3, 4, 5]. See [30, 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.

III.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 [30]. 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.

III.1.4 Organization of the article

Section III.2 briefly presents the data and their distribution. Section III.3 describes the statistical objective of this study. It lists four main challenges that we face and how

we take them up. Section III.4 summarizes the specifics of the implementation, illustrates the resulting algorithm to rank GCs globally by passive safety, and validates its use. Section III.5 concludes the article with a discussion.

III.2 Data and their distribution

III.2.1 Simplification

We refer the reader to [30, Sections 2.1 and 2.2] 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 [30, Section 4] shows how to carry out estimation as if we observed the individual, smaller data-strucures 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 [30, Lemma 1, Section 4] to use *all* observations.

III.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 [30, Section 2.3], regrouped in six themes: general, when and where, what roadway, what collision, which driver, which occupant.

III.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.

III.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

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

The above RHS expression allows to interpret the LHS one as a contextualized-out (see the outer expectations) ranking error because, whichever is the context of accident W, “$Z_x = 1$ and $Z'_{x'} = 0$” teaches us that GC x' is safer than GC x in context W.

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

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

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, (III.1) yields

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

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

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

(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 $x \in \mathcal{X}$,

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

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, (III.3) straightforwardly implies that

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

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

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

and

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

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

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

(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 $\widetilde{P}_W = \widetilde{n}^{-1} \sum_{i=1}^{\widetilde{n}} \text{Dirac}(\widetilde{W}_i)$,

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

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 (III.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.

III.3.2 Statistical roadmap

Statistical objective (in the real world). The causal argumentation developed in Section III.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 III.3.1.

Let us summarize what is the statistical problem. As stated in Section III.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 $s_0(x) = E_P[Q(x,W)]$ (III.6) where $Q : \mathcal{X} \times \mathcal{W} \to [0,1]$ is charaterized by $Q(X,W) = E_P(Z|X,W)$ (III.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

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

(for a user-supplied measure μ on $\mathcal{X}$; we omit the measurability issues) where the loss function L^1 is given in (III.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 $(x, x') \mapsto \mathbf{21}\{s_0(x) < s_0(x')\} - 1$ 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 RHS integrand in (III.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

$$\ell^1_{Q,\mu}(f, W) \;=\; \int_{\mathcal{X}} \Big[L^1_{Q,x}(f, W) + Q(x,W)\log(Q(x,W))$$

$$+(1-Q(x,W))\log(1-Q(x,W))\Big]d\mu(x)$$

$$= \int_{\mathcal{X}} \Lambda(Q(x,W), f(x))d\mu(x). \tag{III.10}$$

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

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

equals (III.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 (III.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 [30, Section 6.2] by super learning [38, 33] 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 (III.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 irrealistic 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

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

   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 (III.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, $P_{n,B_n,0} = (n(1-p))^{-1}\sum_{i=1}^n \mathbf{1}\{B_n(i) = 0\}\mathrm{Dirac}(O_i)$ and $P_{n,B_n,1} = (np)^{-1}\sum_{i=1}^n \mathbf{1}\{B_n(i) = 1\}\mathrm{Dirac}(O_i)$ 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

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

   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 [30, Section 6.2] is so considerable that it cannot be iterated across the successive folds.

Finally, we estimate s_0 with the score function

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

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

III.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 [30, 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 (III.8). The other one yields the empirical measure P_n which is referred to in Section III.3.2. It is thus used to identify the best working model and to train it as in (III.14).

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

III.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 III.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 III.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 (III.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 (III.14).

III.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 III.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 and date of entry into service increase: 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. 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.

Finally, one should interpret this rankings cautiously. In particular, it is not possible to disentangle the effects of better industrial design, more stringent safety regulations, and wear due to time into service. The construction of the score function S_n (III.14) notably involves an empirical distribution of context of accidents (from the 2012 BAAC* data set). In this light, the score $S_n(x)$ of a GC x which was designed in 1994 (like our synthetic GC M1 in Table III.1) quantifies the global safety of x with respect to a distribution of context which may differ significantly from the distribution of context we would have derived from, say, a 1995 BAAC* data set.

III.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

	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 III.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.

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.

Comparisons between Euro NCAP test results and real-world crash data have already been done [23, 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 III.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. Morever, 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 $\{(S_n(x_j), S_{\text{NCAP}}(x_j)) : j \in \mathcal{J}_k\}$ 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 $\{S_{\text{NCAP}}(x_j) : j \in \mathcal{J}_k\}$ 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 $\{S_{\text{NCAP}}(x_j) : j \in \mathcal{J}_k\}$, 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 $S_{\text{NCAP}}(x) \leq S_{\text{NCAP}}(x')$, 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, 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 in-

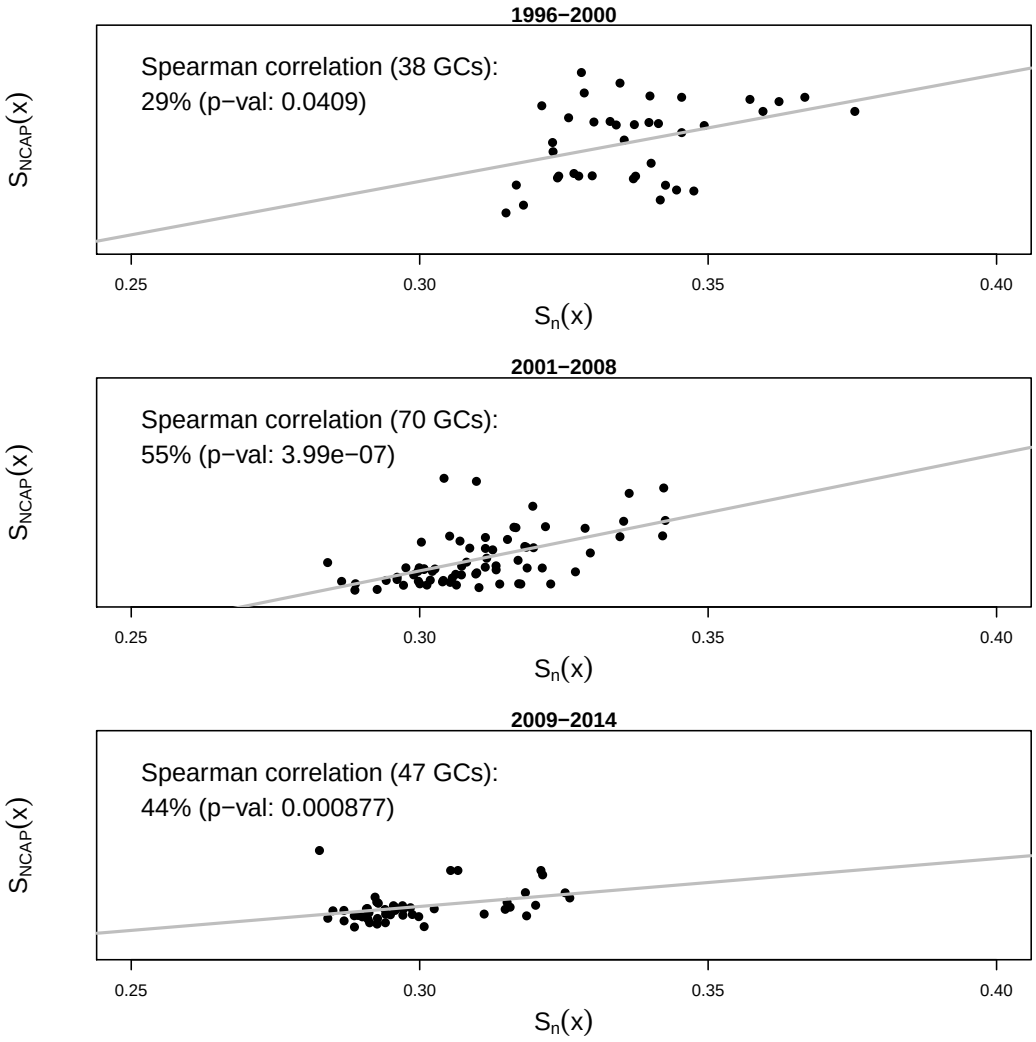

Figure III.1 – Comparing the scores $S_n(x)$ yielded by our method with scores $S_{\mathrm{NCAP}}(x)$ derived from Euro NCAP consumerist ratings for 155 different GCs x. Each plot corresponds to a period during which the Euro NCAP methodology did not change significantly for our purpose. We also report 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.

jury 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 III.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.

1. When $\text{card}(\mathcal{I}_1)$ and $\text{card}(\mathcal{I}_0)$ 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.

III.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 [30] 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 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. 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. 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.

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 [2, 3, 4, 5] 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.

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 III.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.

Chapitre IV

Conclusion et perspectives

IV.1 Conclusion

Les accidents de la route restent une priorité de santé publique aux niveaux mondial, européen et français. La voiture est 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 classer en termes de sécurité offerte les "classes générationnelles" (CG) de véhicules.

Dans la littérature, nous trouvons deux types de classements des modèles automobiles. D'un côté, il y a des méthodes prédictives de classement, qui évaluent la sécurité offerte par les véhicules automobiles en utilisant les données de crash-tests (exemple : Euro-NCAP). De l'autre côté, il y a des méthodes rétrospectives de classement, qui évaluent la sécurité offerte par les véhicules automobiles en utilisant les données d'accidentologie. Notre approche se situe à l'intersection de ces deux principaux courants.

Nous exploitons les données nationales d'accidents corporels de la route appelées BAAC (acronyme de Bulletin d'Analyse des Accidents Corporels). Un bulletin BAAC est rempli par les forces de l'ordre suite à un accident de la route où il y a eu au moins un blessé léger. Nous nous restreignons aux accidents impliquant un seul ou deux véhicules légers. A ce titre, notre approche est rétrospective. Associées à des données de parc, les données BAAC permettent d'identifier la CG de chaque véhicule. Ce sont les CGs que nous classons.

Nous élaborons deux types de classement, un classement contextuel et un classement global. Les deux classements reposent sur le principe de scoring. Pour le classement contex-

tuel, nous cherchons une fonction de scoring qui associe à tout contexte d'accident et toute CG un nombre réel, plus ce nombre est petit plus la CG est jugée sûre dans le contexte d'accident donné. Afin de choisir la meilleur fonction de scoring, nous nous appuyons sur le principe de Super Learner : nous élaborons, à partir de 49 algorithmes, un méta-algorithme qui fait aussi bien que le meilleur que nous ne connaissons pas a priori. Une inégalité oracle justifie théoriquement cette affirmation. Pour le classement global, nous procédons également par scoring : nous cherchons une fonction de scoring 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 (classement contextuel) en s'affranchissant du contexte.

Les deux procédures de classement permettent de classer des CGs existant dans le parc, et également des CGs synthétiques ne correspond à aucun modèle de véhicules dans le parc. A ce titre, notre approche est prédictif.

Les résultats obtenus par les deux procédures de classement sont conformes aux attentes des experts en accidentologie. Nous montrons que notre procédure de classement global est corrélée positivement avec la procédure de classement Euro NCAP. Ce dernier classement est souvent utilisé pour évaluent l'impact des évolutions de conception. Cette corrélation avec notre classement est donc rassurante pour les équipes de conception qui se voient ainsi confirmer qu'un meilleur classement Euro NCAP est positivement corrélé à un meilleur classement fondé sur des données accidentologiques.

IV.2 Discussion et perspectives

Enrichissement des données. La description du contexte de l'accident et la description de la CG ne sont pas figées. Elles peuvent être enrichie si d'autres informations sont disponibles (la vitesse de choc ou les systèmes de sécurité embarqués dans le véhicule) : les procédures de classement (contextuel et global) s'adapteront facilement.

Sécurité active. Dans cette étude, nous nous focalisons sur le classement de CGs en termes sécurité secondaire. Néanmoins, une partie de la sécurité active a été évalué par notre approche. En effet, les CGs associées à des véhicules équipés de systèmes d'assistance à la conduite offrent une meilleure protection (sécurité secondaire). Des nouveaux travaux de recherche vont commencer prochainement pour classer les CGs en termes de sécurité active.

Sous-enregistrement du BAAC. La fonction de scoring, que ce soit dans le classement contextuel ou le classement global, est construite à partir de la loi des accidents enregistrés dans la base de données BAAC, et non pas la loi de la vie réelle de l'accidentologie française. Des études dans la littérature identifient des biais de selection de données BAAC et estiment les nombres de blessés non-enregistrées, en utilisant la méthode capture-recapture. La même méthode est applicable dans le cadre de notre étude, en appliquant les coefficients de correction appropriés.

La réalité accidentologique. L'interprétation des résultats du classement doit être faite avec prudence. En particulier, la distinction entre les effets d'une meilleure conception industrielle et d'une usure due au vieillissement est difficile. La fonction de scoring repose sur la loi empirique du contexte d'accidents (BAAC 2012). Dans ce cadre, le score d'une CG d'un véhicule conçu dans les années 90 quantifie la sécurité globale sachant la distribution du contexte d'accidents de BAAC 2012 qui peut être significativement différente de la distribution de contexte d'accidents de BAAC 1990.

Appendix A

Annexe du Chapitre 2

A.1 ROC curve, AUC, and proofs of results stated in Chapter II Section III.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 [17]. 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 α, *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 β_{Q_0} necessarily dominates β_s [12, 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 (II.3), (II.4), (II.5), (II.6).

Following [12, Example 1], note that

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

$$
\begin{aligned}
&= E_{P^{\otimes 2}}\left(\mathbf{1}\{r(Y,Y')=1\}Q_0(Y)(1-Q_0)(Y')\right.\\
&\qquad \left.+\mathbf{1}\{r(Y,Y')=-1\}(1-Q_0)(Y)Q_0(Y')\right). \quad \text{(A.2)}
\end{aligned}
$$

The above RHS expression is minimized at $r = r_0$, hence (II.4) and the LHS inequality in (II.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 (II.3). It remains to prove the RHS inequality in (II.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 (A.2) implies

$$
\begin{aligned}
&E_{P^{\otimes 2}}\left(L^0(r_s,O,O')\right) - E_{P^{\otimes 2}}\left(L^0(r_0,O,O')\right)\\
&\quad = E_{P^{\otimes 2}}\left((\mathbf{1}\{r_s(Y,Y')=1\} - \mathbf{1}\{r_0(Y,Y')=1\})Q_0(Y)(1-Q_0)(Y')\right.\\
&\qquad\qquad \left.+(\mathbf{1}\{r_s(Y,Y')=-1\} - \mathbf{1}\{r_0(Y,Y')=-1\})(1-Q_0)(Y)Q_0(Y')\right)\\
&\quad = E_{P^{\otimes 2}}\left(\mathbf{1}\{(s(Y)-s(Y'))(Q_0(Y)-Q_0(Y'))<0\}\right.\\
&\qquad\qquad \times\left(\mathbf{1}\{Q_0(Y)\le Q_0(Y')\}(Q_0(Y')-Q_0(Y))\right.\\
&\qquad\qquad\qquad \left.\left.+\mathbf{1}\{Q_0(Y)\ge Q_0(Y')\}(Q_0(Y)-Q_0(Y'))\right)\right)\\
&\quad = E_{P^{\otimes 2}}\left(\mathbf{1}\{(s(Y)-s(Y'))(Q_0(Y)-Q_0(Y'))<0\}|Q_0(Y)-Q_0(Y')|\right). \qquad \text{(A.3)}
\end{aligned}
$$

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

$$
\begin{aligned}
|Q_0(Y)-Q_0(Y')| &= Q_0(Y')-Q_0(Y) = Q_0(Y')-s(Y')+s(Y')-Q_0(Y)\\
&\le Q_0(Y')-s(Y')+s(Y)-Q_0(Y) = |Q_0(Y')-s(Y')+s(Y)-Q_0(Y)|\\
&\le |Q_0(Y')-s(Y')|+|s(Y)-Q_0(Y)|, \quad \text{(A.4)}
\end{aligned}
$$

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

$$
\begin{aligned}
|Q_0(Y)-Q_0(Y')| &= Q_0(Y)-Q_0(Y') = Q_0(Y)-s(Y)+s(Y)-Q_0(Y')\\
&\le Q_0(Y)-s(Y)+s(Y')-Q_0(Y') = |Q_0(Y)-s(Y)+s(Y')-Q_0(Y')|\\
&\le |Q_0(Y)-s(Y)|+|s(Y')-Q_0(Y')|. \quad \text{(A.5)}
\end{aligned}
$$

Combining (A.3), (A.4) and (A.5) 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 (II.5).

We now turn to (II.6). As already mentioned, the inequality $\mathrm{AUC}_s \le \mathrm{AUC}_{Q_0}$ is a direct by-product of [12, 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 [12, Proposition B.2]. Set $p = P(Z = 1)$ hence $p(1-p) = P^{\otimes 2}(Z = 1, Z' = 0)$. Obviously,

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

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

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

hence the equality in (II.6).

A.2 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, Δ_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, (Δ_1, Δ_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:

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

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

This completes the proof.

A.3 Proofs of Proposition 2 and 3

Proof of Proposition 2. We start with the a series of inequalities and equalities. Inequality (A.6) follows from (II.8) and (II.10); it is valid to replace $\widehat{k}_n$ with $\widetilde{k}_n$ as we do in the last RHS term of (A.7) because $\widehat{\mathcal{R}}_n(\widehat{k}_n) \leq \widehat{\mathcal{R}}_n(k)$ for all $1 \leq k \leq K_n$; (A.8) is obtained from (A.7) by rearranging terms:

$$
\begin{aligned}
0 &\leq \widetilde{\mathcal{R}}_n(\widehat{k}_n) - \mathcal{R}(\theta_0) && \text{(A.6)} \\
&= E_{B_n}\left(P^{\otimes 2}\left(\ell(\widehat{\Psi}_{\widehat{k}_n}[\mathbb{P}_{n,B_n,0}]) - \ell(\theta_0)\right)\right) \\
&\quad -(1+\delta) E_{B_n}\left(P_{n,B_n,1}^{\otimes 2}\left(\ell(\widehat{\Psi}_{\widehat{k}_n}[\mathbb{P}_{n,B_n,0}]) - \ell(\theta_0)\right)\right) \\
&\quad +(1+\delta) E_{B_n}\left(P_{n,B_n,1}^{\otimes 2}\left(\ell(\widehat{\Psi}_{\widehat{k}_n}[\mathbb{P}_{n,B_n,0}]) - \ell(\theta_0)\right)\right) \\
&\leq E_{B_n}\left(P^{\otimes 2}\left(\ell(\widehat{\Psi}_{\widehat{k}_n}[\mathbb{P}_{n,B_n,0}]) - \ell(\theta_0)\right)\right) \\
&\quad -(1+\delta) E_{B_n}\left(P_{n,B_n,1}^{\otimes 2}\left(\ell(\widehat{\Psi}_{\widehat{k}_n}[\mathbb{P}_{n,B_n,0}]) - \ell(\theta_0)\right)\right) \\
&\quad +(1+\delta) E_{B_n}\left(P_{n,B_n,1}^{\otimes 2}\left(\ell(\widehat{\Psi}_{\widetilde{k}_n}[\mathbb{P}_{n,B_n,0}]) - \ell(\theta_0)\right)\right) && \text{(A.7)}
\end{aligned}
$$

$$= \quad (1+2\delta)\left(\widetilde{\mathcal{R}}_n(\widetilde{k}_n) - \mathcal{R}(\theta_0)\right) + E_{B_n}(U_{\widehat{k}_n} + V_{\widetilde{k}_n}), \tag{A.8}$$

where we introduce, for each $1 \le k \le K_n$,

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

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 $\max_{1\le k\le K_n} U_k$ and $\max_{1\le k\le K_n} V_k$ given B_n and $\mathbb{P}_{n,B_n,0}$ are both smaller than half the RHS term in (II.14) to derive (II.14) from (A.8).

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 [39]. Let T_k be equal to either U_k or V_k for all $1 \le k \le K_n$. The (conditional) expectation under $\mathbb{P}$ of $\max_{1\le k\le K_n} T_k$ can be written $E_{P^{\otimes 2}}(\max_{1\le k\le K_n} T_k)$. Arbitrarily set $t > 0$, $1 \le k \le K_n$, 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 [6, Theorem 2] yields that

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

But

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

hence

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

Consequently, Lemma 8.1 in [39] implies

$$E_{P^{\otimes 2}}\left(\sqrt{np}\max_{1\le k\le K_n} T_k\right) \le 8\left(\max_{1\le k\le K_n}\left(\frac{v_k}{\lambda_k^\alpha}\right)^{1/(2-\alpha)} + b\right) \times \left(\log\left(1+4K_n\right)\right)^{1/(2-\alpha)}. \tag{A.9}$$

By assumption, $\max_{1\le k\le K_n}(v_k/\lambda_k^\alpha) \le 4(1+\delta)^2 c_2/(\delta\sqrt{np})^\alpha$. Hence, using $2-\alpha \ge 1$, (A.9) yields the following, slightly looser inequality:

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

This completes the proof. □

Proof of Proposition 3. Condition (II.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

$$\begin{aligned}
|\Delta\ell(\theta)(O,O')| &= |(\mathbf{1}\{\theta(Y) \le \theta(Y')\} - \mathbf{1}\{Q_0(Y) \le Q_0(Y')\}) \times \mathbf{1}\{Z = 1, Z' = 0\} \\
&\quad + (\mathbf{1}\{\theta(Y) > \theta(Y')\} - \mathbf{1}\{Q_0(Y) > Q_0(Y')\}) \times \mathbf{1}\{Z = 0, Z' = 1\}| \\
&= |\mathbf{1}\{\theta(Y) \le \theta(Y')\} - \mathbf{1}\{Q_0(Y) \le Q_0(Y')\}| \times \mathbf{1}\{Z \neq Z'\} \\
&= \mathbf{1}\{(\theta(Y) - \theta(Y'))(Q_0(Y) - Q_0(Y')) < 0\} \times \mathbf{1}\{Z \neq Z'\} \\
&\le \mathbf{1}\{(\theta(Y) - \theta(Y'))(Q_0(Y) - Q_0(Y')) < 0\}. \qquad \text{(A.10)}
\end{aligned}$$

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

$$\begin{aligned}
&\mathrm{Var}_P\left(E_{P^{\otimes 2}}\left(\Delta\ell(\theta)(O,O')\middle|O\right)\right) \\
&\quad \le E_P\left(E_{P^{\otimes 2}}(\Delta\ell(\theta)(O,O')\middle|O)^2\right) \qquad \text{(A.11)} \\
&\quad \le E_P\left(E_{P^{\otimes 2}}\left(|\Delta\ell(\theta)(O,O')|\middle|O\right)^2\right) \\
&\quad \le E_P\left(\left[E_{P^{\otimes 2}}\left(\mathbf{1}\{(\theta(Y) - \theta(Y'))(Q_0(Y) - Q_0(Y')) < 0\}\middle|Y\right)\right]^2\right), \qquad \text{(A.12)}
\end{aligned}$$

where the final inequality follows from (A.10) and the independence of Y' with respect to $O = (Y, Z)$. Using the Cauchy-Schwarz inequality now yields

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

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

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

Note that the RHS of the above display can be rewritten $E_{P^{\otimes 2}}\left(\Delta\ell(\theta)(O,O')\right)$ by (A.3). Therefore, we have shown that

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

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

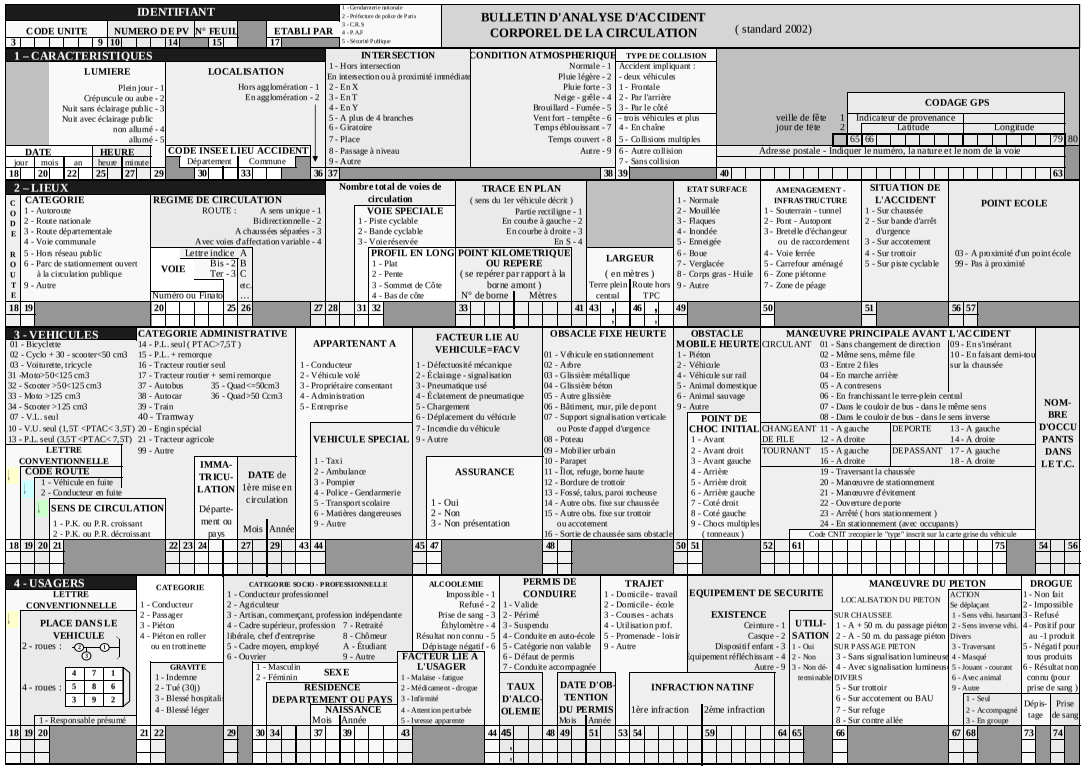

Figure A.1 – Form for the French national file of personal accidents (*Bulletin d'Analyse d'Accident Corporel de la Circulation*, or BAAC form).

j_1	1	2	3	4	5	6	7			
$\widehat{\mathbb{P}}(J_1 = j_1 \vert K = 1)$	40.58	26.75	14.60	10.29	6.41	0.96	0.40			
$\{j_1, j_2\}$	$\{1,1\}$	$\{1,2\}$	$\{1,3\}$	$\{2,2\}$	$\{1,4\}$	$\{2,3\}$	$\{2,4\}$	$\{1,5\}$	$\{3,3\}$	$\{2,5\}$
$\widehat{\mathbb{P}}(\{J_1, J_2\} = \{j_1, j_2\} \vert K = 2)$	53.93	27.02	7.04	3.55	2.51	2.40	1.07	0.83	0.40	0.35
$\{j_1, j_2\}$	$\{3,4\}$	$\{4,5\}$	$\{2,6\}$	$\{1,6\}$	$\{1,7\}$	$\{2,7\}$	$\{3,5\}$	$\{4,4\}$	$\{5,5\}$	$\{5,6\}$
$\widehat{\mathbb{P}}(\{J_1, J_2\} = \{j_1, j_2\} \vert K = 2)$	0.32	0.16	0.08	0.05	0.05	0.05	0.05	0.05	0.03	0.03
$\{j_1, j_2\}$	$\{5,7\}$	$\{3,6\}$	$\{3,7\}$	$\{4,6\}$	$\{4,7\}$	$\{6,6\}$	$\{6,7\}$	$\{7,7\}$		
$\widehat{\mathbb{P}}(\{J_1, J_2\} = \{j_1, j_2\} \vert K = 2)$	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00		

Table A.1 – 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, $\widehat{\mathbb{P}}(\{J_1, J_2\} = \{j_1, j_2\} \vert K = 2)$ is the ratio of number of accidents involving j_1 and j_2 occupants divided by 8,827. Setting $f_1(O) = \mathbf{1}\{\{J_1, J_2\} = \{j_1, j_2\}\}$, $\widehat{\mathbb{P}}(\{J_1, J_2\} = \{j_1, j_2\} \vert K = 2) = E_{\mathbb{P}_n}(\mathcal{W}(f_1)(\mathbb{O}))/\mathbb{P}_n(K = 2)$, 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%.

Annexe B

Bases de données

Bases de données macro d'accidents

Base de données nationales d'accident corporels de la route (BAAC). Ces données sont souvent nommés par données BAAC (Bulletin d'Analyse des Accidents Corporels). Elles répertorient, chaque année, les accidents de la route ayant ont lieu sur les voies publiques français et ayant conduit au moins à un blessé léger. Les bulletins sont établis par les forces de l'ordre. Ils décrivent les conditions générales de l'accident. Ils nous permettent de déterminer :

- quand (date, horaire), où (localisation géographique), dans quelles conditions (nature et état de la route, type de collision) ont eu lieu les accidents ;
- une description du ou des conducteurs impliqués (âge, sexe, catégorie alcoolémie, catégorie socio-professionnelle, validité du permis de conduire) ;
- une description des éventuels passagers (âge, sexe, catégorie socio-professionnelle) ;
- quelles ont été les conséquences de l'accident pour les personnes impliquées (indemne ou blessé léger, blessé grave ou tué).

Le LAB a signé une convention avec ONISR, propriétaire officiel du BAAC, afin de pouvoir y accéder et puisse l'exploiter.

Cette base de données est la source officielle utilisée par les pouvoirs publics pour communiquer sur les chiffres de la réalité accidentologique. Dans [2], [25], les auteurs montrent qu'il existe un écart entre les chiffres de blessés recensés par les forces de l'ordre et la réalité accidentologique. Cet écart a été observé suite à un rapprochement de la base de données BAAC et le registre médicale, qui recense les blessées dans les accidents corporels de la route en département de Rhône. La méthode capture-recapture [2] permet d'estimer le nombre de blessés n'étant enregistrés par aucune des deux sources.

Base de données VOIESUR. Cette base de données prend le nom du projet "Voiture Occupant Infrastructure Etudes de la Sécurité des Usagers de la Route". Les principaux objectifs de ce projet sont de disposer d'un système d'information complet sur les accidents de la route en France et de faire un diagnostic des problèmes de sécurité routière actuels. Ce projet permet de poursuivre le travail fait par le *LAB* en 1990 et 2000. Ce système d'information repose sur l'analyse d'environ 10 000 Procès-Verbaux d'accidents de la route en 2011 : l'exhaustivité des Procès-Verbaux d'accidents mortels en France métropolitaine, l'exhaustivité des Procès-Verbaux d'accidents corporels qui ont eu lieu dans le département du Rhône et 5% des Procès-Verbaux d'accidents corporels, tirés aléatoirement. Cette base de données contient 8 500 accidents. 400 variables présentes dans cette base permettent de réaliser des analyses de sécurité primaire, secondaire et tertiaire.

Bases de données micro d'accidents

Base de données du LAB. C'est une base de données d'accidents de la route orientée vers la sécurité secondaire. Ces études d'accidents de la route sont effectuées soit sur une zone d'enquête représentative "statistiquement" des accidents en France, soit sur le territoire national après une sélection selon des critères spécifiques (véhicules neufs, chocs frontaux, chocs latéraux, accidents impliquant des enfants, ...). Chaque année, 400 nouvelles voitures impliquées en accident sont expertisées par les experts en accidentologie au LAB. Il y a des méthodes de sélection des cas d'accident : une méthode systématique et une méthode ciblée.

La méthode systématique : tous les accidents de la zone Nord-Ouest des Yvelines sont systématiquement étudiés, sans distinction de marques, environ 250 voitures sont étudiées. Les enquêtes sont réalisées avec une identification des véhicules 1 à 2 jours après l'accident. Cette étude d'accident complète les données BAAC avec des données sur la déformation de la voiture et sur le lieu de l'accident (plan, photo, ...).

La méthode ciblée : afin de vérifier la protection offerte par les nouveaux véhicules mis sur le marché, 150 voitures sont analysées en France. Elles sont sélectionnées selon les intérêts des constructeurs automobiles : voitures récentes des 2 groupes, voitures récentes de la concurrence, accidents impliquant des enfants. Les études de cas ciblés se font en différé et l'analyse repose sur la connaissance du choc. Une fiche de synthèse comprenant les circonstances de l'accident, des tableaux de mesures (masse véhicule, enfoncement, intrusion, EES, VR, Dv, ...), les bilans lésionnels des occupants, des commentaires sur l'accident et sur le comportement structurel des véhicules et des photos pertinentes (accompagnées de commentaires), est réalisée à chaque fois.

Les données récoltés par les deux méthodes sont codées, par les enquêteurs, dans une base de données informatique multimédia, qui s'appelle la base LAB.

Base de données de EDA (Études détaillés d'accidents). Ces études détaillées d'accidents sont effectuées sur la scène de l'accident, afin de disposer d'une connaissance

fine des mécanismes accidentelles. Cette connaissance est utilisée pour prédire les besoins en assistance des usagers de la route et pour évaluer l'efficacité attendue (en termes de vies sauvées) des aides à la conduite en développement.

Une équipe spécialisée se rend le plus rapidement possible sur les lieux de l'accident et recherche tous les indices pouvant aider à expliquer ce qu'il s'est passé. Entre 1995 et 2004, tous les accidents qu'ont eu lieu dans la région d'Amiens et à la région d'Evreux sont étudiés. Depuis 2005, l'équipe est installée en région parisienne (Bondoufle, Essonne) où le nombre de véhicules récents est plus important. Les informations à recueillir concernent les impliqués (expérience, actions, ...), les lieux (visibilité, trafic, surface, ...), l'environnement (météo, type de route, luminosité, ...) et les véhicules, par inspection systématique (caractéristiques, défaillances, équipement, ...). Une grande partie des informations est recueillie principalement sur le lieu de l'accident. Les informations complémentaires sont recueillies par des entretiens avec les impliquées à l'hôpital, par inspection des véhicules chez le garagiste et par la consultation de procès-verbal et des bilans médicaux. A partir de toutes ces informations et de leur analyse, une reconstruction de l'accident est réalisée. Une base de données regroupe la majorité de ces informations afin de faciliter le traitement.

Cette base est plus orientée sécurité primaire. Elle contient, à la fin 2015, 1 447 accidents décrit par 1000 variables.

Bases de données étrangères. Il existe d'autres bases de données d'accidents disponibles au LAB. Elles ont plusieurs critères de diversités (origine, équipes d'enquête, la population ciblée, ...) :

- La base de données CCIS (Cooperative Crash Injury Crash) : C'est une base de données britannique d'accidents. Ces données sont récoltées suite à des enquêtes similaires à celles menées par le LAB en France. La population ciblée par l'enquête est les accidents où il y a une voiture particulière (VP) impliqués et de moins de 7 ans, où elle a été remorquée, où l'un de ses occupants a été blessé. 1800 cas par an sont étudiés dans le cadre d'un projet coopératif entre des industriels et le gouvernement britannique. Le LAB dispose des données de phase 2 et 3 du projet. Les données récoltées concernent les impliqués de VP, les causes de blessures (inclus les résultats des autopsies) et les technologies de véhicules.
- La base de données GIDAS (German In depth Investigation Study) : C'est une base d'étude détaillée d'accidents. Elle existe depuis 1995 et couvre 2000 cas par an des régions de Hanovre et de Dresde. La sélection de tous les accidents de la route dans les deux secteurs définis est faite de façon aléatoire. Les données couvrent les impliqués (occupants de tous types de véhicules et piétons), la situation de l'accident, les causes de l'accident, l'origine des blessures, les facteurs humains et les technologies de véhicules.
- Bases de données d'accidents impliquant des enfants : Deux projets européens, CREST et CHILD, sont à l'origine de cette base de données. Le but de ces deux projets est d'évaluer l'efficacité des systèmes de retenue d'enfants (DRE=

Dispositif de Retenue Enfants). Les critères de sélection sont : *(i)* t au moins un enfant de -12 ans impliqué dans l'accident, correctement attaché avec un dispositif de retenue enfant ; *(ii)* au moins un blessé grave dans l'accident, soit l'enfant ou un autre occupant ceinturé ; *(iii)* uniquement des chocs frontaux et latéraux. 669 cas d'accidents, impliquant 1023 enfants ceinturés, sont regroupés dans une base de données communes.

- La base de données internationales : Le projet "Iniative for the Global Harmonisation of Accident Data" (IGLAD) est un projet initié en 2010 par les constructeurs automobiles européens. Le but de ce projet est de mettre en commun des données détaillées d'accidents de la route, provenant de pays différents. La première phase du projet regroupe 1550 accidents qu'ont eu lieu entre 2007 et 2012 dans 10 pays différents (USA, Allemagne, Suède, république tchèque, Inde, Australie, Italie, Autriche, Espagne, France). La deuxième phase de projet regroupe 800 cas d'accidents qu'ont eu lieu en 2012 et 2013 et venant de 8 pays différents (USA, Allemagne, République Tchèque, Chine, Inde, Italie, Autriche, France). Les données françaises sont les données EDA récoltées par les experts d'accidentologie du LAB.

Base de données d'exposition

Base de données de Parc AAA. C'est la base de données des voitures immatriculées en France. Ces données sont gérées par AAA Data, qui une filiale de CCFA (Comité de constructeurs français d'Automobiles) crée en 1959 par les pouvoirs publics sous la forme d'une association 1901. En 2009, les pouvoirs publics ont crée des licences pour l'exploitation statistiques et commerciales de ces données. Ainsi, AAA Data est transformée en société et a acquis ces licences pour exploiter ces données à destination des constructeurs automobiles présents en France et toute entreprise intéressé. Chaque année, le LAB obtient les données des véhicules immatriculées en France pendant les dix dernières années. Cette extraction, achetée par le LAB, contient 22 variables pour décrire les véhicules immatriculées. Ces données sont utilisées en complement des données BAAC pour traiter notre problématique.

Bibliographie

[1] Accidentalité routière 2015 – résultats définitifs. Technical report, Observatoire National Interministérielle de la Sécurité Routière, 2016.

[2] 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`.

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

[4] 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.

[5] 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.

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

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

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

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

[10] M. Cameron, S. Narayan, S. Newstead, T. Ernvall, V. Laine, and K. Langwieder. Comparative analysis of several vehicule safety rating systems. *Proceeding 16th International Technical Conference on the Enhanced Safety of Vehicles*, 2001.

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

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

[13] 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.

[14] 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.

[15] 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.

[16] 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/`.

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

[18] 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.

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

[20] 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.

[21] 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.

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

[23] 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.

[24] K. Langwieder, B. Fildes, T. Ernvall, and M. Cameron. Quality criteria for crashworthiness assessment from real-world crashes. *Proceeding 17th International Technical Conference on the Enhanced Safety of Vehicles*, 2001.

[25] B. Laumon. Insécurité routière : l'apport de l'épidomologie dans le debat public. *Statiqtique et Societé*, 4(1) :21–30, 2016.

[26] 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.

[27] 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/`.

[28] 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.

[29] S. Newstead and M. Cameron. The relatonship between real crach based and barrier test based vehicule safety ratings : A summary and interpretation of three recent studies. *RS 2002 Road Safety Research, Policing and Education Conference Proceedings*, 2002.

[30] 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`.

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

[32] 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.

[33] 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.

[34] 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/`.

[35] 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.

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

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

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

[39] 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.

[40] 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.

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