Mémoire présenté devant l'Institut de Science Financière et d'Assurance pour l'obtention du Diplôme Universitaire d'Actuariat de l'ISFA et l'admission à l'Institut des Actuaires

le 15 Juin 2021

Par : Kevin Ramtohul

Titre : L'apport du "Big Data" dans la tarification de la garantie Responsabilité Civile des flottes ouvertes

Confidentialité : ☐ Non ☒ Oui (Durée : ☐ 1 an ☒ 2 ans)

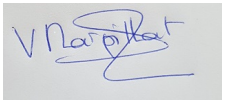
Les signataires s'engagent à respecter la confidentialité ci-dessus

Membres présents du jury de l'Institut des Actuaires :
M. Jérôme SCHAEFFER
M. Alexandre YOU

Membres présents du Jury de l'ISFA :
Mme. Esterina MASIELLO
M. Denys POMMERET

Entreprise :
Nom : AXA France IARD
Signature :

Directeur de Mémoire en entreprise :
Nom : Mme. Véronique MARPILLAT
Signature :

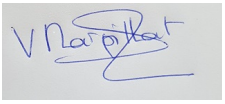


Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels (après expiration de l'éventuel délai de confidentialité)

Secrétariat :

Bibliothèque :

Signature du responsable entreprise



Signature du candidat

Kevin Ramtohul

Résumé

Le marché de l'assurance IARD est soumis depuis quelques années à une très forte concurrence. Cette situation pousse les assureurs à s'adapter en améliorant leur tarif grâce à une segmentation plus fine mais également à une meilleure connaissance du risque. Concernant le marché automobile d'AXA France, on distingue alors deux types de flottes automobiles : les parcs dont la connaissance des véhicules est acquise et les flottes ouvertes pour lesquels nous ne disposions que peu d'informations.

L'avènement du *big data* ainsi que la mise en application du Fichier des Véhicules Assurés (FVA) a permis à la branche automobile entreprise de disposer de nouvelles données permettant de travailler à une meilleure segmentation du tarif de la garantie responsabilité civile des contrats flottes ouvertes. La nouvelle connaissance du numéro d'immatriculation des véhicules composant ce segment, véritable numéro d'identité, a permis l'enrichissement des informations connues sur le contrat, de nouvelles variables tarifaires par le biais de données externes notamment en *open data*.

C'est donc en utilisant la théorie des modèles linéaires généralisés (GLM) que seront proposés plusieurs modèles afin d'améliorer la rentabilité des flottes automobiles par le biais d'innovations tarifaires. Le présent travail aura pour objectif de tester la fiabilité de l'application aux flottes ouvertes du modèle en place sur les parcs. L'étude se concentrera ensuite sur un modèle propre aux flottes ouvertes en approche fréquence $\times$ coût moyen puis en approche prime pure. Enfin, une composante géographique sera intégrée afin d'optimiser le modèle retenu par la mise en place d'un zonier conformément au modèle du parc.

Mots-clés : *Responsabilité Civile*, *Flottes de véhicules*, *Modèles linéaires*, *IARD Entreprise*, *Big Data*, *Open Data*, *Fichier des Véhicules Assurés*, *Sécurité et Réparation Automobile*.

Abstract

Over the last few years, the property and casualty insurance market has been subject to very strong competition. This situation is pushing insurers to adapt their rates by improving them through an even more refined segmentation but also through a better understanding of risk. There are therefore two types of vehicle fleets for AXA France car market : fleets with full knowledge of vehicles (park) and open fleets for which we had sparsely information.

The advent of big data as well as the setting up of the "Fichier des véhicules assurés" (FVA) which list all insured vehicles in France has enabled the automotive business branch to have new data, allowing it to work on better segmentation of the price of civil liability cover for open fleet contracts. The new knowledge of the registration number regarded as a real identity number has enabled the enrichment of known information on the contract with new pricing variables through external data like open data.

By using Generalized Linear Models theory (GLM), several models will be proposed in order to improve the profitability of automobile fleets through pricing innovations. This work will test the reliability of the application of the model used for the parks to open fleets. Then, this research will focus on a specific model for open fleets by using an average cost and frequency approach and then, by using a pure premium approach. Finally, a geographic component will be integrated in order to optimize the model adopted by setting up a risk map district in accordance with the park model.

Keywords : *Civil Liability*, *Vehicle Fleets*, *Linear models*, *Property and Casualty Business Insurance*, *Big Data*, *Open Data*, *FVA*, *SRA*.

Remerciements

Je tiens tout d'abord à remercier tout particulièrement Véronique Marpillat, responsable des équipes Actuariat Produit et Data Science au sein de la Direction Actuariat et Pilotage Entreprise pour sa disponibilité, ses conseils et sa confiance accordée dans l'accomplissement de ce projet personnel et professionnel. Je remercie également tous les membres de l'équipe Actuariat Produit et Data Science pour leur soutien ainsi que leur temps passé à la relecture de ce travail.

J'exprime également toute ma gratitude à l'ensemble du corps professoral de l'Institut de Science Financière et d'Assurances pour la qualité des cours dispensés.

Enfin, je tiens à témoigner toute ma reconnaissance à mes parents pour m'avoir donné les moyens d'accomplir mes projets ainsi que pour leur soutien sans faille. J'adresse également mes plus sincères remerciements à Étienne qui a toujours été présent, même dans les moments de doutes.

Sommaire

Introduction

Avec 220 milliards d'euros de chiffre d'affaire, le marché français de l'assurance est le leader européen devant l'Allemagne et l'Italie. Le secteur assurance de biens et responsabilité, qui regroupe notamment l'assurance automobile, représente plus d'un quart de ces cotisations en affaires directes [1]. Avec l'obligation d'assurance en responsabilité civile, ce sont donc au total, plus de 53 millions de véhicules assurés en France dont la grande majorité concerne les assurances de particuliers. Cependant, la contraction et la forte concurrence de ce marché poussent les assureurs à se tourner vers d'autres segments, notamment du côté des entreprises, qui peuvent amener une certaine rentabilité. C'est le cas en particulier pour le marché des flottes automobiles qui permet à un client de couvrir l'ensemble des véhicules de son entreprise par un unique contrat.

On distingue alors deux grandes catégories de contrats liés aux flottes automobiles différenciées selon leur taille, les flottes fermées et les flottes ouvertes. Si pour les contrats flottes fermées, les informations de chaque véhicule étaient historiquement connues car il s'agit en général de flottes contenant moins de 50 véhicules, ce n'était pas le cas jusqu'à présent pour les contrats flottes ouvertes pour lesquels le nombre de véhicules est beaucoup plus important. En effet, il est plus simple et rapide pour un souscripteur, de gérer une flotte de petite taille et ainsi de connaître l'ensemble des caractéristiques de chaque véhicule que d'avoir à renseigner ne serait-ce que l'immatriculation de tous les véhicules d'une flotte ouverte. Par ailleurs, les systèmes de gestion n'étaient pas non plus en mesure de traiter la quantité de données que pouvait amener une flotte ouverte.

Cependant, dans le cadre du Comité interministériel pour la sécurité routière, le gouvernenement français a décidé la création d'un Fichier des Véhicules Assurés (FVA) afin notamment, de lutter contre la conduite sans assurance en facilitant les contrôles des forces de l'ordre. Depuis le 1^{er} janvier 2019, les compagnies d'assurance ont l'obligation de déclarer l'ensemble des véhicules qu'elles assurent auprès de l'Association pour la Gestion des Informations sur le Risque en Assurance (AGIRA) qui s'occupe de constituer le fichier à destination des institutions. Ce fichier recense notamment les informations relatives aux contrats souscrits par les assurés telles que l'immatriculation du véhicule, le nom de l'assureur, le numéro de contrat ainsi que sa période de validité. En particulier, les compagnies d'assurance doivent connaître la dynamique des contrats qui constituent leurs différents portefeuilles. L'ensemble des flux de véhicules (entrées et sorties) qui composent les contrats doit obligatoirement être renseigné, ce qui implique la connaissance parfaite de l'ensemble des immatriculations liées aux flottes ouvertes. Pour l'heure, ce dispositif ne concerne que les véhicules de moins de 3,5 tonnes. Il sera dans un deuxième temps, étendu aux véhicules de plus de 3,5 tonnes à partir de 2021.

1. Source : Fédération Française de l'Assurance - 2018

Or, la connaissance du numéro d'immatriculation de chaque véhicule est un atout considérable pour le périmètre des flottes ouvertes. En effet, si la tarification pour un contrat de particulier permet de prendre en compte les informations relatives au conducteur telles que son âge ou l'ancienneté de son permis de conduire qui sont des variables fortement discriminantes, ce n'est pas le cas concernant les contrats de flottes ouvertes pour lesquels l'assureur ne dispose d'informations que sur l'état des sinistres, l'usage que fait l'entreprise de ses véhicules ou encore le type principal des véhicules qui composent sa flotte et, désormais, l'immatriculation de chaque véhicule. Il n'existe effectivement aucune obligation légale imposant de désigner un ou des conducteurs attitrés sur le périmètre des flottes automobile. Par nature, les conducteurs sont amenés à changer régulièrement de véhicules dans une flotte. C'est par exemple le cas pour une flotte de voitures de transport avec chauffeur (VTC). L'immatriculation, véritable numéro d'identité du véhicule, permet alors de faire le lien avec de nombreuses données liées aux caractéristiques techniques du véhicule qui peuvent être utilisées dans le cadre de la mise en place d'un modèle de tarification sur le segment des flottes. Par ailleurs, l'absence de désignation d'un conducteur attitré sur ce segment siginifie qu'il ne peut y avoir de système bonus-malus comme cela peut se faire sur les contrats de particuliers. La tarification, aussi bien en affaire nouvelle qu'en renouvellement, est par conséquent, une étape essentielle dans l'atteinte de la rentabilité sur le segment des flottes ouvertes mais également dans le but de proposer à l'assuré un tarif compétitif.

En effet, en tant qu'entreprise, le client cherchera à minimiser les coûts annexes non liés à son développement et en particulier la prime d'assurance couvrant sa flotte de véhicules. La tarification et le renouvellement de son contrat d'assurance seront donc sujets à négociations pour lesquels il n'hésitera pas à faire intervenir la concurrence. Cette étape de la vie du contrat d'assurance flotte ouverte revêt donc une importance particulière à l'heure où la croissance d'AXA France IARD, qui possède presque 30% de parts de marché sur le périmètre des flottes automobiles, est négative. Des moyens innovants sont alors mis en oeuvre permettant de redresser la rentabilité de ce segment. Dans cette attente, une meilleure connaissance du profil des contrats flottes ouvertes sera essentielle et passera assurément par l'acquisition de données complémentaires à l'immatriculation des véhicules.

Par ailleurs, la tarification actuelle sur le segment des flottes ouvertes ne permet pas de distinguer les différentes catégories de véhicules. Aussi bien en affaire nouvelle qu'en renouvellement, les tarifs et majorations sont établis à partir du type majoritaire qui compose la flotte. Cette segmentation ne permet pas de proposer un tarif compétitif au client. L'enjeu de ce travail est donc multiple. Il a pour ambition d'aider au redressement de la rentabilité de la branche automobile mais également de retrouver de la croissance sur ce périmètre par le biais de l'établissement d'une nouvelle tarification bénéficiant d'une meilleure segmentation du risque.

Avant de pouvoir présenter ce nouveau tarif, il convient de présenter en détail les enjeux de cette étude en décrivant le marché de l'assurance de flottes automobiles ainsi que l'apport de la réglementation à ce travail. Dans un deuxième temps et avec le développement du *big data* et de l'*open data*, nous détaillerons les données utilisées et apportées par la connaissance de l'immatriculation. Enfin, le cadre théorique des modélisations sera fixé avant de décrire les différents modèles de tarification testés.

Chapitre 1

Présentation de l'étude

En France, l'assurance automobile est un marché fortement concurrentiel. S'il est nécessaire pour un assureur d'avoir un tarif attractif lors de la mise en place d'une affaire nouvelle, il est capital pour lui d'avoir un tarif précis lors du renouvellement du contrat s'il veut pouvoir défendre son portefeuille face à la concurrence. Afin d'apprécier au mieux les enjeux de ce travail, nous allons présenter dans ce chapitre, le marché de l'assurance des flottes automobiles mais aussi les évolutions réglementaires qui nous ont permises d'affiner la tarification des contrats flottes automobiles en affaire nouvelle et en renouvellement.

1.1 L'assurance des flottes automobiles

1.1.1 Le marché des flottes automobiles

Avec 21,9 milliards d'euros de primes acquises en affaires directes, le marché de l'assurance automobile est le premier marché des assurances de biens et responsabilité [1] et représente un peu plus de 39% des cotisations de ce secteur en 2018. Si l'assurance du particulier représente une part prépondérante du marché de l'assurance automobile, il est désormais difficile de faire progresser ce secteur en terme de rentabilité et de part de marché du fait de la forte concurrence qui y règne. Les assureurs se sont donc tournés vers le marché des flottes automobiles qui représente plus de 10% des cotisations du secteur et qui bénéficie d'une variation positive de +5,1% des cotisations entre 2017 et 2018. En effet, le marché est en croissance régulière depuis 2010 comme le montre la figure 1.1, ce qui laisse quelques perspectives de développement aux assureurs.

1. Ensemble des assurances non vie, à l'exception des assurances maladie et accidents corporels (autres qu'automobile).

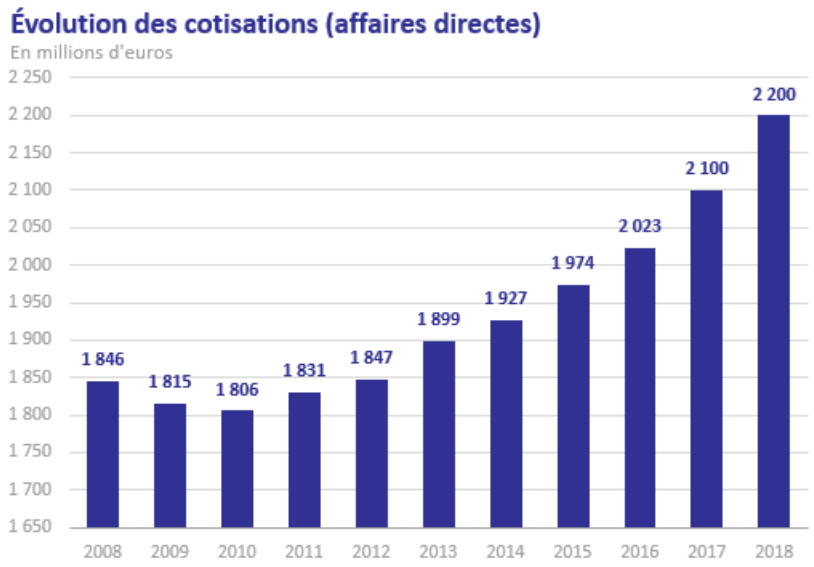


FIGURE 1.1 – Evolution des cotisations en affaires directes des contrats flottes automobiles. Source : *FFA - Données clés 2018*

Il faut donc distinguer au sein du marché de l'automobile en France, l'assurance automobile du particulier et l'assurance automobile des entreprises. Au total, ce sont approximativement 4,1 millions de véhicules assurés au sein d'un contrat flottes automobiles, en hausse de 5% par rapport à 2017 et qui représentent plus de 7% du parc des véhicules assurés en France en 2018 comme le montre la figure 1.2 ci-dessous.

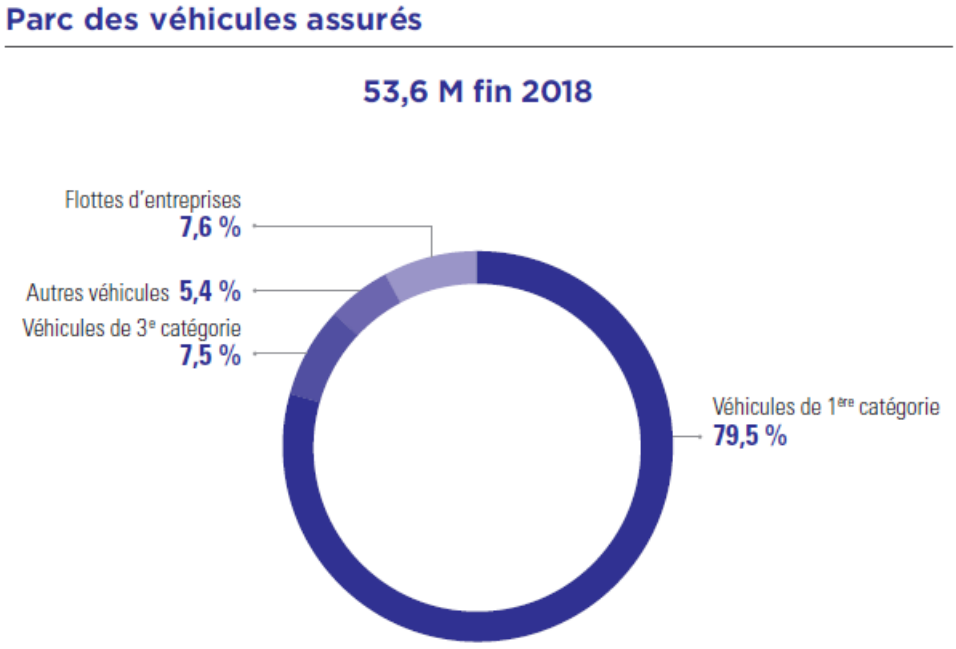


FIGURE 1.2 – Répartition du parc des véhicules assurés en France. Source : *FFA - Données clés 2018*

Chez AXA France, leader du secteur, deux types de contrats liés aux flottes automobiles sont différenciés notamment par la taille de la flotte (plus ou moins de 50 véhicules) :

1. le contrat dit *flotte fermée*, noté PARC dans la suite de ce travail. Il s'agit ici d'un contrat regroupant entre 5 et 50 véhicules, tels que les véhicules d'une petite entreprise de chauffeurs privés. Le nombre de véhicules étant modéré, les informations et caractéristiques propres à chaque véhicule sont connues, renseignées et par conséquent, exploitées. Les contrats flottes fermées bénéficient alors d'une tarification véhicule par véhicule selon les méthodes usuelles ;
2. le contrat dit *flotte ouverte*, noté FLOTTE. Ce type de contrat couvre en général, un regroupement de nombreux véhicules. Il peut s'agir par exemple, des véhicules d'une grande entreprise ou d'entreprises de transports (marchandises ou voyageurs, etc.), ou encore des flottes de collectivités locales. Ces contrats font l'objet d'une tarification différenciée par rapport aux flottes fermées et nous aurons l'occasion de revenir sur ce sujet par la suite. Comme pour les PARCS, ces types de flottes n'ont, en général, pas de conducteurs attitrés. Il ne peut donc y avoir de système de bonus-malus comme pour les voitures de particuliers. En revanche, la connaissance des performances passées est essentielle pour ce type de flotte car il permet d'ajuster le tarif aussi bien en affaire nouvelle qu'en renouvellement. Cette structure est basée sur le calcul d'un tarif par groupe de véhicules composant la flotte. Par ailleurs, le nombre de véhicules composant une flotte ouverte étant important, les caractéristiques propres à chaque véhicules n'étaient pas renseignées - jusqu'à l'application depuis janvier 2019, du *Fichier des Véhicules Assurés*. La mise en place étant particulièrement longue aussi bien au niveau des développements informatiques qu'au niveau du recensement des véhicules, les caractéristiques ne sont, à ce stade, toujours pas connues en totalité et cette étude en bénéficiera pour une part du portefeuille non exhaustive.

On peut noter par ailleurs, la présence de deux autres catégories de contrats au sein de la branche Auto Entreprise AXA :

1. les contrats dits *Mono-véhicules*. On parle alors d'un contrat d'assurance couvrant un unique véhicule majoritairement de plus de 3,5 tonnes (MONO), par exemple, le véhicule d'un artisan ou encore celui d'une petite société. Ces contrats sont alors gérés de la même façon que les contrats d'assurance du particulier. Il existe parfois pour ce type de contrat, un système de bonus-malus et le nombre maximal de véhicules assurés se limite à 4 (Article A 121-1 du Code des Assurances) ;
2. les contrats dits *Garages et Concessions* (GARAGES) qui regroupent les véhicules présents au sein du garage ou de la concession en attente de reprise ou de vente. Il peut également s'agir de véhicules de démonstration.

1.1.2 Les contrats flottes automobiles entreprise

L'assurance des entreprises est donc clairement identifiée comme axe de développement sur lequel les assureurs doivent travailler pour arriver à l'équilibre de rentabilité. Au sein d'AXA Entreprise IARD [1], la branche automobile représente 25% du chiffre d'affaire total à fin 2019. Les deux grandes familles de flottes automobiles ont été présentées et nous allons maintenant décrire les différents types de contrats qui composent ces familles.

1. Incendie, Accidents, Risques Divers

AXA Entreprise IARD [1] analyse son portefeuille automobile de plusieurs manières. En effet, l'analyse des contrats flottes automobiles peut s'effectuer par le mode de gestion. On distingue alors, comme cité précédemment, les flottes fermées (ou à véhicules dénommés) des flottes ouvertes. Dans le premier cas, la composition de la flotte (ou du parc) est connue, ce qui n'est pas le cas pour les flottes ouvertes. Le mix portefeuille peut également s'apprécier par tonnage. En effet, dans l'analyse de la sinistralité par exemple, il est fréquent de distinguer les véhicules de plus de 3,5 tonnes (que nous noterons dans la suite P3T5) des véhicules de moins de 3,5 tonnes (que nous noterons dans la suite M3T5). Le coût des sinistres est généralement plus élevé chez les P3T5. Enfin et surtout, les flottes de véhicules se distinguent également par usage majoritaire :

1. les TPM (Transports Public de Marchandises). Ce sont généralement des véhicules de P3T5. Ces véhicules transportent des marchandises qui n'appartiennent pas à l'assuré. Il peut s'agir notamment d'une entreprise de transport routier. Ce type de flotte est en général peu rentable car les coûts et la fréquence des sinistres sont relativement élevés. En effet, ce sont des véhicules qui circulent quotidiennement sur les routes ;
2. les TPV (Transports Public de Voyageurs) sont composés majoritairement de véhicules de P3T5 mais contiennent une part non négligeable de M3T5. Il peut s'agir ici de transports scolaires, autobus ou encore une flotte de taxis. Ces véhicules circulent également de manière fréquente. Ils cumulent donc un risque de sinistres dommages élevé mais une part plus faible de sinistres de responsabilité. En revanche, ces véhicules transportant des personnes, le coût des sinistres corporels peut être très élevé. On parle alors de sinistres graves voire atypiques ;
3. les TPC (Transports pour Propre Compte) sont principalement des véhicules de P3T5. À la différence des TPM, les véhicules de ce segment transportent les marchandises de l'entreprise assurée. Il est utile de les différencier des TPM car les marchandises transportées peuvent également être assurées par nos soins ;
4. enfin, les AUTRES regroupent les véhicules ne rentrant pas dans les catégories précédentes. Il s'agit principalement de véhicules légers (VL) donc de M3T5 comme notamment les véhicules commerciaux d'une entreprise mais aussi les engins agricoles ou les deux roues.

Si les flottes fermées représentent environ 45% du chiffre d'affaire 2019 de la branche Auto Entreprise d'AXA France comme nous le montre la figure 1.3 ci-dessous, les flottes ouvertes représentent quant à elles presque 35% du chiffre d'affaire qui se répartit majoritairement entre les VL (11,6%) et les TPM (8%).

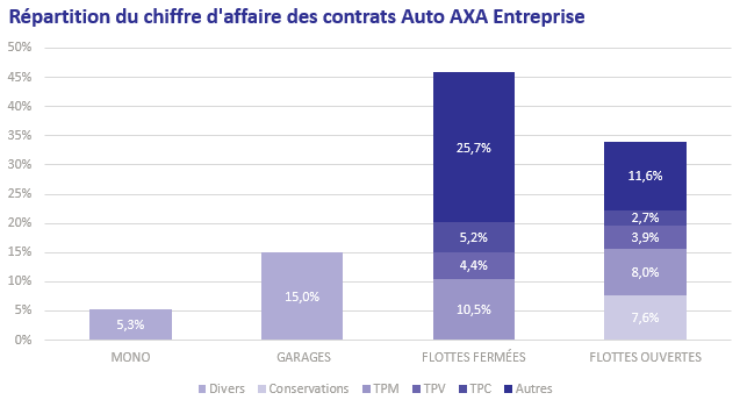


FIGURE 1.3 – Répartition du chiffre d'affaire de la branche automibile entreprise. Source : *AXA - Données à fin 2019*

Nous pouvons noter enfin, un segment qui, au sein d'AXA France, fait l'objet d'un suivi particulier : les conservations, qui représentent 7,6% du chiffre d'affaire Auto Entreprise en 2019. En effet, ces contrats spécifiques sont liés à de grandes flottes ouvertes pour lesquelles une sinistralité récurrente est observée. Une partie de la sinistralité est alors conservée par l'assuré qui bénéficie d'un gain purement fiscal dans le calcul de sa prime commerciale. Pour plus de détails, le lecteur pourra se reporter à Nguyen (2008) (page 8).

Enfin, dans le cadre de ce travail, il est important de distinguer également les flottes de véhicules par leur catégorie :

1. catégorie 1 : Véhicules légers (VL). Comme cité précédemment, il s'agit de véhicules de M3T5 ;
2. catégorie 2 : Camions et remorques P3T5. Il peut s'agir de TPV, TPM ou encore de TPC P3T5. Les remorques de plus de 750 kg doivent bénéficier d'une assurance propre ;
3. catégorie 3 : Bus et cars. Cette catégorie ne concerne que les TPV P3T5 ;
4. catégorie 4 : Engins de chantier. Même si ces engins sont considérés au sein d'une flotte ouverte, ils bénéficient d'une catégorisation particulière au sein d'AXA France ;
5. catégorie 5 : Engins agricoles. De la même façon que les engins de chantier, les engins agricoles sont isolés au sein d'une flotte ;
6. catégorie 6 : Deux roues. Il peut y avoir des 2 roues au sein d'une flotte. Il peut s'agir par exemple de taxis motos ;
7. catégorie 7 : Il s'agit ici des remorques M3T5. Elles sont souvent incluses dans les véhicules de catégorie 1.

En effet, comme l'aborde le paragraphe 1.2, tous les véhicules assurés ne font pas l'objet d'une obligation légale de déclaration au *Fichier des Véhicules Assurés*. Les véhicules agricoles ou les remorques dont le PTAC [1] excède 750 kg par exemple, font l'objet d'une déclaration facultative jusqu'en 2021.

1.1.3 Les garanties d'un contrat flotte automobile

Les différents types de véhicules ayant été décrits, il est utile pour la suite de ce travail, de consacrer ce paragraphe aux définitions des différentes garanties qui peuvent composer une police d'assurance flotte automobile. En effet, si certaines sont obligatoires, d'autres ne le sont pas mais sont souscrites de manière esssentielle en fonction de la typologie de la flotte.

Un contrat d'assurance flotte automobile est composé de deux types de garanties :

1. Une garantie obligatoire : la garantie Responsabilité Civile (RC) constitue la base de tout produit d'assurance automobile en France. Elle est en effet prévue par l'article L211-1 du Code des Assurances. Cette garantie, lorsqu'elle est appelée, permet d'indemniser les victimes des dommages corporels (RCCORP) ou matériels (RCMAT).
2. Une ou plusieurs garanties facultatives :
   (a) les garanties Dommages permettent d'indémniser l'assuré dans le cadre d'un dommage matériel survenu sur un véhicule de sa flotte. Il peut s'agir de garantie dommages accidentels (DOMA), de garantie vol (VOL), de garantie incendie (INC). Les garanties matérielles

1. Poids Total Autorisé en Charge

peuvent également couvrir les dommages survenus en cas d'explosion, d'attentats ou de catastrophes naturelles. Enfin dans certains cas, les effets et objets personnels (EOP) ou les bris de glaces (BDG) peuvent notamment être couverts ;

(b) la garantie Protection Juridique (PJ) permet de défendre les intérêts de l'assuré. On parle de défense lorsque la responsabilité de l'assuré est engagée par le dommage dont il serait l'auteur et de recours lorsque l'assuré est victime d'un dommage causé par un tiers ;

(c) la garantie Sécurité Du Conducteur (SDC) permet d'indemniser le conducteur du véhicule assuré à hauteur de montants plafonnés par le contrat ;

(d) la garantie Assistance (ASS) permet de prendre en charge les frais de remorquage par exemple et qui permet d'organiser l'assistance liée au véhicule endommagé.

Enfin, il est important de noter qu'au sein d'une flotte de véhicules, les différentes catégories de véhicules peuvent bénéficier de garanties optionnelles différentes. En effet, au sein d'une flotte composée principalement de TPM (P3T5) mais comprenant également quelques VL, il peut être nécessaire de n'assurer que les VL contre le BDG par exemple.

1.2 Le fichier des véhicules assurés

Nous allons maintenant nous intéresser au cadre légal qui a amené ce travail. En effet, comme décrit précédemment, les différents véhicules composant les flottes ouvertes au sein d'AXA France n'étaient pas connus jusqu'à la mise en place du Fichier des Véhicules Assurés (FVA). Cela tient principalement au fait que les flottes assurées au sein d'AXA France sont majoritairement de grosses flottes de véhicules et qu'il est compliqué de recenser chaque véhicule d'une flotte. En effet, le nombre de véhicules d'une entreprise peut varier fréquemment et nous n'avons pas, dans la majorité des cas, le même nombre de véhicules lors de la signature du contrat et lors de la tarification par exemple.

1.2.1 La réglementation

En France, 235 personnes ont trouvé la mort en 2016 dans un accident routier impliquant un véhicule non assuré [1]. C'est l'équivalent de 7% de la mortalité routière. Par ailleurs, le Fonds de Garantie des Assurances Obligatoires de dommages (FGAO), qui est notamment en charge d'indemniser les victimes d'accidents de la circulation provoqués par des personnes non assurées ou non identifiées, estime à 750 000 le nombre de véhicules qui circuleraient sans assurance en France. Il s'agit de près de 2% des véhicules en circulation. En avril 2016, le FGAO constatait que le nombre de dossiers de non-assurance qu'il traitait avait augmenté de 40% depuis 2009. En 2018, le fonds a ainsi pris en charge plus de 36 000 victimes dont 72% concernaient la non-assurance comme le montre la figure 1.4 ci-après, et à qui l'organisme a versé 153 millions d'euros.

1. Source : Ministère de l'Intérieur

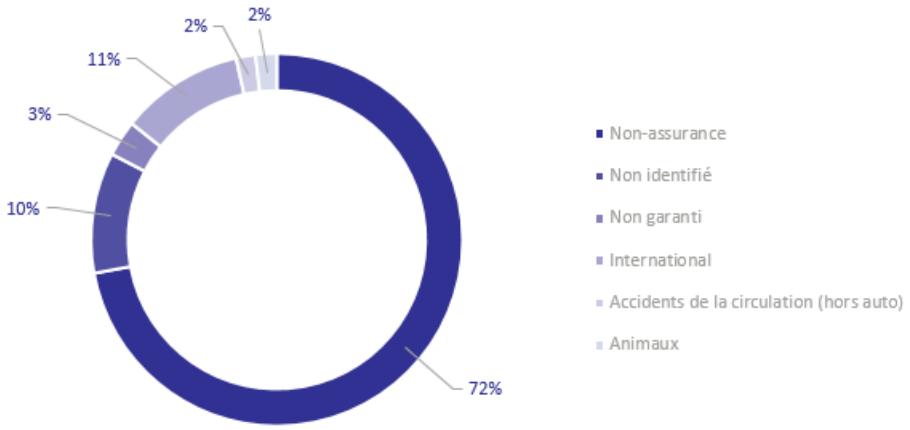


FIGURE 1.4 – Répartition des dossiers traités par le FGAO en 2018. Source : *FGAO - Données à fin 2018*

C'est suite à ce constat et dans le cadre de la Loi de modernisation de la justice du XXI^{e} siècle que la Loi n^{o} 2016-1547 du 18 novembre 2016 [1] prévoit l'obligation pour l'ensemble des assureurs d'alimenter un fichier de tous les véhicules assurés (FVA) pour une mise en application au 1er janvier 2019. Le décret n^{o} 2018-644 du 20 juillet 2018 [2] précise les modalités de constitution et d'alimentation de ce fichier que nous allons détailler par la suite.

La création de ce fichier vise à répondre à plusieurs objectifs :

1. lutter contre la non-assurance, en recensant tous les véhicules immatriculés disposant d'un contrat d'assurance en cours de validité, de manière à fournir les informations nécessaires lors des contrôles des forces de l'ordre ;
2. faciliter les recherches et l'identification des véhicules en cas de délit de fuite, de vol ou de trafic ;
3. remplacer le dispositif actuel de l'Organisme d'Information (OI) pour améliorer la qualité des réponses fournies ;
4. permettre à la profession et aux pouvoirs publics, à partir du fichier des véhicules assurés croisé avec le Système d'Immatriculation des Véhicules [3] (SIV), de fiabiliser et de disposer d'une meilleure connaissance statistique du parc automobile. Ce dernier objectif permettra par conséquent, l'élaboration d'un fichier des véhicules non assurés.

À partir du 1er janvier 2019, tous les véhicules assurés et immatriculés devront donc être connus et renseignés informatiquement par les assureurs et leurs délégataires. En particulier, AXA devra être en mesure de recenser les quelques 1,2 millions de véhicules d'entreprises qu'il assure ainsi que leurs mouvements (entrées, sorties, modifications de parc assuré, etc.). S'agissant des catégories Monos et Parcs, comme nous l'avons décrit précédemment, cela est chose aisée car chaque véhicule est connu. En revanche, concernant les Flottes et les Garages, la connaissance de chaque véhicule n'étant pas acquise, il a fallu développer des outils qui seront détaillés dans la 3ème partie de ce travail.

1. https://www.legifrance.gouv.fr/eli/loi/2016/11/18/JUSX1515639L/jo
2. https://www.legifrance.gouv.fr/eli/decret/2018/7/20/INTS1805978D/jo/texte
3. Fichier des immatriculations de la Préfecture

Ces déclarations devront se faire en quasi-temps réel auprès de l'Association pour la Gestion des Informations sur le Risque en Assurance (AGIRA) qui est en charge de la constitution et de l'exploitation du FVA, puisque le délai d'alimentation retenu par les Pouvoirs Publics est de 72h après la prise d'effet de la garantie et devra suivre le schéma 1.5 présenté ci-dessous.

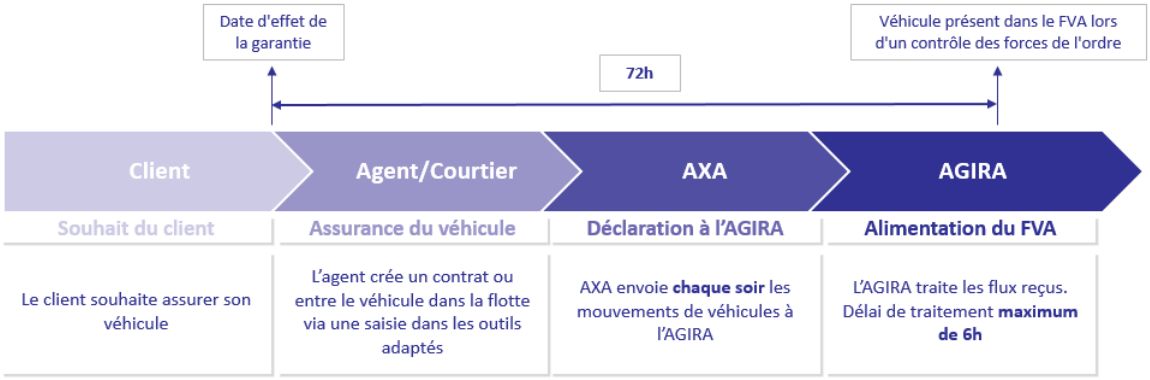


FIGURE 1.5 – Schéma de déclaration des véhicules assurés par AXA à l'AGIRA

1.2.2 Les véhicules concernés par la réglementation

Tous les véhicules assurés et immatriculés devront donc être déclarés à l'AGIRA. Une grande majorité des véhicules composant une flotte ouverte est donc concernée par l'obligation de déclaration au FVA.

Il s'agit alors de tous les véhicules terrestres à moteur (ou assimilés commes les remorques / semi-remorques) circulant sur la voie publique :

- les voitures particulières, les véhicules utilitaires ;
- les deux-roues et scooters, y compris de moins de 50 cm^3, motos, quads, cyclomoteurs, tricycles, quadricycles homologués ;
- les tracteurs et autres engins agricoles ;
- les engins de chantiers à caractère routier (catégorie 1) ;
- les remorques dont le PTAC est supérieur à 500 kg (y compris les remorques agricoles) ;
- les semi-remorques.

Nous pouvons noter ici le cas particulier des remorques, qui disposent d'une immatriculation propre dès lors que leurs PTAC est supérieur à 500 kg. Néanmoins, l'usage pour les contrats d'assurance automobile est que l'assurance du véhicule tracteur couvre sans obligation de déclaration toute remorque tractée dont le PTAC est inférieur à 750 kg. Il n'y a pas de différenciation selon que la remorque soit immatriculée à l'ancien ou au nouveau format. Ainsi, les remorques de PTAC inférieur à 500 kg ne disposent pas d'immatriculation propre et ne sont pas concernées par l'obligation de déclarer au FVA. Les remorques de PTAC supérieur à 500 kg et inférieur à 750 kg, même si elles disposent d'une immatriculation propre (format FNI [1] ou format SIV [2]), peuvent être déclarées au

1. Ancien format d'immatriculation (véhicules immatriculés avant 2009)
2. Nouveau format d'immatriculation en place depuis 2009

FVA. Toutefois, cette déclaration est facultative pendant la période transitoire de la mise en place du FVA, c'est-à-dire, jusqu'en 2021. Il en est de même pour les véhicules agricoles.

Concernant les garages, les immatriculations à déclarer au FVA concernent les véhicules relevant de la flotte classique garage. Il peut s'agir par exemple, de véhicules utilitaires ou encore de dépanneuses. Les véhicules de courtoisie prêtés aux clients par le garage ainsi que les véhicules de dirigeants désignés ou autres (associés ou conjoint) sont également à déclarer. En revanche, même immatriculés, les véhicules confiés par les clients pour entretien et utilisé en essai circulation, les véhicules neufs ainsi que les véhicules en déclaration d'achat (occasion) ne sont pas à déclarer au FVA. On notera le cas particulier des plaques dites *W Garage* dont le garage peut faire la demande et servant dans les cas suivants :

— les véhicules neufs tels que les prototypes à l'essai ou les véhicules en attente de mise en circulation (démonstrations aux clients, présentation à la presse, etc.) ;
— les véhicules d'occasion dont la mise en circulation porte sur des essais techniques liés à une réparation, le transport entre ateliers ou vers un centre de contrôle technique, la revente du véhicule ou encore le remorquage après accident notamment ;
— les véhicules utilisés par les coopératives agricoles et dans les établissements de formation des mécaniciens.

Ces immatriculations "temporaires" doivent également être déclarées aux FVA.

Enfin, AXA ne possédant pas la Libre Prestation de Service (LPS) en branche Responsabilité Civile (RC) Auto, il ne peut assurer les véhicules immatriculés à l'étranger, en dehors du cas où le propriétaire aura réalisé les démarches pour le faire immatriculer en France. Par conséquent, les plaques temporaires d'immatriculation (de type WW) sont également soumises à l'obligation de déclaration au FVA.

1.2.3 Les sanctions en cas de non-assurance

D'une manière générale, la déclaration au FVA s'applique à tout véhicule terrestre à moteur, immatriculé et ayant son stationnement habituel sur le territoire français, et soumis à ce titre à une obligation d'assurance RC. Les autorités ont donc à leur disposition, depuis le 1^{er} janvier 2019, toutes les informations nécessaires à la verbalisation en cas de non-assurance. Les usagers de la route pourront donc être controlés, et le cas échéant, verbalisés lors des contrôles en bord de route réalisés par les forces de l'ordre.

Par ailleurs, il est pertinent de noter que les dispositions du code des assurances relatives à l'attestation d'assurance (R. 211-14 et suivants) deumeurent cependant inchangées : la carte verte en guise d'attestation d'assurance et le certificat d'assurance apposé sur le véhicule de moins de 3,5 tonnes restent obligatoires et font foi en cas de vérification. L'attestation d'assurance continue donc à prévaloir sur le FVA lors des contrôles mais l'absence de présentation de l'attestation d'assurance est par conséquent verbalisable.

Dans un second temps, les usagers de la route roulant sans assurance pourront être verbalisés à la suite d'un délit (excès de vitesse ou non-respect d'un feu rouge par exemple) constaté par

un dispositif de contrôle sanction automatique lorsque l'immatriculation du véhicule concerné par l'infraction ne présente pas de garantie active dans le FVA le jour de l'infraction.

Enfin, la conduite sans assurance est sanctionnée par une amende forfaitaire d'un montant de 750 € minorée à 600 € en cas de paiement dans les 15 jours et majorée à 1500 € en cas de paiement au-delà de 45 jours. Le tableau récapitulatif 1.1 ci-dessous présente la répartition de ces montants dont une partie est allouée au fonds de garantie (FGAO).

TABLE 1.1 – Récapitulatif des amendes en cas de non-assurance

	Minoré Paiement dans les 15 jours [1]	**Normal**	**Majoré** Paiement au-delà de 45 jours [1]
Amende forfaitaire	400 €	500 €	1000 €
+ Contribution obligatoire au fonds de garantie	200 €	250 €	500 €
= Montant à payer	600 €	700 €	1500 €

Le contrevant s'expose par ailleurs à d'autres sanctions, notamment en cas de récidive, comme le précise l'article L324-2 du code de la route :

— 3750 € d'amende ;
— la suspension ou l'annulation du permis de conduire ;
— la saisie du véhicule ;
— une peine de travail d'intérêt général ;
— l'interdiction de conduire certains véhicules terrestres à moteur, y compris les véhicules sans permis, pour une durée de cinq ans au plus ;
— l'obligation d'accomplir, à ses frais, un stage de sensibilisation à la sécurité routière.

1.3 L'apport du FVA à l'étude - Conclusion

Nous l'avons décrit précédemment, plusieurs catégories de véhicules peuvent composer une flotte ouverte chez AXA France IARD. Par ailleurs, l'article L 211-1 du Code des Assurances impose à tout propriétaire de véhicule terrestre à moteur, qu'il l'utilise ou non, de souscrire une assurance minimum obligatoire, dite au tiers, couvrant au moins sa responsabilité civile en cas d'accident. Cette garantie couvre les dommages que le véhicule peut occasionner tels que la blessure d'un passager ou encore les dégâts causés à un autre véhicule par exemple mais une telle garantie ne pourra couvrir le conducteur du véhicule ou le responsable de l'accident pour les dommages qu'ils ont subis. Les produits d'assurance flotte automobile chez AXA France peuvent donc concerner tous types de véhicules qu'ils soient immatriculés ou non, tels que :

— les véhicules légers ou les poids lourds ;
— les engins de chantier à caractères non routiers (catégorie 2) ;
— les matériels de travaux publics et les véhicules et matériels agricoles attachés à une exploitation agricole ou forestière, à une entreprise de travaux agricoles ou à une coopérative

1. Délai prolongé de 15 jours en cas de paiment dématérialisé

d'utilisation de matériel agricole ;
— les engins de déplacement personnel (EDPM), non immatriculés mais circulant sur la voie publique.

Par ailleurs, la mise en place du FVA impose aux assureurs ainsi qu'à leurs délégataires de recenser et de répertorier tous les véhicules disposant d'une plaque d'immatriculation. D'une manière générale, la déclaration au FVA s'applique à tout véhicule terrestre à moteur immatriculé et ayant son stationnement habituel sur le territoire français, et soumis à ce titre à une obligation d'assurance RC. Les remorques de moins de 750 kg et les véhicules agricoles font l'objet d'une déclaration à ce stade et ne deviendra obligatoire qu'à compter du 1^{er} janvier 2021.

Le FVA contraint donc les assureurs à être réactifs dans la saisie des mouvements de véhicules mais également à être plus vigilants dans l'établissement du contrat avec le client, notamment pour qu'il ne soit pas verbalisé sur les garanties fournies à effet rétroactif ou les retards de paiement entraînant des suspsensions de garanties par exemple.

En revanche, la mise en place du FVA est également une opportunité, en particulier pour AXA France, qui aura une parfaite connaissance des risques assurés, notamment sur les flottes ouvertes et ainsi apporter un meilleur service au client en cas de sinistre ou de demande d'assistance. Elle permettra également d'optimiser les processus de gestion, en particulier celui du renouvellement, grâce à la connaissance des flottes assurées et ce, à tout moment. D'un point de vue actuariel, le FVA permettra d'établir une tarification précise des flottes ouvertes en renouvellement et ainsi de se rapprocher de la tarification effectuée sur le PARC.

Cependant, la mise en place du FVA ne concerne, pour l'heure, que les véhicules immatriculés. Même si AXA France a fait le choix de recenser dès à présent, tous les véhicules assurés par ses soins, les véhicules assurés pouvant être déclarés par l'intermédiaire d'un délégataire, tel qu'un courtier qui utiliserait ses propres outils de déclaration, il peut également se poser la question de la qualité du recensement des véhicules dans le FVA et de la complétude de la donnée. C'est pourquoi ce travail se concentrera uniquement sur les données dont AXA France est gestionnaire et en particulier sur le réseau des agents généraux. Nous détaillerons ce point par la suite dans le paragraphe 2.1.3.

Plus généralement, l'utilisation quotidienne du FVA permettra, à terme, de supprimer les cartes vertes pour la circulation des véhicules en France et, concernant AXA France, d'avoir une meilleure connaissance des véhicules qu'elle assure en terme de caractéristiques techniques. En effet, la connaissance de l'immatriculation via le FVA permet de faire le lien avec différentes bases de données existantes et c'est ce que nous allons aborder dans la seconde partie de ce mémoire.

Chapitre 2

Les bases de données

La mise en place du FVA a permis à AXA France ainsi qu'à l'ensemble des assureurs du marché d'obtenir l'immatriculation de chaque véhicule qu'elle assure. La connaissance de l'immatriculation, véritable numéro d'identité du véhicule, permet via des sources de données internes ou externes à l'entreprise, d'intégrer aux analyses de tarification et de rentabilité de nouvelles variables explicatives. Point de départ de tout exercice de tarification, nous allons présenter dans ce chapitre, les différentes bases de données qui ont été utilisées dans la construction de ce travail.

2.1 Les bases internes

Cette partie détaillera les différentes bases issues des systèmes de gestion et de souscription internes d'AXA France et utilisées dans la construction du modèle de données de tarification.

2.1.1 La base contrat AXA

La première étape dans la construction du modèle de données consiste en la récupération du portefeuille d'assurance flottes automobiles ouvertes d'AXA France. Les données relatives au contrat sont saisies par le souscripteur lors de la création ou la modification d'un contrat via les outils de gestion et de souscription (AXAPAC, OSE TPC [1] ou encore OSE FND [2]) et redescendent ainsi dans les systèmes d'informations d'AXA France (Infocentre) dont les données sont extraites mensuellement sous forme de bases au format SAS. En particulier, la base contrat recence l'ensemble des contrats AXA en cours sur l'année étudiée ainsi que les contrats résiliés durant les deux années précédentes. Cette base, par numéro de contrat, permet d'accéder à des informations telles que :

— Les dates d'émission et d'effet du contrat ainsi que l'émission et l'effet de la résiliation éventuellement ;
— Les informations liées au client (numéro de client, nom de l'entreprise, numéro SIRET [3],

1. OSE TPC : Outil de Souscription Entreprise Tout Produit Complet.
2. OSE FND : Outil de Souscription Entreprise Flottes Non Dénommées.
3. SIRET : Système d'Identification du Répertoire des Établissements

activité du client, etc.) ;
— Les différentes primes du contrat (primes émises, primes acquises ou encore la cotisation potentielle annuelle qui permet d'avoir une vision annuelle des primes pour un contrat mensualisé par exemple).

À ce stade et pour des raisons qui seront détaillées dans le paragraphe 2.2.1, il est important de pouvoir inclure les données concernant le distributeur du contrat. En effet, AXA France bénéficie d'un grand réseau d'agents généraux mais travaille également en étroite collaboration avec plusieurs courtiers. Un intermédiaire gère un certain nombre de portefeuilles. Chaque contrat possède donc un numéro de portefeuille qui permet de faire le lien avec les bases distributeurs (agent, courtier) qui recencent donc par numéro de portefeuille, les informations relatives aux distributeurs. Il peut s'agir par exemple :

— Du nom de l'agent ou du courtier ;
— Du numéro d'identification de l'agent ou du numéro ORIAS [1] du courtier ;
— La localisation du distributeur (région, code postal, adresse, etc.).

À noter que les bases contrats sont également archivées par année, ce qui facilite le recensement des contrats flottes ouvertes sur la période étudiée. Une première base de données des contrats flottes ouvertes d'AXA France a donc été construite reprenant l'ensemble des variables présentes dans la table 2.1 ci-dessous :

TABLE 2.1 – Liste des variables issues des bases contrats

Nom variable	**Description**
cnt_numero	Numéro de contrat
cnt_categ	Catégorie du contrat
cnt_sscateg	Sous-catégorie du contrat
cnt_sscateg_plr	Sous-catégorie du contrat en vision PLR [2]
cnt_cdpost	Code postal de souscription du contrat
cnt_fract	Fractionnement du contrat
cnt_moisa	Mois d'anniversaire du contrat
cnt_naf	Code NAF [3] du contrat
cnt_dtfan	Date d'affaire nouvelle du contrat
cnt_dtres	Date de résiliation du contrat
cnt_cot_rc	Cotisation RC du contrat

Ces données sont donc rencensées à la maille contrat. En effet, ces informations sont communes à chacun des véhicules qui composent un contrat flotte entreprise.

1. ORIAS : Organisme pour le Registre unique des Intermédiaires en Asurance, banque et finance. Le code ORIAS est le numéro d'identification unique désignant un intermédiaire.
2. PLR : Permissible Loss Ratio. La définition est donnée dans le chapitre 3.
3. NAF : Nomenclature d'Activité Française.

2.1.2 La base sinistre AXA

Afin de pouvoir appliquer les méthodes actuarielles de tarification, il est nécessaire de pouvoir récupérer les sinistres liés à chaque contrat sur la période étudiée. Ces données permettront notamment de modéliser la fréquence ainsi que le coût moyen des sinistres RC. De même que pour les bases contrats, les données sinistres sont issues des systèmes de gestion des sinistres et redescendues dans des bases SAS mensuellement, également archivées de manière annuelle. Chaque sinistre déclaré possède un numéro de sinistre spécifique qui est lié à un numéro de contrat. Il est donc possible d'agréger par contrat, le nombre et le montant des sinistres survenus sur une année (année de survenance) et lier ces informations à la base contrat par le numéro de contrat. Les bases sinistres contiennent en effet plusieurs informations telles que :

— Le numéro du sinistre ;
— La date d'ouverture ainsi que la date de survenance du sinistre ;
— Le montant du sinistre ;
— La nature du sinistre ;
— L'immatriculation du véhicule sinistré, ce qui permettra de faire le lien avec le recensement des immatriculations par contrat ;
— Éventuellement des informations sur un tiers impliqué dans le sinistre.

Nous pouvons dès à présent noter que la charge globale d'un sinistre peut être répartie entre charge hors graves, charge hors atypique ou encore charge écrêtée. Ce point sera détaillé dans le chapitre 3.

Les variables décrites dans la table 2.2 ci-après ont donc été intégrées dans la base de données du modèle :

TABLE 2.2 – Liste des variables issues des bases sinitres

Nom variable	**Description**
veh_immat	Immatriculation du véhicule sinistré
sin_rc_chg	Charge sinistres RC
sin_rc_chg_gr30	Charge sinistres RC graves
sin_rc_chg_hg30	Charge sinistres RC hors graves
sin_rc_nb	Nombre de sinistres RC
sin_rc_nb_gr30	Nombre de sinistres RC graves
sin_rc_nb_hg30	Nombre de sinistres RC hors graves
sin_vision	Année de survenance du sinistre

La charge ainsi que le nombre des sinistres RC ont été déclinées en charge et nombre RCCORP et RCMAT. La distinction sinistre responsable / non responsable a également été ajoutée pour l'ensemble des sous-catégories. La fusion avec les données contrats et véhicules se fait par immatriculation.

2.1.3 La base véhicule AXA

Comme il a été précisé précédemment, l'innovation apportée par la mise en place du FVA est la connaissance des immatriculations des véhicules qui composent une flotte ouverte. La base véhicule d'AXA France comprenant déjà l'ensemble des immatriculations liées aux contrats Parc et afin de faciliter la saisie en masse des immatriculations des contrats flottes ouvertes, il a été nécessaire de développer un outil de gestion, OSE GdV Flottes & Garages [1], qui a permis de faire le lien avec la base véhicule.

En effet, la saisie des véhicules via cet outil peut se faire en important une liste d'immatriculations fournie par le client ou encore en rapatriant les véhicules d'une entreprise via son numéro SIRET. Cette saisie pouvant également se faire manuellement, nous pouvons constater dans cette base véhicule, quelques différences de format au niveau de l'immatriculation. Il peut arriver par exemple que certains gestionnaires ne saisissent pas les tirets présents dans les immatriculations.

Enfin, ce mode de saisie peut également être source d'erreur (mauvaise immatriculation, mauvaise affectation de la catégorie du véhicule, etc.). Lorsqu'elles sont saisies, les informations redescendent immédiatement dans la base véhicule qui est actualisée mensuellement et les immatriculations sont complétées de différentes données qui font notamment référence :

- Au numéro de contrat associé ;
- À la date d'entrée et éventuellement de sortie du véhicule ;
- À la date de mise en circulation ;
- À différentes informations sur la carrosserie, l'énergie ou encore la puissance fiscale du véhicule.

L'objectif étant de pouvoir associer à chaque ligne de contrat de notre modèle de données, le nombre de véhicules mais aussi leurs caractéristiques, il a été nécessaire d'harmoniser le format de l'immatriculation afin de pouvoir intégrer des données complémentaires au modèle, et notamment externes.

D'autre part, il est utile de préciser ici que les garanties couvertes par le contrat d'assurance ne sont précisées dans cette base que pour les contrats Parcs. En effet, le nombre de véhicule composant un parc étant limité, il est obligatoire pour le gestionnaire de renseigner pour chaque véhicule, les garanties associées. En revanche, le nombre de véhicules composant un contrat Flotte étant important, le gestionnaire ne peut associer à chaque véhicule, toutes les garanties couvrant ce dernier. C'est pourquoi ce travail ne se concentrera pour l'heure, qu'à l'étude de la garantie obligatoire RC. La modélisation des autres garanties pourrait faire l'objet d'une étude particulière ultérieure.

Les immatriculations associées à un contrat donné ont donc été récupérées. La fusion avec les bases contrat et sinistre a pu être réalisée par numéro de contrat. Les variables décrites dans le tableau 2.3 ci-après ont donc été ajoutées à la base de données servant à l'élaboration du modèle :

1. OSE GdV : Outil de Souscription Entreprise Gestion Des Véhicules

TABLE 2.3 – Liste des variables issues de la base véhicule

Nom variable	Description
cnt_numéro	Numéro de contrat
veh_immat	Immatriculation du véhicule
veh_activite	Activité principale du véhicule
veh_carross	Carrosserie du véhicule
veh_cdgenrn	Code genre du véhicule
veh_ctg	Catégorie du véhicule
veh_cylindre	Cylindrée du moteur du véhicule
veh_danger	Top de transport de matières dangereuses
veh_dtent	Date d'entrée du véhicule
veh_dtsort	Date de sortir du véhicule
veh_dtmcirc	Date de mise en circulation du véhicule
veh_libmar	Libellé de la marque du véhicule
veh_modeachat	Mode d'achat du véhicule
veh_modefin	Mode de financement du véhicule
veh_pfisc	Puissance fiscale du véhicule
veh_ptac	Poids Total Autorisé en Charge
veh_valass	Valeur assurée
veh_valven	Garantie valeur vénale
veh_vhtype	Type du véhicule
veh_vhvan	Valeur à neuf du véhicule

2.2 Les bases externes

Une fois l'immatriculation connue, il est facile de compléter le modèle de données d'informations externes. En effet, une multitude de bases de données sont maintenues et mises à disposition parfois gratuitement, notamment par l'État ou encore par des associations.

2.2.1 Le Fichier des Véhicules Assurés (FVA)

Bien que les données soient issues principalement des outils de déclaration au FVA internes d'AXA, certain courtiers ont fait le choix de développer et d'utiliser leurs propres applications déclaratives de véhicules. Il s'agit principalement de grands courtiers dont les contrats flottes recensent plusieurs milliers de véhicules. C'est pourquoi cette base de donnée est considérée comme externe.

En effet, l'obligation de déclaration s'impose à tous les assureurs ainsi qu'à leurs intermédiaires et en particulier, les courtiers. AXA France ainsi que les courtiers avec lesquels elle travaille transmettent donc les flux de véhicules assurés à l'AGIRA qui, dans le but de contrôler les déclarations, a mis en place un système de flux retour mensuel (état de parc des véhicules déclarés et mouvements dans le mois). Ce flux retour permet à AXA France de récupérer l'ensemble des véhicules qu'elle assure soit par ses propres moyens, soit par le biais de ses intérmédaires.

Chez AXA France et concernant le FVA, les intermédiaires peuvent donc être catégorisés en 3 groupes non nécessairement disjoints :

1. Les agents généraux AXA utilisent uniquement les outils de déclaration internes d'AXA France. Les données sont redesendues directement dans la base véhicule décrite précédemment. Ce groupe est indépendant des deux autres groupes.
2. Les courtiers habilités à utiliser les outils déclaratifs d'AXA France. Ces intermédiaires peuvent, s'ils le souhaitent, utiliser les outils AXA afin de déclarer les véhicules de leurs portefeuilles. En revanche, ces courtiers bénéficient également d'habilitations à déclarer directement à l'AGIRA ce qui provoque parfois la présence de doublons dans le flux retour de l'AGIRA.
3. Les courtiers non habilités sont les intermédiaires qui ont fait le choix d'utiliser leurs propres outils déclaratifs et ne passent donc pas via les plateformes AXA. Concernant ces courtiers, la question de la qualité de la donnée peut se poser. En effet, AXA n'a pas le contrôle des déclarations effectuées par ces courtiers et le flux retour de l'AGIRA a mis en évidence des anomalies déclaratives puisque certains véhicules sont attribués à un mauvais numéro de contrat notamment.

En conséquence et étant donné la récente mise en application du FVA, ce travail ne se concentrera que sur les agents généraux ainsi que les courtiers qui ont fait le choix d'utiliser la plateforme déclarative d'AXA et dont les déclarations peuvent être contrôlées. Les déclarations qui n'ont pas été faites par le biais des outils AXA devront faire l'objet de contrôles de la part des courtiers mais également d'AXA France qui reste responsable de tous les véhicules qu'elle assure, même par le biais d'intermédiaires.

Le flux retour de l'AGIRA a donc permis d'intégrer au modèle un contrôle sur les immatriculations déclarées ainsi que sur le numéro de contrat. Le FVA a également permis l'ajout des données concernant la date d'application de la garantie RC (via la date d'entrée et de sortie du véhicule) ou encore des données relatives au modèle du véhicule et extraites du SIV. En effet, lorsque l'information était manquante dans les bases véhicules AXA, il a été possible d'apporter une certaine complétude à nos données via ce flux. Enfin, la présence dans cette base du numéro ORIAS des courtiers permettra à terme de mettre en place un suivi des déclarations des courtiers et ainsi appliquer le modèle décrit dans ce travail.

2.2.2 Le Système d'Immatriculation des Véhicules (SIV)

Remplaçant l'ancien Fichier National des Immatriculations (FNI) depuis avril 2009, le Système d'Immatriculation des Véhicules (SIV) s'inscrit dans le cadre d'une harmonisation européenne concernant la gestion des formalités administratives exigées pour la circulation des véhicules (nouvelles plaques d'immatriculation, carte grise, etc.). Le SIV, géré par l'Agence Nationale des Titres Sécurisés (ANTS) sous l'égide du Ministère de l'Intérieur, repose sur 4 grands principes :

1. Un numéro d'immatriculation est attribué à vie pour chaque nouveau véhicule mis en circulation ;
2. La simplification de la gestion par les services de l'État des pièces administratives pour la circulation des véhicules ;

3. La simplification des démarches administratives pour les propriétaires de véhicules via des procédures télétransmises ;
4. La gestion des habilitations des professionnels du commerce de l'automobile, des huissiers de justice, des experts, des assureurs, des démolisseurs-broyeurs et des sociétés de crédit.

Le SIV référence notamment :

— Les données relatives au titulaire du certificat d'immatriculation du véhicule (nom, prénom, sexe, date et lieu de naissance, raison sociale, numéro SIRET, adresse, etc.) ;
— Les données relatives au véhicule et à l'autorisation de circuler (immatriculation, numéro VIN [1], caractéristiques techniques du véhicule, déclaration d'achat ou de cession du véhicule, date de mise en circulation, etc.) ;
— Les données relatives à l'identification des professionnels habilités à transmettre des données au SIV (nom, prénom, sexe, date et lieu de naissance, raison sociale, numéro SIRET, adresse, type d'habilitation et mode d'accès au SIV, etc.)

À ce titre, le Règlement Général sur la Protection des Données (RGPD) impose que les informations relatives à l'identification du titulaire du certificat d'immatriculation et au véhicule soient conservées 5 ans à compter de la date de la destruction du véhicule. Il en est de même concernant les données relatives aux professionnels habilités dont les données sont conservées 5 ans à compter du retrait ou de la résiliation de l'habilitation. Les données du SIV, qui ne sont donc pas en accès libre, sont mises à disposition d'un utilisateur après octroi par le Ministère de l'Intérieur d'une licence qui vaut agrément et paiement d'une redevance. Ces données peuvent soit être utilisées pour un usage interne dans le cadre d'une activité économique par exemple soit dans le but de vendre une prestation liée à l'exploitation des données.

La connaissance de l'immatriculation a donc permis l'enrichissement de la base de données par les variables décrites dans la table 2.4 suivante :

Table 2.4 – Liste des variables issues du SIV

Nom variable	**Description**
veh_immat	Immatriculation du véhicule
siv_denominationcommerciale	Dénomination commerciale du véhicule
siv_genre	Genre du véhicule
siv_marque	Marque du véhicule
siv_puissanceadministrative	Puissance administrative du véhicule
siv_typeenergie	Type d'énergie du véhicule

Certaines données peuvent paraître redondantes par rapport aux données précédemment décrites mais elles ont permis, soit de compléter les données manquantes, soit de vérifier l'exactitude des renseignements. La qualité de la donnée est, en effet, fondamentale dans une étude de tarification.

1. Numéro VIN : *Vehicle Identification Number*

2.2.3 La base Sécurité et Réparations Automobiles (SRA)

Le SRA est un organisme professionnel qui possède le statut d'association sous la loi 1901 dont toutes les entreprises d'assurance automobiles sont adhérentes. Le SRA a vocation de promouvoir, au sein de la profession et avec tous les acteurs de l'automobile, toutes études ou tous moyens utiles pouvant contribuer à la limitation du nombre et du coût des sinistres dans l'intérêt des assurés. L'une des principales missions du SRA est de tenir à jour un fichier qui recense toutes les caractéristiques techniques et commerciales des véhicules. Ces données ne concernent que les véhicules à moteur de M3T5. Il s'agit des 2 roues tels que les motos, les 3 roues tels que certains types de scooters et enfin, les 4 roues (voiture, camionnettes, etc.). Un code d'identification SRA est affecté à tous les véhicules importés ou fabriqués en France. Ce code SRA est également mentionné sur la carte grise des véhicules.

Outre le maintien du fichier de données, le but de l'association est avant tout de maitriser le mieux possible le coût des indemnisations des sinistres. En effet, via ses études, le SRA est en mesure de définir les véhicules les plus sûrs, les options à privilégier ou encore les modèles les plus sécurisés. Chaque trimestre, les bulletins publiés par le SRA indiquent l'évolution des coûts de réparation et des pièces détachées. Les entreprises adhérentes peuvent ainsi librement utiliser les données techniques et commerciales des véhicules et ainsi les lier aux données concernant le conducteur et l'usage. Il a donc été possible de lier à un certain nombre de véhicule de catégorie 1 à son code SRA (ex GTA - Groupement Technique d'Assurance) et ainsi récupérer ses caractéristiques techniques. Concernant les véhicules dont le code SRA n'était pas renseigné dans les bases à notre disposition, il a été nécessaire d'implémenter un algorithme sous python permettant la récupération du code par l'immatriculation via le site internet de souscription d'AXA.

Les variables décrites dans la table 2.5 ci-dessous ont donc été intégrées à la base de données utilisée pour le modèle :

Table 2.5 – Liste des variables issues du SRA

Nom variable	Description
codauto	Code SRA
gta_carros	Carrosserie SRA
gta_classe	Classe SRA
gta_classeprix	Classe de prix SRA
gta_classerepar	Classe de réparation SRA
gta_emissionco2	Emission de CO_2 du véhicule
gta_energie	Type d'énergie du véhicule
gta_genre	Genre SRA du véhicule
gta_groupe	Groupe SRA
gta_libmar	Libellé SRA de la marque du véhicule
gta_modele	Modèle SRA
gta_nomcom	Nom commercial du véhicule
gta_segment	Segment SRA
gta_vitmaxi	Vitesse maximale du véhicule
gta_poidsvd	Poids à vide du véhicule
gta_puisscee	Puissance réelle du véhicule
gta_nbplace	Nombre de place du véhicule

L'ensemble des données recensées ne feront pas toutes partie du modèle final. Néanmoins, il a été nécessaire d'étudier l'information qu'apportait chacune de ces données afin de ne sélectionner que les plus pertinentes.

2.3 Le modèle final - Conclusion

La connaissance de l'immatriculation est le point de départ de la construction du modèle de données. En effet, le numéro d'identité du véhicule a permis d'intégrer de nombreuses informations et notamment les caractéristiques techniques des véhicules. Il a également permis à AXA France de connaître en détail les véhicules qui composent chaque flotte ouverte. Bien que présentes sur un nombre limité de véhicules du au fait que toutes les catégories de véhicule ne sont pas représentées, les données référencées dans la table 2.6 ci-après permettront d'affiner la tarification des flottes ouvertes chez AXA France.

Par ailleurs, l'intégration de données externes complémentaires permettront d'anticiper certaines évolutions réglementaires. En effet, même si l'Assemblée Nationale a voté l'interdiction de la vente de voitures thermiques à partir de 2040, la Convention Citoyenne pour le Climat prévoit de faire évoluer rapidement le parc automobile français et propose d'interdire dès 2025, la commercialisation de véhicules neufs très émetteurs de dioxyde de carbone (plus de 110 g de CO_2/km). Cette donnée, présente dans les bases SRA permettra de mettre en place des contrôles mais aussi une tarification adapatée en fonction des émissions de CO_2.

Enfin, des contrôles et une harmonisation sur les données devront être mis en place par AXA mais également par les courtiers qui déclarent les véhicules directement à l'AGIRA. La crise sanitaire liée à la Covid en France a impliqué le report de certains travaux de fiabilisation du FVA de la part de l'AGIRA. En effet, l'organisme propose de mettre en place un dispositif de suivi et de pilotage de la qualité des données par acteur. Ce dispositif en plusieurs étapes visera à améliorer la fiabilité du FVA en proposant à chaque société d'assurance, un accompagnement qualité. Aujourd'hui encore, l'AGIRA constate un grand nombre de réclamations de la part de propriétaires de véhicules assurés mais non encore déclarés ainsi que des temps de déclaration trop long. Des outils de vérification ont donc été développés afin de fluidifier les processus entre les assureurs et l'organisme gestionnaire.

TABLE 2.6 – Liste des variables du modèle final

Nom variable	Description
cnt_nafret	Code NAF de l'entreprise
cnt_sscateg_plr	Sous-catégorie du contrat en vision PLR
cnt_fract	Fractionnement du contrat
gta_segment	Segment SRA du véhicule
gta_classeprix	Classe de prix SRA du véhicule
gta_classrepar	Classe de réparation SRA du véhicule
gta_classe	Classe SRA du véhicule
gta_energie	Energie du véhicule
gta_trans	Transmission du véhicule
gta_groupe	Groupe SRA du véhicule
gta_pfisc	Puissance fiscale SRA du véhicule
gta_poidsvd	Poids à vide du véhicule
gta_puisscee	Puissance réelle du véhicule
gta_nbplace	Nombre de place du véhicule
gta_emissionco2	Emission de CO_2 du véhicule
veh_pfiscret	Puissance fiscale du véhicule
veh_ageveh	Âge du véhicule
veh_carrosret	Carosserie du véhicule
veh_risqmarq	Catégorisation de la marque du véhicule
veh_ageveh	Âge du véhicule
veh_modefin	Mode de financement du véhicule
veh_regpuis	Regroupement de la puissance fiscale du véhicule
veh_marque_RC	Catégorisation de la marque du véhicule
veh_ptac	Poids Total Autorisé en Charge du véhicule
veh_vhvan	Valeur à neuf du véhicule
veh_cdgenrn	Code genre du véhicule

Par ailleurs, le schéma 2.1 ci-après permet d'avoir une vision de la concaténation des différentes bases décrites précédemment en fonction des variables de la table 2.6.

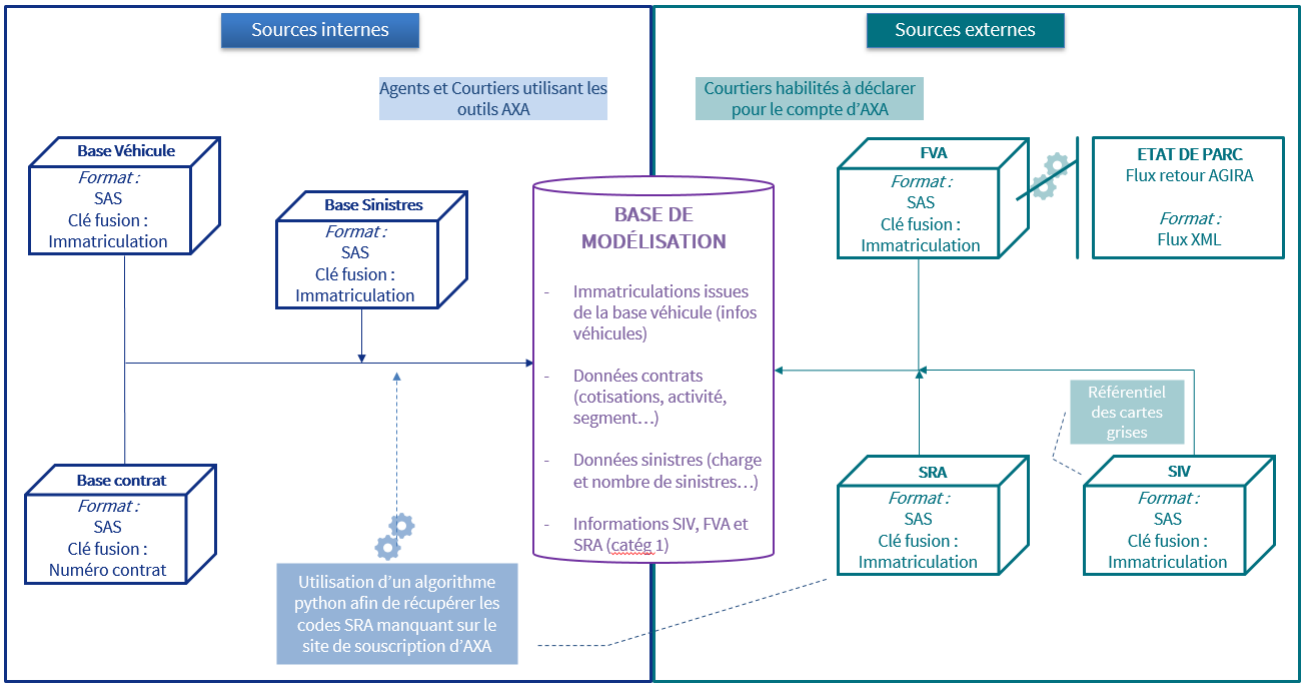


FIGURE 2.1 – Schéma de création de la base de données du modèle

Chapitre 3

La théorie de la modélisation du risque

Chez AXA France, des outils sont constamment développés ou améliorés afin d'affiner la tarification des flottes automobiles. La tarification actuellement implémentée dans les calculettes à destination des souscripteurs est basée sur un modèle de fréquence et de coût moyen que nous allons décrire dans ce chapitre. Cette tarification peut être utilisée aussi bien en affaire nouvelle qu'en renouvellement bien que dans ce dernier cas, la sensibilité du souscripteur à l'affaire et sa connaissance de la rentabilité de la flotte peuvent être des facteurs influents pour le tarif. Nous décrirons dans ce chapitre, les méthodes usuelles de tarification en assurance. Nous aborderons ensuite, par le biais de l'analyse de la rentabilité du segment des flottes ouvertes, les méthodes de tarification en renouvellement et notamment les notions de théorie de la crédibilité. Enfin, dans le cadre de cette étude qui conduira à une segmentation plus fine de la tarification actuelle, nous conclurons sur les différences entre individualisation du risque et segmentation.

3.1 Les modèles de tarification

Cette partie présentera les principaux modèles de tarification utilisés en assurance automobile afin de caractériser au plus juste la prime pure, c'est-à-dire l'espérance du coût des sinistres. Nous verrons dans un deuxième temps, comment modéliser cette prime pure.

3.1.1 Le modèle Fréquence × Coût Moyen

Afin de déterminer la loi ainsi que la combinaison optimale des variables explicatives caractérisant au mieux la prime pure, nous allons utiliser la décomposition de l'espérance du coût des sinistres en fréquence × coût moyen.

En effet, la définition de la cotisation d'assurance payée par l'assuré peut être effectuée par le biais de la charge sinistre. L'estimation de cette charge revient à minimiser l'écart quadratique moyen :

$$\mathbb{E}[(S-P)^2]$$

Avec S, variable aléatoire caractérisant la charge sinistre totale pour une période donnée et P constante définissant la prime pure que l'assuré devrait payer en théorie. Cette prime doit donc permettre à l'assureur de régler la totalité des sinistres survenus durant la même période. Il faut donc trouver la constante qui minimisera la distance de S à P. Nous avons :

$$\begin{aligned}\mathbb{E}[(S-P)^2] &= \mathbb{E}[(S-\mathbb{E}[S]+\mathbb{E}[S]-P)^2] \\ &= \mathbb{E}[(S-\mathbb{E}[S])^2] + 2(\mathbb{E}[S]-P)\mathbb{E}[S-\mathbb{E}[S]] + (\mathbb{E}[S]-P)^2\end{aligned}$$

Or $\mathbb{E}[S-\mathbb{E}[S]] = 0$, d'où :

$$\mathbb{E}[(S-P)^2] = \mathbb{E}[(S-\mathbb{E}[S])^2] + (\mathbb{E}[S]-P)^2$$

On remarque alors que cette équation est minimisée lorsque :

$$P = \mathbb{E}[S]$$

Soient N le nombre de sinistres survenus pendant la période étudiée et X_i suite de variables réprésentant le montant du i$^{\text{ème}}$ sinistre et supposées indépendantes et identiquement distribuées, nous pouvons définir la charge totale des sinistres telle que :

$$S = \sum_{i=1}^{N} X_i$$

Sous réserve de l'indépendance entre le nombre de sinistre N et le coût des sinistres X_i et parce que les X_i sont supposées i.i.d. [1], on a :

$$\begin{aligned}\mathbb{E}[S] &= \sum_{k=1}^{+\infty} \mathbb{P}[N=k] \times \mathbb{E}\left[\sum_{i=1}^{k} X_i\right] \\ &= \left(\sum_{k=1}^{+\infty} \mathbb{P}[N=k] \times k\right) \times \mathbb{E}[X_1] \\ &= \mathbb{E}[N] \times \mathbb{E}[X_1]\end{aligned}$$

On voit ainsi que la charge moyenne des sinistres s'exprime en fonction du nombre moyen des sinistres, autrement dit, la fréquence des sinistres et du coût moyen.

L'indépendance entre les coûts moyens et la fréquence des sinistres est une hypothèse forte qui peut être contestée car en pratique pas toujours vérifiée mais elle reste négligeable par rapport aux autres incertitudes liées à la tarification et elle permet d'obtenir la décomposition de la prime pure en fréquence × coût moyen ainsi que la propriété qui suit sur la variance.

En effet, en utilisant la décomposition de la variance, on obtient :

$$Var[S] = \mathbb{E}[Var[S|N]] + Var[\mathbb{E}[S|N]]$$

1. i.i.d. : Indépendantes et identiquement distribuées

Or, les X_i étant i.i.d. on a :

$$\begin{aligned}\mathbb{E}[S|N] &= \mathbb{E}\left[\sum_{i=1}^{N}\frac{X_i}{N}\right] \\ &= N \times \mathbb{E}[X_1]\end{aligned}$$

Puis en utilisant l'indépendance des fréquences et des coûts moyens, on a :

$$\begin{aligned}Var[S|N] &= Var\left[\sum_{i=1}^{N} X_i\right] \\ &= \sum_{i=1}^{N} Var[X_i] \\ &= N \times Var(X_1)\end{aligned}$$

D'où :

$$Var[S] = \mathbb{E}[N] \times Var[X_1] + Var[N] \times \mathbb{E}[X]^2$$

Nous pouvons ainsi remarquer que la variance de la charge totale des sinistres s'exprime en fonction de la variance due à l'incertitude concernant le nombre de sinistres et d'autre part, en fonction de la variance consécutive à l'incertitude sur les montants des sinistres.

3.1.2 Le modèle de Prime Pure

Une autre approche permettant de modéliser directement la prime pure consiste à ne considérer que le coût des sinistres. Cette méthode permet de prendre en compte l'ensemble des assurés qu'ils aient ou non un sinistre. En effet, un grand nombre de contrats sont non sinistrés et ont par conséquent, une charge nulle. Les contrats sinistrés ont alors une charge continue et positive.

La distribution de Tweedie permet d'estimer directement la prime pure d'un contrat d'assurance sans faire intervenir la décomposition classique fréquence × coût moyen. En effet, cette distribution présente la particularité d'avoir une masse de probabilité en 0, autrement dit, une valeur positive en 0 et une densité continue sur $]0;\infty[$. La distribution de Tweedie appartient à la classe des modèles de dispersion exponentielle que nous définirons dans la partie 3.2.2. Elle est alors caractérisée par sa densité qui s'exprime sous la forme :

$$f_Y(y|\theta,\phi) = exp\left(\frac{y\theta - v(\theta)}{u(\phi)} + w(y,\phi)\right), \quad y \in \mathbb{R}$$

avec :

$$\begin{aligned}\theta &= \frac{\mathbb{E}(Y)^{1-p}}{1-p} \\ u(\phi) &= \phi \\ v(\theta) &= \frac{\mathbb{E}(Y)^{2-p}}{2-p} \\ w(y,\phi) &= ln\left(\sum_{n=1}^{+\infty}\frac{\beta^{n\alpha}\lambda^n y^{n\alpha}-1}{\Gamma(n\alpha)n!}\right)\end{aligned}$$

où :

$$\beta = \frac{1}{\phi(1-p)} \quad \text{et} \quad \lambda = \frac{1}{\phi(2-p)}$$

ainsi que par la relation particulière entre sa variance et son espérance :

$$V(Y) = \phi\left[\mathbb{E}(Y)\right]^p$$

où ϕ est le paramètre de dispersion et $p \in \mathbb{R}$ est le paramètre de forme de la distribution ou paramètre de puissance.

La famille de Tweedie contient ainsi plusieurs lois usuelles en fonction de la valeur du paramètre p.

— Si $p = 0$, on retrouve une loi Normale. En effet, si $Y \sim \mathcal{N}(\mu, \sigma^2)$, on a :

$$\phi\left[\mathbb{E}(Y)\right]^0 = \phi = \sigma^2 = Var(Y)$$

— Si $p = 1$, on retrouve une loi de Poisson. En effet, si $Y \sim \mathcal{P}(\lambda)$, on a :

$$\phi\left[\mathbb{E}(Y)\right]^1 = \mathbb{E}(Y) = \lambda = Var(Y) \quad \text{avec} \quad \phi = 1$$

— Si $p = 2$, on retrouve une loi Gamma. En effet, si $Y \sim \mathcal{G}(p, \lambda)$, on a :

$$\phi\left[\mathbb{E}(Y)\right]^2 = \frac{1}{p} \times \frac{p^2}{\lambda^2} = \frac{p}{\lambda^2} = Var(Y)$$

— Si $p = 3$, on retrouve une loi Gaussienne Inverse. En effet, si $Y \sim \mathcal{IG}(\mu, \lambda)$, on a :

$$\phi\left[\mathbb{E}(Y)\right]^3 = \frac{1}{\lambda} \times \mu^3 = \frac{\mu^3}{\lambda} = Var(Y)$$

Pour les valeurs de $p > 3$, les distributions sont toujours définies mais sont plus compliquées à estimer car elles ne peuvent pas être écrites sous une forme finie.

Le cas $1 < p < 2$ est le cas le plus pertinent pour notre étude car les distributions sont alors continues pour $Y > 0$ mais avec une quantité positive lorsque $Y = 0$. La distribution de Tweedie, qui fait alors intervenir une loi de Poisson composée avec une loi Gamma, permet donc de ne pas imposer d'hypothèse d'indépendance entre la fréquence et le coût moyen mais également de modéliser la présence de nombreux contrats avec une charge nulle tout en tenant compte des sinistres dont la charge serait strictement positive, c'est-à-dire :

$$S = \sum_{i=0}^{N} Y_i, \quad \text{avec} \quad Y_0 = 0$$

3.1.3 Application : La Calculette Flottes Ouvertes

Au sein d'AXA France IARD, les souscripteurs flottes ouvertes de la branche automobile bénéficient d'outils spécifiques pour la souscription leur permettant d'établir un tarif rapidement. Cette calculette, développée par les actuaires, leur permettent de déterminer la prime commerciale à partir du calcul d'une prime pure établie selon la méthode sélectionnée.

En effet, trois méthodes basées sur le calcul de la fréquence et du coût moyen sont proposées aux souscripteurs :

— La méthode 1 fait intervenir la fréquence du client ainsi que le coût moyen de ses sinistres. Cette méthode peut être utilisée pour l'ensemble des garanties à souscrire (RC, DOMA, BDG...). Concernant la garantie RC, la fréquence du contrat est calculée à partir des véhicules à moteur composant la flotte. Le coût moyen est calculé à partir de la charge écrêtée et majorée d'un coefficient de mutualité afin de tenir compte de la marge nécessaire à l'indémnisation des sinistres graves. Le seuil d'écrêtement et le chargement pour graves, que nous décrivons en détail dans la partie 3.4, dépendent de l'usage principal de la flotte et de sa taille. En effet, ces deux indicateurs varient fortement selon qu'il s'agit d'une petite flotte de véhicules légers ou d'une grosse flotte de poids lourds par exemple. Cette méthode, faisant intervenir les fréquences et coûts moyens propres au client est la plus robuste pour les flottes de taille significative. En effet, plus la taille est importante, plus la crédibilité à apporter aux données est élevée.

— La méthode 2 tient également compte de la fréquence du client mais se réfère au coût moyen AXA. Il est calculé par catégorie de véhicule et par garantie à partir des données observées du portefeuille PARC et FLOTTE d'AXA France. Cette méthode est préconisée pour les flottes de faible taille qui peuvent présenter une forte disparité de charges d'un exercice à l'autre et rendant le coût moyen du client peu exploitable.

— La méthode 3 est fondée uniquement sur le référentiel AXA, tant en fréquence qu'en coût moyen. Elle n'est pas proposée en RC où les données du client sont fondamentales. Elle n'est destinée qu'aux garanties de type DOMA lorsque le contrat en vigueur ne comportait pas la garantie ou lorsque les antécédents sinistres et la flotte associée à ces garanties ne sont pas connus.

En RC, le choix de la méthode 1 suppose que l'on ait non seulement un volume de sinistre suffisant pour que la fréquence et la charge client soient représentatifs, ce qui implique une flotte de taille suffisamment importante, mais également une vision statistique exhaustive de la charge. La méthode 2 nécessite pour sa part, d'avoir une connaissance précise du nombre de sinistres. Afin d'aider le souscripteur dans son choix de méthode, des indicateurs sont mis en place afin de mettre en exergue les "anomalies" rencontrées telles qu'une fréquence supérieure de plus de 20% à la fréquence constatée sur le portefeuille AXA ou un coût moyen supérieur également à 20% du coût moyen du portefeuille AXA.

Lors de l'établissement des référentiels AXA, la non connaissance du nombre de véhicules sur le périmètre des flottes ouvertes a impliqué l'utilisation des statistiques sinistres du PARC. Les fréquences du référentiel ont été établies à partir de la sinistralité observée sur les contrats PARC sur une période de 3 ans afin de minimiser la volatilité induite par les années considérées comme atypiques. Les coûts moyens ont été calculés sur les périmètres des PARCS et FLOTTES afin d'avoir une grande stabilité, à partir de données constatées sur une durée de 10 ans. Ces référentiels, présentés en annexe A.1, ont été segmentés par garanties, en fonction de la responsabilité du sinistre et en dissociant les segments de véhicule par tonnage.

Ces méthodes d'estimation de la prime pure pose plusieurs problèmes. En effet, la segmentation mise en place dans la calculette ne permet pas de tenir compte des spécificités de tous les véhicules qui composent une flotte ouverte et affecte la fréquence et le coût moyen du segment majoritaire à l'ensemble des véhicules. Or, il peut arriver qu'une entreprise qui souhaite couvrir l'ensemble des véhicules qui composent sa flotte ait différents segments de véhicules. Le résultat de ce travail pourra

donc répondre à cette problématique en intégrant dans l'outil de tarification des flottes ouvertes, une nouvelle méthode permettant l'établissement d'un tarif véhicule par véhicule.

3.2 Les Modèles Linéaires

Les modèles linéaires généralisés (GLM) introduits par Nelder & Wedderburn (1972) et développés ensuite par McCullagh & Nelder (1989) sont largement utilisés de nos jours par les statisticiens et les actuaires afin de modéliser la fréquence et le coût moyen des sinistres. Généralisation des modèles linéaires gaussiens, ils ont l'avantage de permettre l'utilisation de lois non nécessairement gaussiennes et permettent ainsi de modéliser des événements qui ne suivent pas de distributions normales.

3.2.1 Le modèle linéaire gaussien

Avant de pouvoir définir les modèles linéaires généralisés, il convient de comprendre en premier lieu, le modèle linéaire gaussien. En effet, ce modèle permet de mettre en relation une variable quantitative Y dite à expliquer (ou encore variable réponse, exogène ou dépendante) avec p variables quantitatives $X_1, \ldots, X_p$ dites explicatives (ou encore de contrôle, endogènes, indépendantes). L'équation du modèle s'écrit alors :

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j \times X_{ij} + \varepsilon_i, \quad i = 1, \ldots, n$$

où :

— $X_{i1}, \ldots, X_{ip}$: variables explicatives associées à l'individu i ;
— $\beta_0, \ldots, \beta_p$: paramètres inconnus à estimer et supposés constants ;
— $\varepsilon_1, \ldots, \varepsilon_n$: termes d'erreur, non observés et i.i.d., de moyenne nulle et de variance constante.

Le modèle est dit gaussien lorsque l'on suppose que $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ pour $i = 1, \ldots, n$ et où σ^2 paramètre inconnu à estimer.

3.2.2 Le modèle linéaire généralisé

Comme le décrivent Denuit & Charpentier (2005), le modèle linéaire généralisé est caractérisé par les 3 composantes suivantes :

— La variable à expliquer Y est la **composante aléatoire** à laquelle on associe une loi de probabilité ;
— Les variables explicatives $X_1, \ldots, X_p$ correspondent à la **composante déterministe** ;
— Une **fonction de lien** qui décrit la relation entre les variables explicatives et $\mathbb{E}[Y]$.

Composante aléatoire

La composante aléatoire permet d'identifier la loi de la variable à expliquer Y. Les lois possibles pour la variable à expliquer doivent appartenir à la famille exponentielle, c'est-à-dire que la densité (ou mesure de probabilité dans le cas discret) doit s'écrire :

$$f_Y(y|\theta,\phi) = exp\left(\frac{y\theta - v(\theta)}{u(\phi)} + w(y,\phi)\right) \tag{3.1}$$

où :

- $\theta \in \mathbb{R}$ est le paramètre de la moyenne supposé inconnu. θ est également appelé paramètre naturel de la famille exponentielle ;
- $\phi \in \mathbb{R}$ est le paramètre de dispersion supposé connu ;
- u, v, w sont des fonctions spécifiques à la famille exponentielle telles que u définie sur $\mathbb{R}$ non nulle, v définie sur $\mathbb{R}$ deux fois dérivable et w définie sur $\mathbb{R}^2$

Différentes distributions peuvent donc appartenir à la famille exponentielle. Il peut s'agir de la loi de Poisson, Bernoulli ou Binomiale pour des variables discrètes ou encore de la loi Normale ou Gamma pour des variables continues. Le lecteur trouvera des exemples d'utilisation de la formule 3.1 en annexe A.2.

Il convient de noter ici qu'il est possible d'affecter un poids connu ω_i aux observations tel que $u(\phi) = \frac{\phi}{\omega_i}$. Dans un soucis de simplification, on suppose ici que $\omega_i = 1$. L'expression 3.1 peut alors s'écrire sous sa forme canonique en posant :

$$\begin{aligned} Q(\theta) &= \frac{\theta}{\phi} \\ a(\theta) &= exp\left\{-\frac{v(\theta)}{\phi}\right\} \\ b(y) &= exp\{w(y,\phi)\} \end{aligned}$$

On obtient alors :

$$f_Y(y|\theta,\phi) = a(\theta)b(y)exp\{yQ(\theta)\} \tag{3.2}$$

Pour les lois de la famille exponentielle, l'espérance et la variance sont données par les formules suivantes :

$$\mathbb{E}(Y) = v'(\theta)$$

$$Var(Y) = u(\phi)v''(\theta)$$

On remarque que la variance de Y s'exprime comme le produit de deux fonctions :

- $u(\phi)$, fonction indépendante de θ et ne dépend que de ϕ ;
- $v''(\theta)$ qui dépend uniquement du paramètre θ. C'est la fonction variance.

Si on note $\mu = \mathbb{E}(Y)$, on remarque que le paramètre θ est lié à la moyenne μ. La fonction variance citée précédemment peut donc être définie en fonction de μ et est notée $V(\mu)$.

Les composantes des principales distributions de la famille exponentielle sont présentées dans le tableau 3.1 ci-après. Les densités associées sont présentées en annexe A.3.

TABLE 3.1 – Les principales composantes de la famille exponentielle

Distribution	$\boldsymbol{\theta}$	$\mathbf{v}(\boldsymbol{\theta})$	$\mathbf{u}(\boldsymbol{\phi})$	$\mathbf{w}(\mathbf{y},\boldsymbol{\theta})$
Normale $\mathcal{N}(\mu,\sigma^2)$	μ	$\frac{\theta^2}{2}$	σ^2	$-\frac{1}{2}ln(2\pi\sigma^2)-\frac{y^2}{2\sigma^2}$
Bernoulli $\mathcal{B}(1,p)$	$log\left(\frac{p}{1-p}\right)$	$log(1+e^\theta)$	1	0
Binomiale $\mathcal{B}(n,p)$	$log\left(\frac{p}{1-p}\right)$	$nlog(1+e^\theta)$	$\frac{1}{n}$	$log\binom{n}{y}$
Poisson $\mathcal{P}(\lambda)$	$log(\lambda)$	$exp(\theta)$	1	$-log(y!)$
Gamma $\mathcal{G}(p,\lambda)$	$-\frac{1}{p}$	$-log(-\theta)$	$\frac{1}{\lambda}$	$\left(\frac{1}{\phi}-1\right)log(y)-log\left(\Gamma\left(\frac{1}{\phi}\right)\right)$

Composante déterministe

La composante déterministe est définie comme la fonction linéaire des variables explicatives $X_1,\ldots,X_p$, utilisées comme prédicatrices dans le modèle. Le prédicteur linéaire η est donc défini comme combinaison linéaire des variables explicatives et des paramètres du modèle de sorte que :

$$\eta_i=\beta_0+\beta_1X_{i1}+\cdots+\beta_pX_{ip},\quad i=1,\ldots,n$$

On note également sous forme matricielle :

$$\eta=X\beta$$

Où X est la matrice du modèle représentant les valeurs X_{ij} des variables explicatives $X_1,\ldots,X_p$, β est le vecteur des paramètres du modèle avec $\beta=\begin{pmatrix}\beta_1\\ \vdots\\ \beta_p\end{pmatrix}$ et où on suppose que $p\leq n$

Fonction de lien

La troisième composante d'un modèle linéaire généralisé est la fonction de lien, strictement monotone et différentiable, qui exprime une relation fonctionnelle entre la composante aléatoire et le prédicteur linéaire η. En effet, contrairement aux modèles linéaires classiques, les valeurs prédites ne correspondent pas à la prédiction moyenne d'une observation mais à une transformation, par une fonction mathématique, de cette observation.

Si on note $\mu_i=\mathbb{E}(Y_i)$ pour $i=1,\ldots,n$, on a :

$$\eta_i=g(\mu_i)=g(v'(\theta_i)),\quad i=1,\ldots,n$$

Définie sur $\mathbb{R}$, g est appelée fonction de lien. L'espérance de Y correspondant à une transformation linéaire du prédicteur linéaire, la fonction g permet donc d'expliquer comment évolue l'espérance de Y en fonction des variables explicatives.

La fonction de lien étant monotone et différentiable, elle est donc inversible. On peut alors également définir la fonction de lien canonique qui associe la moyenne μ_i au paramètre naturel et qui vérifie la relation :

$$\mu_i = g^{-1}(\theta_i) \iff \theta_i = \eta_i$$

Le choix de la fonction de lien est important car elle permet de s'assurer que les valeurs prédites restent dans des limites raisonnables et respectent la nature des valeurs d'origine de la variable à expliquer. En effet, dans le cadre par exemple de données de comptage, il est nécessaire de s'assurer que les prédictions du modèle ne donne que des valeurs positives ou nulles. Dans ce cas, une fonction de lien de type logarithme est appropriée car les valeurs ajustées sont exponentielles du prédicteur linéaire et donc positives ou nulles. Dans le cadre de ce travail, cette fonction de lien permet notamment d'éviter la modélisation de primes pures négatives.

Enfin, si plusieurs fonctions de lien peuvent convenir pour la variable à expliquer, il faudra veiller à choisir la fonction de lien qui permet de minimiser les écarts entre les valeurs prédites et les valeurs réelles, ce que nous verrons dans le paragraphe 3.3.3. Les fonctions de lien les plus couramment utilisées sont listées dans le tableau 3.2 ci-dessous.

TABLE 3.2 – Exemples de fonction de lien

Lien	**Fonction de lien**
Log	$g(\mu_i) = log(\mu_i)$
Logit	$g(\mu_i) = log\left(\frac{\mu_i}{1-\mu_i}\right)$
Identité	$g(\mu_i) = \mu_i$
Inverse	$g(\mu_i) = \mu_i^- 1$
Probit	$g(\mu_i) = \Phi^{-1}(\mu_i)$ où $\Phi(.)$ fonction de densité de $\mathcal{N}(0,1)$

3.2.3 Estimation des paramètres du modèle

Les caractéristiques des modèles linéaires généralisés ayant été définies, il faut maintenant s'intéresser à l'estimation des différents coefficients de régression qui composent le modèle. Ces estimations sont réalisées grâce à la méthode du maximum de vraisemblance, c'est-à-dire en maximisant la fonction de log-vraisemblance.

Si on considère la suite de variables à expliquer $(Y_i)_{1\leq i\leq n}$ indépendantes et issues d'une famille exponentielle, l'expression de la vraisemblance est de la forme :

$$L(y_1, \ldots, y_n; \theta_i, \phi) = exp\left\{\sum_{i=1}^{n} \frac{y_i\theta_i - v(\theta_i)}{u(\phi)} + w(y_i, \phi)\right\}$$

Si on note $L = L(y_1, \ldots, y_n; \theta_i, \phi)$, alors on a :

$$log(L) = \sum_{i=1}^{n} \frac{y_i\theta_i - v(\theta_i)}{u(\phi)} + w(y_i, \phi) \tag{3.3}$$

On maximise alors l'expression 3.3 en commençant par remarquer que θ dépend de β puis en calculant sa dérivée en fonction des paramètres β_j. On a alors :

$$\frac{\partial log(L)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial}{\partial \beta_j} \left(\frac{y_i\theta_i - v(\theta_i)}{u(\phi)} + w(y_i, \phi) \right)$$

Maximiser 3.3 revient donc à chercher les coefficients $\beta_1, \ldots, \beta_p$ qui vérifient les équations :

$$\frac{\partial log(L)}{\partial \beta_j} = 0, \quad j = 1, \ldots, p$$

Or, on a pour tout $i = 1, \ldots, n$:

$$\frac{\partial log(L)}{\partial \beta_j} = \frac{\partial log(L)}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_j}{\partial \beta_j}$$

De plus, comme $\mu_i = v'(\theta_i)$, on a :

$$\begin{cases} \frac{\partial log(L)}{\partial \theta_i} = \frac{y_i - v'(\theta_i)}{u(\phi)} = \frac{y_i - \mu_i}{u(\phi)} \\ \frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{\frac{\partial \mu_i}{\partial \theta_i}} = \frac{1}{v''(\theta_i)} \\ \frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{g'(\mu_i)} \\ \frac{\partial \eta_i}{\partial \beta_i} = x_{ij} \end{cases}$$

Étant donné que pour tout $i = 1, \ldots, n$ on a :

$$\frac{1}{v''(\theta_i)} = \frac{u(\phi)}{Var(Y_i)}$$

Les équations de vraisemblance à résoudre sont :

$$\frac{\partial log(L)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{Var(Y_i)g'(\mu_i)} = 0, \quad j = 1, \ldots, p$$

La résolution de ces équations non linéaires en β doit se faire numériquement car elle nécessite des méthodes itératives telles que la méthode de Newton-Raphson qui fait intervenir le Hessien ou encore la méthode de Fisher qui utilise la matrice d'information. Pour cette étude, nous allons décrire succinctement la méthode de Newton-Raphson qui est la méthode utilisée par défaut par le logiciel SAS, un des logiciels utilisés dans le cadre de ce travail.

L'algorithme, basé sur un développement de Taylor d'ordre 1, repose sur le processus récurrent suivant :

- Initialisation : choix du point initial $\beta^{(0)}$
- Récurrence : $\beta^{(k+1)} = \beta^{(k)} - H^{-1}(\beta) \times \nabla Log(L(\beta))$

— Condition d'arrêt : lorsque la différence entre $\beta^{(k+1)}$ et $\beta^{(k)}$ est plus petite qu'un certain seuil, autrement dit, $\beta^{(k+1)} \approx \beta^{(k)}$

Avec :

$$H(\beta) = \frac{\partial^2 log(L)}{\partial\beta\partial\beta'}, \quad \text{Matrice Hessienne de } Log(L)$$

Et :

$$\nabla Log(L(\beta)) = \begin{pmatrix} \frac{\partial log(L)}{\partial\beta^{(1)}} \\ \vdots \\ \frac{\partial log(L)}{\partial\beta^{(k)}} \end{pmatrix}, \quad \text{Vecteur gradient de } Log(L)$$

Le lecteur pourra se référer à Gonnet (2010) pour une description détaillée de la méthode itérative de maximisation.

3.2.4 Paramétrage des modèles

Les différentes méthodes permettant de modéliser la fréquence, le coût moyen ou encore la prime pure ayant été décrites et avant de s'intéresser aux critères de sélection des modèles, il est nécessaire d'effectuer un paramétrage des modèles, concernant à la fois la base de modélisation, mais également les paramètres de calibrage tels que les variables à prédire ou à expliquer.

Validation croisée

Afin de vérifier la qualité des modélisations, il est nécessaire de partionner la base de données. Une partie de ces données (10%) permettra de mesurer les performances de prédiction et sont présentes dans la base de validation. La base d'apprentissage contient quant à elle, l'autre partie des données (90%) qui serviront au calibrage des modèles.

La validation croisée permet alors de minimiser le risque de sur-apprentissage. En effet, cette méthode consiste à découper la base d'apprentissage en k échantillons de façon aléatoire et d'utiliser $k-1$ échantillons afin de calibrer le modèle. Ces échantillons de base, entrainés sur le modèle, sont appelés "base train". L'échantillon restant permettra de valider le modèle et est appelé "base test". Cette technique, appelée *k-fold*, permet d'assurer la stabilité du modèle en vérifiant l'homogénéité des indicateurs de qualité d'ajustement.

La figure 3.1 ci-après montre la validation croisée par un *4-fold*, méthode utilisée dans l'ensemble des modélisations de ce travail.

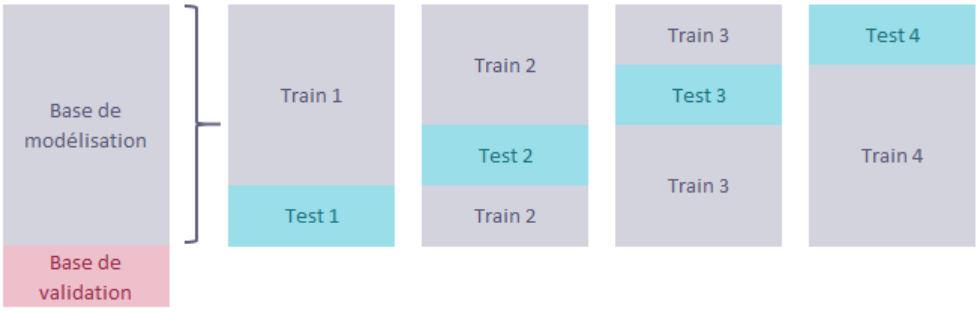


FIGURE 3.1 – Schéma représentant une validation croisée par un *4-fold*

Choix des paramètres de calibrage du modèle

Le choix des paramètres de calibrage du modèle est une étape essentielle à la modélisation. En effet, il s'agit de définir au préalable la variable à expliquer. Il peut s'agir du nombre de sinistres si l'on souhaite étudier un modèle de fréquence ou encore de la charge des sinistres dans le cadre d'une étude concernant un modèle de coût moyen. Il est également nécessaire de définir les variables explicatives en fonction des données dont on peut disposer, tout en tenant compte des différentes corrélations afin d'isoler le pouvoir explicatif de chacune d'entre elles sur la variable à prédire.

Dans cette étude, il a également été nécessaire d'intégrer l'exposition au risque de chaque observation. En effet, afin de ne pas biaiser l'estimation de la fréquence ou encore de la prime pure, le temps de présence de chaque véhicule a été ajouté dans le but d'attribuer un poids à chaque observation.

Enfin, dans une approche prime pure par la loi de Tweedie, il faut également définir le paramètre de puissance p qui peut être calibré automatiquement de sorte à s'ajuster le plus précisément aux données. Il est également possible d'imposer des contraintes sur les coefficients du modèle afin de corriger une erreur de calibrage ou intégrer un lissage plus important.

3.3 Les critères de sélection du modèle

Le cadre théorique des modèles linéaires généralisés ayant été présenté, il faut maintenant s'intérreser à la qualité de l'ajustement du modèle. En effet, plusieurs critères doivent être pris en compte tels que la significativité des paramètres estimés, l'information apportée par le modèle ou encore l'adéquation du modèle aux observations. Ce paragraphe traitera donc des différents facteurs permettant de juger de l'adéquation et permettant ainsi de sélectionner le modèle le mieux adapté.

3.3.1 Adéquation d'un modèle et tests de significativité

La déviance

La qualité d'ajustement d'un modèle peut être mesuré grâce à la statistique dite des écarts, également appelée déviance. En effet, cette statistique permet de comparer le modèle construit au modèle dit saturé qui comporte autant de paramètres à estimer que d'observations et qui font que ce modèle représente exactement les données. Le principe de la déviance est donc de comparer ces deux modèles en faisant intervenir leur vraisemblance respective.

En effet, si on note L la vraisemblance du modèle construit et L_{sat} la vraisemblance du modèle saturé, on définit alors le rapport :

$$\Lambda = \frac{L}{L_{sat}}$$

Plus Λ sera proche de 1 plus la qualité d'ajustement du modèle sera bon. En appliquant le logarithme, on a alors :

$$log(\Lambda) = Log(L) - Log(L_{sat})$$

Si ϕ désigne le paramètre de dispersion, on définit ainsi la statistique suivante :

$$\begin{aligned} D &= -2\phi log(\Lambda) \\ &= -2\phi \left[log(L) - log(L_{sat})\right] \end{aligned}$$

Qui peut également s'exprimer :

$$D = \phi D^* \quad \text{où} \quad D^* = -2\left[log(L) - log(L_{sat})\right] \quad \text{désigne la déviance standardisée.}$$

Ici, plus l'écart entre les log-vraisemblances sera faible, autrement dit, plus D sera petit, plus l'ajustement sera bon puisque la distance entre les valeurs modélisées et les valeurs observées sera faible. À l'inverse, une valeur élevée de D supposera une description des données de mauvaise qualité

Enfin, lorsque l'ajustement est de qualité, la déviance standardisée D^* suit asymptotiquement une loi χ^2_{n-p} où n correspond au nombre d'observations de la variable réponse, ou de manière équivalente, au nombre de paramètres du modèle saturé et p correspond au nombre de paramètres du modèle construit. Il sera donc possible, à partir de cette statistique, de construire un test permettant d'accepter ou de rejeter le modèle étudié. C'est l'objet du paragraphe suivant.

Le test du rapport des vraisemblances

Ce test permet de se rendre compte de la significativité des variables en calculant pour chaque variable explicative, l'écart entre la vraisemblance du modèle complet et la vraisemblance du modèle sans la variable à tester.

En effet, si on considère la statistique S définie ci-après et associée à la $j^{ème}$ variable du modèle, on a, si L_j représente la vraisemblance du modèle privé de la $j^{ème}$ variable à tester :

$$S = \frac{2\left[log(L) - log(L_j)\right]}{\phi} \sim \chi^2_{p-1} \quad \text{sous } H_0 : \beta_j = 0$$

On peut alors effectuer le test suivant :

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$$

En définissant la p-value p associée telle que :

$$p = \mathbb{P}(X > S) \quad \text{où} \quad X \sim \chi^2_{p-1}$$

La variable est alors significative lorsque $p < 5\%$ en règle générale.

Le test de Wald

Même si une variable peut paraître significative grâce au test du rapport des vraisemblances, toutes les modalités que cette variable peut prendre ne le sont pas nécessairement. Le test de Wald permet alors de comparer l'estimateur du maximum de vraisemblance $\hat{\beta}$ du paramètre β à une valeur β_0. Dans le cas univarié, le test revient à poser :

$$\begin{cases} H_0 : \beta = \beta_0 \\ H_1 : \beta \neq \beta_0 \end{cases}$$

Et :

$$W = \frac{(\hat{\beta} - \beta_0)^2}{Var(\hat{\beta})} \sim \chi^2_1$$

La modalité est significative lorsque $p < 5\%$ en règle générale et où :

$$p = \mathbb{P}(X > S) \quad \text{où} \quad X \sim \chi^2_1$$

Le cas multivarié, que nous ne décrirons pas dans ce travail, fait intervenir la matrice de covariance des paramètres β_j, l'inverse de la matrice d'information ainsi que la matrice de l'ensemble H_0 des hypothèses à tester.

3.3.2 Les critères AIC et BIC

À l'inverse du critère de la déviance qui permet de comparer deux modèles ayant les mêmes lois, les critères AIC et BIC permettent de comparer les modèles n'ayant pas les mêmes caractéristiques. Ils peuvent en effet être de lois différentes ou encore de fonction de lien différentes.

AIC

Introduit par Hirotugu Akaike en 1973, le critère d'information d'Akaike (AIC) permet d'apprécier de la qualité d'un modèle en comparant les modèles entre eux. En effet, en utilisant la log-vraisemblance des modèles, l'AIC pénalise les modèles dont le nombre de variables est trop

important et qui peuvent dans certains cas, avoir des effets de sur-ajustement. Si on note k, le nombre de paramètres du modèle, l'AIC est alors défini par :

$$AIC = -2log(L) + 2k$$

Lorsque le nombre d'observations n est trop petit par rapport au nombre de paramètres k du modèle, on peut également utiliser l'AIC corrigé défini par :

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1}$$

Par définition, plus l'AIC sera petit, meilleur sera le modèle.

BIC

Le critère d'information Bayésien (BIC), introduit par Gideon Schwartz en 1978, est un dérivé de l'AIC. Ici, la pénalité dépend de la taille n de l'échantillon et pas seulement du nombre k de paramètres du modèle. Il est défini par :

$$BIC = -2Log(L) + k \times ln(n)$$

Le meilleur modèle sera celui qui minimise le critère BIC.

3.3.3 Analyse des résidus

L'étude des critères présentés jusqu'à présent ne suffit pas pour valider un modèle. Il est en effet nécessaire d'analyser les écarts entre les valeurs prédites $\hat{y}_i$ et les valeurs observées y_i du modèle. L'analyse des résidus permet alors de détecter les valeurs abérantes pénalisant le modèle. Nous allons présenter dans cette partie, les différents résidus les plus couramment utilisés lors de la validation des modèles.

Les résidus bruts

Les résidus bruts r_i sont définis pour chaque observation i comme l'écart entre cette observation et sa valeur prédite. On a alors :

$$r_i = y_i - \hat{y}_i$$

Ce cadre est, en général, inapdapté car les $(r_i)_{1 \leq i \leq n}$ n'ont pas la même variance. On définit alors leur version standardisée.

Les résidus de Pearson

Les résidus de Pearson sont obtenus en effectuant le rapport entre les résidus bruts et l'écart-type estimé de $\hat{y}_i$ noté s_i. On a alors :

$$\begin{aligned} r_{P_i} &= \frac{y_i - \hat{y}_i}{s_i} \\ &= \frac{y_i - \hat{y}_i}{\sqrt{\widehat{Var(\hat{y}_i)}}} \end{aligned}$$

La version standardisée de ces résidus est donnée grâce à l'estimation des écarts-types s_i et on a :

$$r^*_{P_i} = \frac{y_i - \hat{y}_i}{s_i\sqrt{h_{ii}}}$$

où h_{ii} correspond au i$^{\text{ème}}$ terme diagonal de la matrice chapeau H définie par $H = W^{\frac{1}{2}}X(X^tWX)^{-1}X^tW^{\frac{1}{2}}$ avec W matrice diagonale dont le i$^{\text{ème}}$ terme diagonal est $w_i = \frac{1}{Var(\widehat{y}_i)(g'(\hat{y}_i))^2}$ où g est la fonction de lien.

Graphiquement, plus les résidus seront centrés en 0, plus faible sera l'erreur de modélisation.

Les résidus de la déviance

De la même manière, il est possible de définir les résidus de la déviance. Ces résidus permettent de mesurer la contribution de chaque observation à la déviance du modèle par rapport au modèle saturé. On définit alors pour chaque observation :

$$r_{D_i} = signe(y_i - \hat{y}_i)\sqrt{D^*_i}$$

avec $D^*_i = -2\left[log(L(y_i;\hat{y}_i,\phi)) - log(L(y_i;y_i,\phi))\right]$.

La version standardisée est quant à elle définie par :

$$r^*_{D_i} = signe(y_i - \hat{y}_i)\sqrt{\frac{D^*_i}{h_{ii}}}$$

Graphiquement, si l'histogramme des résidus de la déviance est centré alors l'ajustement du modèle aux données sera bon.

3.3.4 Les autres indicateurs

D'autres indicateurs permettent de comparer les modèles entre eux. Il est possible de citer parmis eux, les indicateurs "graphiques" tels que l'indice de Gini qui fait intervenir la courbe de Lorenz ou encore la *lift curve*. Enfin, dans le cadre de ce travail, des indicateurs d'écarts seront également étudiés.

L'indice de Gini

Contrairement aux indicateurs décrits précédemment, l'indice de Gini ne permet pas de mesurer la qualité d'ajustement du modèle aux données. En revanche, c'est un indicateur de la dispersion d'une distribution dans une population. Souvent utilisé en économie afin de mesurer l'inégalité des richesses au sein d'un pays, il est utilisé en statistiques afin d'évaluer la qualité de la segmentation d'un modèle :

— Dans un modèle de fréquence, l'indice de Gini permet de vérifier que les assurés ayant le moins de sinistre sont modélisés par un risque plus faible que les assurés ayant un plus grand nombre de sinistres.
— Dans un modèle de coût moyen, il permet de vérifier que les sinistres les plus coûteux sont modélisés par une charge plus élevée que les sinistres moins onéreux.

Le calcul de l'indice de Gini se fait à partir de la courbe de Lorenz, contruite par l'économiste américain Max Otto Lorenz en 1905 qui souhaitait représenter graphiquement les inégalités de revenus.

Dans le cadre de cette étude et concernant le risque de fréquence par exemple, il est possible de représenter cette courbe sur le graphique 3.2 en plaçant sur l'axe des abscisses la part cumulée des contrats triés en ordre croissant par rapport au nombre de sinistres et en ordonnée, la part cumulée du nombre de sinistres observés. La bissectrice représentée sur le graphique 3.2 représente la situation d'égalité parfaite où x% des assurés ont x% des sinistres.

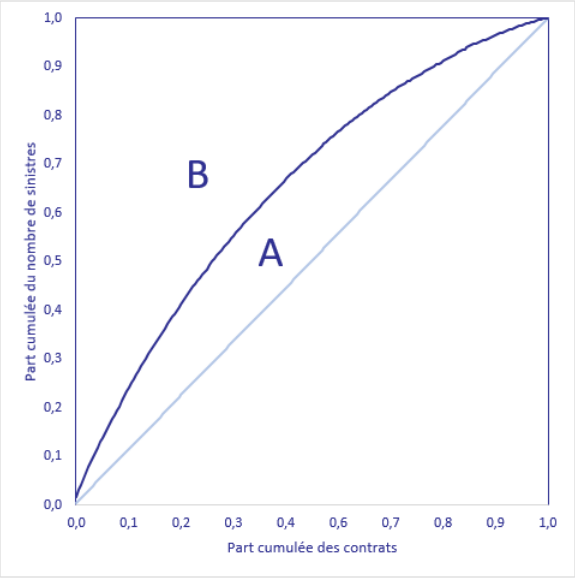


FIGURE 3.2 – Représentation graphique de la courbe de Lorenz et de l'indice de Gini

L'indice de Gini mesure l'aire entre la bissectrice et la courbe de Lorenz. Si A représente l'aire comprise entre la courbe de Lorenz et la première bissectrice et B l'aire au-dessus de la courbe de Lorenz, alors l'indice de Gini G est défini par :

$$G = \frac{A}{A+B}$$

L'indice de Gini est alors compris entre 0 et 1 et on remarque que lorsque le modèle tend vers l'observé, l'indice de Gini tend vers 1. On aura donc tendance à privilégier les modèles ayant un indice de Gini proche de 1 qui permettent d'éviter le mauvais classement des assurés en fonction de leur risque et par conséquent, éviter l'antisélection.

La *lift curve*

En complément de l'indice de Gini, la *lift curve* permet de visualiser l'adéquation du modèle aux données. En effet, la construction de cette courbe repose sur la segmentation en quantiles des risques prédits (fréquence ou coût moyen) et classés en ordre croissant. Pour chaque quantile, nous calculons alors l'observation moyenne du risque modélisé que nous comparons à la prédiction.Un écart d'adéquation des données sur les premiers quantiles permet de savoir s'il y a une sous-modélisation ou une sur-modélisation des contrats moins sinistrés tout comme les écarts sur les derniers quantiles permet de capter ce phénomène sur les contrats les plus sinistrés. Enfin, il est nécessaire de bien analyser les *lift curves* des différentes bases de données, qu'il s'agisse de la base de modélisation, de la base test ou encore de la base de validation afin de s'assurer qu'il n'y a pas de sur ou de sous modélisation dû à une spécification dans l'une de ces bases.

La *lift curve* du graphique 3.3 ci-après a été créé à partir d'une base de données triée en ordre croissant et segmentée en 20 quantiles. Les moyennes observées par quantiles sont représentées par la courbe violette et les modélisations moyennes sont représentées par la courbe jaune. On voit ici que le modèle s'ajuste correctement aux données.

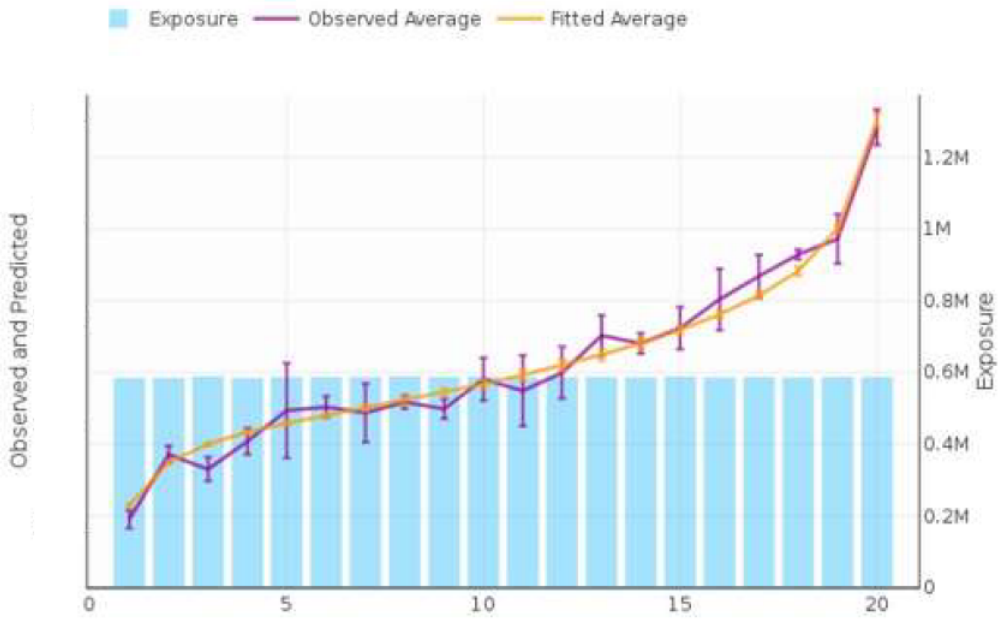


FIGURE 3.3 – Représentation graphique de la *Lift Curve*

Les indicateurs d'écarts

Pour conclure ce chapitre sur les critères de sélection du modèle, nous pouvons définir deux indicateurs supplémentaires courramment utilisés dans l'analyse des écarts au modèle.

L'erreur quadratique moyenne (MSE [1]) est une mesure d'erreur classique correspondant à la somme des carrés des écarts entre les prédictions ($\hat{y}_i$) et leurs observations (y_i) rapporté au nombre N des observations :

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

On définit également usuellement, la racine carré du MSE : $RMSE = \sqrt{MSE}$

Enfin, l'erreur absolue moyenne (MAE [2]) correspond à la somme des valeurs absolues des écarts entre les prédictions et leurs observation, rapportée au nombre des observations :

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$$

Le *spread*

Une fois les variables du modèle choisies selon leur corrélation, il est important de vérifier le degré d'importance de chacune d'entre elles. Le *spread* permet alors de mesurer l'importance de la discrimination des variables au travers de la dispersion des coefficients associés. Il correspond au ratio entre le coefficient le plus élevé et le coefficient le moins élevé associé à une variable et mesure ainsi l'écart maximal entre les deux coefficients des modalités d'une variable :

$$Spread_{var} = \frac{\text{Coefficient le plus élevé}_{var}}{\text{Coefficient le plus faible}_{var}}$$

Il est alors possible de dissocier le *spread 100/0%*, qui prend l'ensemble des données du portefeuille du *spread 95/5%* qui permet de tenir compte des valeurs extrêmes en excluant 5% des données du portefeuille ayant les coefficients les plus élevés et les plus faibles.

3.3.5 Les méthodes de sélection des variables tarifaires

Le nombre k de paramètres du modèle pouvant être très grand, il est difficile de tester l'exhaustivité des 2^k sous-ensembles possibles parmi ces paramètres. Il faut donc recourir à des méthodes dites pas à pas qui utilisent les tests statistiques décrits précédemment tels que la déviance ou encore

1. MSE : Mean Square Error
2. MAE : Mean Absolute Error

les critères de sélection AIC ou BIC. Nous explicitons ici 3 algorithmes de sélection des variables courrament utilisées dans l'application des modèles linéaires généralisés.

La méthode *backward*

La méthode *backward* (descendante) consiste à prendre en compte l'ensemble des variables explicatives dans le modèle puis éliminer la variable la moins significative à chaque pas de telle sorte que :

- Si la méthode utilise le test de la déviance par exemple, la variable explicative ayant la p-value associée au test la plus élevée est retirée. L'algorithme continue d'éliminer les variables ainsi jusqu'à ce que toutes les variables explicatives soient retirées si aucune variables n'expliquent le modèle ou jusqu'à ce que la p-value soit inférieure à un certain seuil défini au préalable.
- Si la méthode utilise un critère de choix tel que l'AIC ou le BIC par exemple, la variable explicative qui permet la plus grande diminution du critère est retirée du modèle. Cette étape est réitérée jusqu'à ce que toutes les variables soient retirées ou lorsque aucun autre retrait n'entraine de diminution du critère de choix.

La méthode *forward*

Contrairement à la méthode *backward*, la méthode *forward* (ascendante) consiste à prendre en compte le modèle avec la variable la plus significative et intègre à chaque pas, la variable qui contribue le plus au modèle :

- Si la méthode utilise le test de la déviance, la variable explicative dont la p-value associée au test qui compare les 2 modèles emboîtés est la plus faible est ajoutée au modèle. L'algorithme se poursuit jusqu'à ce que toutes les variables soient ajoutées au modèle ou jusqu'à ce que la p-value du modèle soit plus grande qu'un certain seuil défini au préalable.
- Si la méthode utilise un critère de choix tel que l'AIC ou le BIC, la variable explicative qui permet la meilleure optimisation du critère de choix sera ajoutée. Cette étape est réitérée jusqu'à ce que toutes les variables soient ajoutées ou jusqu'à ce qu'aucun ajout n'entraine d'augmentation du critère de choix.

La méthode *stepwise*

La méthode *stepwise* (progressive) est une combinaison des 2 méthodes précédemment décrites. En effet, elle permet, comme la méthode *forward*, d'ajouter des variables significatives au modèle et dans le même temps d'éliminer une variable déjà introduite qui serait moins significative avec l'ajout de cette nouvelle variable.

3.4 Le processus de renouvellement

Contrairement aux véhicules de particuliers, il n'est pas possible de mettre en place de système de bonus-malus sur le segment des flottes de véhicules. En effet, il n'y a généralement pas de conducteur attitré pour chacun des véhicules qui composent une flotte. Cette distinction implique donc une différenciation dans le processus de renouvellement du contrat flotte automobile car il fait intervenir un processus de majoration spécifique utilisant le ratio de sinistralité sur cotisation (S/C).

Cependant, la récente connaissance de l'immatriculation, peut modifier considérablement ce processus de renouvellement en permettant l'établissement d'un tarif "par véhicule" en fonction des caractéristiques propres à chaque véhicule composant la flotte. Afin de pouvoir apréhender ce changement, il convient donc de décrire le processus de renouvellement existant en abordant le sujet de la rentabilité des flottes ouvertes au sein d'AXA ainsi que la théorie sur laquelle est basée ce processus.

3.4.1 Rentabilité des flottes ouvertes AXA

Afin de comprendre l'importance du processus de renouvellement dans la vie d'un contrat flotte automobile chez AXA France, il convient d'apporter quelques notions sur l'analyse de la rentabilité.

En effet, AXA France a mis en place depuis plusieurs années, un indicateur de rentabilité conforme à la règlementation Solvabilité II : l'*Economic Combined Ratio* (ECR) ou Ratio Combiné Économique. Ce ratio permet de prendre en compte l'ensemble des produits et des charges inhérents à un contrat, un produit, un segment ou encore à une branche de portefeuille. Comme le décrit Rebadj (2016), l'ECR est donc défini comme la somme sur les n contrats du segment étudié, du rapport entre les charges diminuées des produits financiers générés d'une part, et d'autre part, les primes encaissées sur ce même segment pour une période donnée :

$$ECR^{ptf} = \sum_{i=1}^{n} \frac{(Charges_i^{ptf} - Produits_i^{ptf})}{Primes_i^{ptf}}$$

L'ECR peut donc être calculé sur une segmentation plus ou moins fine mais également pour tenir compte d'une période spécifique. En effet, il peut être nécessaire de ne pas tenir compte d'une année atypique en terme de sinistralité comme par exemple l'année 2020 et sa sinistralité particulière induite par les mesures de confinement pendant la pandémie liée à la COVID. Il est alors possible d'établir qu'un segment ayant un ratio d'ECR inférieur à 1 est rentable. À l'inverse, un segment ayant un ratio supérieur à 1 est considéré comme non rentable car déficitaire pour AXA.

La définition de l'ECR ayant été donnée et afin de comprendre l'utilisation du S/C dans le processus de renouvellement, il faut maintenant s'intéresser aux différentes variables qui composent cette formule. L'idée ici n'est pas de décrire en détail toute la théorie derrière la création des composantes de l'ECR mais de permettre au lecteur d'avoir une idée des méthodes utilisés.

En effet, l'ECR tient compte au titre des charges de :

— La sinistralité attritionnelle ou charge hors graves. Elle comprend l'ensemble des sinistres dont le montant n'excède pas 30 000 €, à l'exception des sinistres climatiques qui font l'objet d'une modélisation à part entière selon le modèle interne AXA France. Le seuil de 30 000 € est appelé seuil d'écrêtement et correspond au montant à partir duquel la charge sinistre sur-crête est considérée comme charge grave. Ce niveau de seuil, spécifique à la branche automobile, a été déterminé lors de précédentes études utilisant la modélisation de la charge sinistre ainsi que les quantiles de la loi associée. Le lecteur pourra se référer à Boyer-Chammard (2008) pour plus de détails concernant la construction de ce seuil.

— La sinistralité grave qui est composée de la part des sinistres hors climatiques compris entre 30 000 € et 2M €. Elle est modélisée à partir des observations sur 5 ans des sinistres graves, qui sont projetées en charges finales prévisibles à partir de triangles de projection utilisant la méthode de Chain Ladder sur un historique de 10 ans. Un taux de chargement pour graves est alors déterminé par rapport à la charge attritionnelle de la même période. Ce taux de chargement est ensuite décliné en fonction des segments. Le montant en euros de charges graves d'un segment peut alors être retrouvé en multipliant le taux de grave du segment par sa charge attritionnelle. L'hypothèse utilisée ici est que plus la sinistralité attritionnelle sera conséquente, plus le risque d'avoir des sinistres graves sera important.

— La sinistralité atypique correspond aux sinistres dont la charge excède 2M €. Comme pour la sinistralité grave, elle est considérée hors sinistres climatiques. Un niveau de charges atypiques finales prévisibles est également déterminé. Cette charge atypique est ensuite ventilée en fonction des segments selon leur poids respectif de charges sinistres supérieure à 500K € et cumulée sur 10 ans afin d'éviter d'intégrer de la volatilité et dans un esprit de mutualisation.

— La sinistralité climatique est une composante indépendante des autres composantes liées à la sinistralité. En effet, elle bénéficie d'une modélisation spécifique du fait de sa fréquence réduite (bien qu'en augmentation ces dernières années) et de sa probable sévérité. Historiquement, le taux de chargement pour sinistres climatiques était calculé à partir des charges climatiques finales prévisibles. Hors ce taux étant proche de 0, il a quand même été nécessaire d'intégrer cette dimension dans le calcul de l'ECR afin lisser ce que cette charge peut représenter pour AXA et notamment pour les garages et concessions qui portent l'essentiel du risque climatique automobile. Il a donc été décidé de modifier cette formule de calcul en tenant compte de l'historique de la sinistralité climatique sur 10 ans afin de conserver une certaine stabilité puis de répartir cette charge en fonction du poids en prime de chaque segment. Certains segments étant trop peu représentés, il a également été nécessaire de mutualiser quelques groupes. Les garages et concessions ont bénéficié d'un chargement spécifique. Les segments de véhicules de M3T5 ont été regroupés entre eux (VL, Missions, Convention), de même que les segments généralement composés de véhicules de P3T5 (TPM, TPV...).

— Le coût lié aux versements d'une rente par la compagnie d'assurance aux victimes d'accidents. Ce coût est estimé à partir du montant de rentes versées sur 10 ans rapporté à la charge finale prévisible cumulée également sur 10 ans. Cette charge est ensuite répartie selon le poids de la charge RCCORP de chaque segment.

— Le coût de la réassurance est également une composante à prendre en compte au titre des

charges. En effet, ce coût correspond au montant que paie AXA France afin de se couvrir en cas de sur-sinistralité sur le portefeuille de la branche automobile dans le cadre de ses traités en excédant placés auprès des réassureurs. Il s'agit donc bien d'une charge à répartir entre les différents segments. Cette répartition s'effectue selon la sinistralité observée au prorata du chiffre d'affaire de chaque segment.

— Le montant des frais généraux est une donnée communiquée par la direction financière. Ce montant correspond aux frais engagés par AXA France dans la gestion et le développement de ses produits. Il peut s'agir de frais de gestion de sinistres ou de frais liés à la production tels que les frais informatiques ou de marketing. Cette charge est répartie en fonction du chiffre d'affaire et du nombre d'actes de gestion de chaque segment. En effet, il est logique qu'un produit nécessitant plus d'actes de gestion ait un montant de frais généraux plus important.

— Le montant des commissions prélevées par les intermédiaires d'assurance. Ce montant est propre à chaque contrat.

— Le coût du risque qui correspond au coût lié à la capacité à souscrire les contrats d'assurance. En effet, selon la règlementation Solvabilité II, chaque compagnie d'assurance doit détenir un montant de capital pour chaque risque souscrit. Ce capital est rémunéré au-delà du taux sans risque et correspond à la rémunération attendue de la part de l'actionnaire.

— Les taxes et impôts payés par AXA au titre des revenus financiers générés par les actifs détenus par la compagnie d'assurance.

Du côté des produits, l'ECR tient compte de :

— L'escompte qui reflète le niveau des produits financiers générés par les provisions techniques pour sinistres.

— Les primes d'assurances payées par les assurées et incluses dans le ratio S/C.

Le schéma 3.4 ci-dessous permet de visualiser la construction de l'ECR :

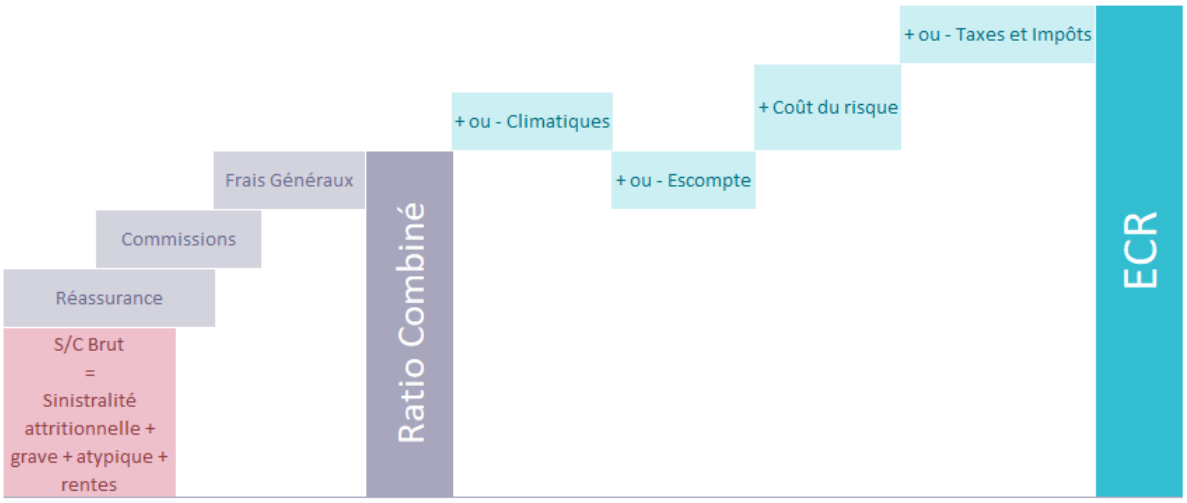


FIGURE 3.4 – Décomposition de l'ECR

L'ECR permet aux directions techniques (actuariat, comptes, direction technique automobile, COMEX [1],...) d'avoir une vision de la rentabilité d'une branche, d'un segment ou d'un produit tout en tenant compte de la fiscalité, du résultat d'investissement et de la rémunération du capital. En revanche, son utilisation par les équipes de souscription lors d'un renouvellement par exemple, semble peu aisée. Par ailleurs, l'ECR est un indicateur de rentabilité a posteriori. Lors de la souscription, il peut être difficile de lier cet indicateur aux différents indicateurs de performance d'un contrat. Il a donc été nécessaire d'introduire un indicateur spécifique à la souscription et équivalent de l'ECR d'un point de vue de la rentabilité a priori d'un contrat. AXA a donc mis en place le *Permissible Loss Ratio* ou le ratio sinistre à prime d'équilibre.

Il s'agit de déterminer un seuil maximal de S/C à partir duquel, l'ECR sera égal à 1, tout en tenant compte des différentes composantes liées à la sinistralité ou aux frais économiques (réassurance, commissions, frais généraux, événements climatiques, escompte...). Autrement dit, il s'agit du seuil maximal acceptable pour qu'une affaire soit à l'équilibre, compte tenu du taux de commission de l'affaire, des frais généraux, des produits financiers et d'évènements modélisés, comme les évènements climatiques ou les sinistres atypiques. Ce seuil permet au souscripteur de déterminer la rentabilité d'une affaire a priori, tout en intégrant les différentes charges, grâce au seul S/C. Ainsi, lorsque le S/C d'une affaire est supérieur au PLR, l'affaire est considérée comme non rentable. À l'inverse, un contrat ayant un S/C inférieur au PLR engendre des profits. Tout comme l'ECR, le PLR peut se décliner selon une branche, un segment ou encore un produit.

L'ECR se décompose donc en somme de plusieurs éléments :

- Le S/C Brut qui comprend les charges liées à la sinistralité attritionnelle, grave, atypique et aux versements des rentes (S/C Brut) ;
- Les coûts liés à la réassurance (REASS) ;
- Les commissions prélevées par l'intermédiaire d'assurance (COM) ;
- Les frais généraux (FGX) ;
- Le coût lié aux sinistres climatiques (CLIM) ;
- L'escompte (ESC) ;
- Le coût du risque (CoC) ;
- Les taxes et impôts (TAX).

Afin de comprendre comment le S/C d'une affaire peut être lié à l'ECR et par conséquent, au PLR, il est nécessaire d'étudier le passage de l'ECR au PLR. On a alors la formule suivante de l'ECR (en % de la prime) :

$$ECR = S/CBrut + \frac{FGX + COM + REASS + CLIM + ESC + CoC + TAX}{Prime}$$

Afin de déterminer un PLR, on cherche donc la majoration (dans le cadre d'un ECR supérieur à 1) ou la minoration (dans le cadre d'un ECR inférieur à 1) à appliquer à la prime afin d'obtenir un ECR égal à 1. En effet, le seul poste sur lequel il est possible d'influer afin de remettre à l'équilibre une affaire est le poste de prime du S/C. À partir du montant de prime majorée ou minorée selon la situation, on déduit les niveaux de S/C Brut, de frais liés aux rentes, de frais généraux, de commissions et de coût de réassurance. L'escompte, variant en fonction du S/C Brut y compris rentes, se déduit des calculs précédents ainsi que les taxes calculées sur la base du ratio combiné y compris escompte.

1. COMEX : Comité Exécutif

Ces éléments ayant été déterminés, les souscripteurs peuvent alors utiliser les PLR selon qu'ils regardent le S/C écrêté ou le S/C y compris graves définis par les formules suivantes :

$$\begin{cases} PLR_{écrêté} = S/CBrut_{PLR} - S/C_{Atyp} - S/C_{Graves} \\ PLR = S/CBrut_{PLR} - S/C_{Atyp} \end{cases}$$

Avec :

$$S/CBrut_{PLR} = \frac{100 - CoC - TAX - FGX - COM - REASS}{1 + ESC} - S/C_{Rentes}$$

Où :

- S/C_{Rentes} est le coût lié aux versements des rentes ;
- S/C_{Clim} est la charge liée aux sinistres climatiques ;
- S/C_{Atyp} est la charge liée aux sinistres atypiques ;
- S/C_{Graves} est la charge liée aux sinistres graves.

L'exemple décrit sur la figure 3.5 ci-dessous montre le passage de l'ECR au PLR. Dans ce cas, l'ECR est supérieur à 1. Une affaire dont le S/C écrêté est inférieur au PLR écrêté de 50% sera considérée comme "équilibrée".

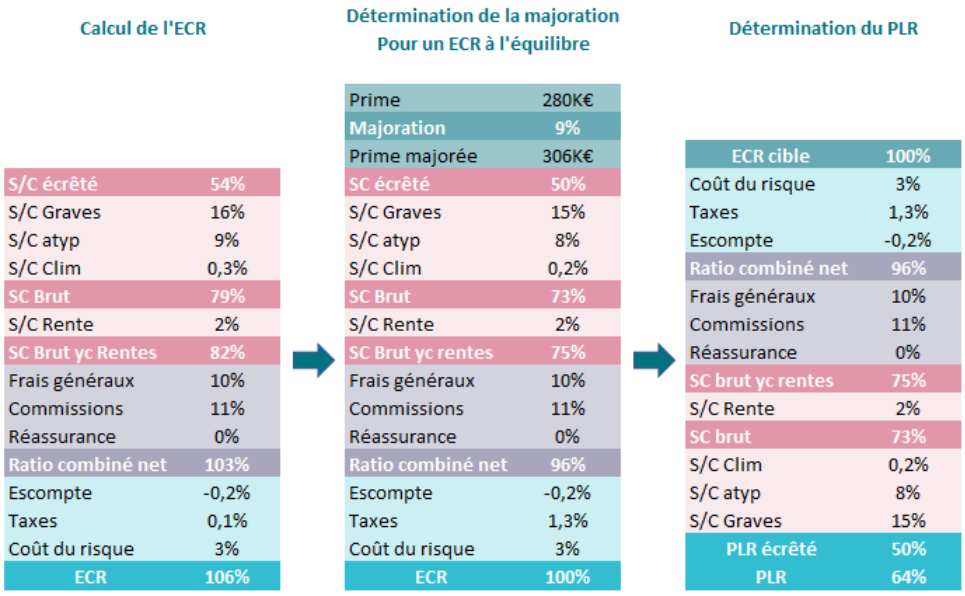


FIGURE 3.5 – Passage de l'ECR au PLR

3.4.2 Processus de majoration

L'utilisation seule du ratio de sinitres sur cotisations (S/C) ne suffit pas. En effet, le S/C peut varier fortement d'une année sur l'autre si par exemple, un contrat a eu un sinistre grave voir atypique au cours de l'année. C'est régulièrement le cas lors d'accidents corporels induits par un accident de cars par exemple, et dont le sinistre peut être évalué à plusieurs millions d'euros. Il est donc nécessaire de recourir à différentes méthodes afin de rendre plus stable cet indicateur,

notamment en ayant recours à des méthodes d'écrêtement comme nous l'avons vu précédemment afin de ne prendre en compte que la sinistralité attritionnelle sur plusieurs années afin de rendre le S/C encore plus stable mais également en ayant recours à des modèles de crédibilité.

Chez AXA, le choix a été fait d'analyser la sinistralité passée sur les trois dernières années afin de limiter la volatilité induite par une fréquence et un coût moyen en dehors de la moyenne. Même si l'écrêtement permet de réduire la variance du coût moyen et de mutualiser la part de graves à l'ensemble du segment étudié, le recours à la théorie de la crédibilité est indispensable afin de limiter l'instabilité du nombre et des coûts des sinistres : on considère alors le S/C d'un contrat comme une moyenne de son S/C propre et de celui de son segment voir de son portefeuille. En effet, pour un contrat donné, considérer son seul S/C serait dangereux du fait de la volatilité de ce dernier et lui affecter le S/C de son segment serait absurde dans la mesure où l'on observe une grande hétérogénéité au sein même d'un segment. La crédibilité permet alors de combiner ces deux approches en intégrant dans le calcul du S/C crédibilisé du contrat, un facteur Z dit de crédibilité, qui mesurera la fiabilité de l'information apportée par le S/C propre du contrat d'une part, et d'autre part, l'information apportée par le S/C du segment ou du portefeuille.

On a alors :

$$S/C_{crédibilisé} = Z \times S/C_{contrat} + (1 - Z) \times S/C_{segment}$$

- Lorsque $Z = 0$, le S/C crédibilisé correspond au S/C du segment. On considère alors que l'information apportée par le contrat est non fiable et on se fie uniquement au résultat du segment.
- Lorsque $Z = 1$, le S/C crédibilisé correspond au S/C du contrat. On considère alors que l'information apportée par le contrat est complètement fiable et suffit entièrement pour expliquer la sinistralité du contrat.

Le facteur de crédibilité peut être choisi selon plusieurs critères. On peut par exemple utiliser l'ancienneté d'un contrat, plus un contrat sera ancien et plus on pourra accorder de l'importance à ses statistiques de sinistralité. Concernant le périmètre des flottes, un des critères les plus discriminant est la taille de la flotte.

En effet, plus une flotte de véhicules sera importante, moins sa fréquence et le coût de ses sinistres variront car plus grand sera le nombre de ses sinistres. On observe par exemple sur certain gros contrats, une charge de sinistre récurrente et incompressible d'année en année sur lesquels il est possible d'appliquer une conservation. Pour ce type de contrat, on peut décider d'accorder une plus forte crédibilité soit au contrat s'il ne s'agit pas d'une conservation, soit au segment si le contrat est un contrat en conservation.

La théorie de la crédibilité a été développée par des actuaires américains au début du XX^{e} siècle qui avaient mis en place une technique appelée "théorie de la fluctuation limitée" ou "crédibilité américaine" qui consistait à tarifer un contrat au sein d'un groupe en pondérant les informations apportées par les données propres au contrat d'une part et par les données du groupe d'autre part. Cette méthode manquant de fondements mathématiques, l'approche actuelle repose sur les travaux d'Hans Bühlmann qui, dans les années 1960, a posé les bases, mathématiquement incontestables, en définissant clairement les hypothèses de cette théorie de la crédibilité dite "européenne" ou "moderne".

Sans entrer dans le détail car ce n'est pas l'objet de ce travail, il est possible de citer les modèles de crédibilité les plus couramment utilisés de nos jours en actuariat. En effet, le modèle initial de Bühlmann qui supposait tous les contrats identiques, fut amélioré afin de prendre en compte la différence de poids entre les contrats et inclure ansi l'information liée à la taille du contrat, c'est le modèle de Bühlmann-Straub.

Cependant, ce modèle suppose encore que tous les contrats sont identiques à un facteur de taille près. Or la taille du contrat n'est pas le seul facteur discriminant. On pourrait souhaiter par exemple, segmenter les contrats selon leurs catégories de véhicules ou en fonction du tonnage. En d'autres termes, on souhaiterait tenir compte des différentes segmentations d'un portefeuille donné. C'est ce que permet le modèle mis en place par Jewell en 1975. En effet, il s'agit d'un modèle hiérarchique qui divise le portefeuille initial en catégories de portefeuilles homogènes, chaque catégories de portefeuilles étant elles-mêmes divisées en plusieurs sous-catégories et ainsi de suite. Le schéma 3.6 ci-dessous illustre le modèle proposé par Jewell : la taille de chaque groupe diminuant de niveau en niveau, permet d'augmenter l'homogénéité de chaque classe. En revanche, il faut faire attention à l'individualisation de chaque groupe qui peut entraîner une certaine volatilité.

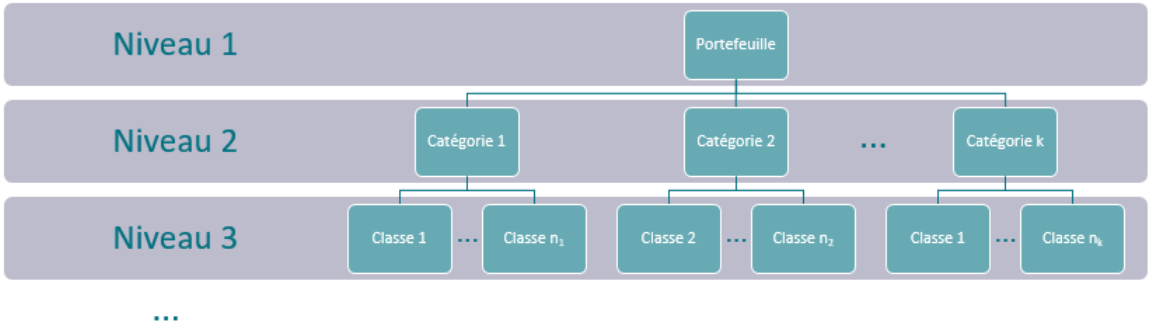


FIGURE 3.6 – Illustration du modèle de Jewell

Dans le cadre des renouvellements de la branche automobile, le processus de majoration est segmenté selon le réseau de distribution et l'ancienneté du contrat, par tranche de chiffre d'affaire ainsi que par segment (VL, TPC, TPM, TPV, Garages et Concessions...). Le calcul du S/C écrêté 3 ans et la comparaison avec le PLR de chaque segment permet de classer les contrats selon les segments. Lorsque l'ancienneté du contrat ne permet pas de déterminer de S/C 3 ans, le S/C est calculé depuis la souscription. La direction technique automobile fixe alors le niveau de majoration en fonction des résultats de chaque segment en tenant compte des indicateurs "marché". Le souscripteur récupère ensuite cette préconisation qui va l'aider dans le renouvellement du contrat.

3.5 L'intérêt d'un nouvelle méthode tarifaire - Conclusion

En France, les tarifs d'assurance sont libres. L'actuaire peut presque librement segmenter son tarif. Même si des variables telles que le sexe ne permettent plus de différencier un tarif, l'explosion de la *data* permet d'enrichir les modèles d'une multitude d'autres variables pouvant être discriminantes telles que la marque d'une voiture, la zone du risque ou encore l'âge du véhicule.

Même si l'assurance repose intrinsèquement sur le principe de mutualisation des risques entre les assurés au sein d'une population de risques homogènes, l'univers concurrentiel du marché de l'assurance automobile pousse les assureurs à sophistiquer leur tarif afin d'avoir le modèle permettant d'estimer la prime pure d'un contrat le plus précisément possible. L'espérance mathématique de la charge annuelle ne suffit pas. Certains assureurs parlent même d'individualisation du tarif. C'est pourquoi la modélisation de la fréquence et du coût moyen est une étape essentielle dans l'élaboration d'un tarif. Il faut alors trouver le bon équilibre entre la segmentation et la robustesse des modèles statistiques liée à la mutualisation.

Or, le modèle actuellement en place au sein d'AXA concernant les flottes ouvertes ne permet pas d'atteindre la rentabilité attendue. La tarification actuelle, bien que se basant sur les fréquences et les coûts moyens des segments du portefeuille, voir du client selon les méthodes sélectionnées dans la calculette mise à disposition des souscripteurs, n'est pas concurrentielle. On peut en effet le constater par la très forte volatilité des fréquences et des coûts moyens par segment et par garantie. L'assuré, en recherche constante d'optimisation de son tarif, cherchera le meilleur prix. La clé réside donc dans la segmentation de la flotte qui peut être composée de véhicules de différentes catégories. La connaissance de l'immatriculation grâce à la mise en place du FVA ainsi que des différentes données qui s'y attachent prend tout son sens. L'idée principale étant de mettre en place un tarif véhicule par véhicule comme cela se fait sur les petites flottes dites fermées. Cette segmentation est loin de la segmentation ultime car la connaissance des classes de risques n'est pas parfaite mais elle permettra d'améliorer le tarif actuel pour lequel la segmentation n'est pas suffisante sans pour autant atteindre l'individualisation du tarif.

Il est quand même nécessaire de nuancer la mise en place de ce nouveau tarif car l'environnement de taux bas diminue l'escompte et augmente par conséquent le coût du risque. Cet environnement agit directement sur l'ECR qui croit à mesure que les taux diminuent et durcit les conditions de souscription. Le tarif sera donc toujours lié au contexte économique et au marché. Par ailleurs, la décomposition de l'ECR permet de ne modéliser que la charge hors grave que nous couplerons au modèle de fréquence et c'est ce que nous allons aborder dans le chapitre suivant. Afin d'obtenir un prime commerciale, il suffira alors d'appliquer à la prime pure, les différentes composantes de l'ECR afin d'avoir un tarif chargé de frais, de sinistralité grave, de sinistralité atypique et des autres charges, notamment financières.

Chapitre 4

Tarification et résultats de la modélisation

Avec la mise en place du FVA, la nouvelle connaissance de l'immatriculation permet d'intégrer une multitude d'informations liées à chaque véhicule. Ces nouvelles données seront étudiées selon leur importance dans les modèles et permettront l'établissement d'un tarif propre au périmètre des flottes ouvertes. Cependant, avant de s'intéresser à un modèle propre aux flottes ouvertes, il a paru essentiel de vérifier que le modèle actuellement en place pour les flottes fermées puisse être adapté sur les flottes de plus grande taille. En terme d'implémentation opérationnelle, cela faciliterait grandement l'intégration des flottes ouvertes car il s'agira d'appliquer le même tarif sans distinction du type du contrat (PARC ou FLOTTE) et c'est ce que nous essaierons de déterminer dans la première partie de ce chapitre.

Par la suite, nous tenterons d'établir un modèle propre aux flottes ouvertes afin de ne considérer que les variables les plus discriminantes, tout en tenant compte des contraintes opérationnelles de mise en place d'un tarif (nombre et disponibilité des variables par exemple). En revanche, il est important de préciser ici que la connaissance de l'immatriculation ne remonte qu'à 2019, date de la mise en vigueur du FVA. Par ailleurs, afin de ne pas biaiser les estimations et modèles de tarification, il a été nécessaire d'exclure l'année 2020 de nos données afin de ne pas prendre en compte la baisse de fréquence constatée sur les flottes automobiles et liée à la mise en place d'un confinement pendant la période de pandémie de la COVID. En conséquence, il faudra s'assurer de la stabilité des différents modèles afin qu'il n'y ait pas de sur-apprentissage suite à une quantité de données peu élevée selon les segmentations.

Enfin, dans la suite de ce travail et comme il a été précisé précédemment, nous tenterons de modéliser la charge sinistre hors graves puisque l'application des taux de graves, d'atypiques et de charges financières liés à la construction de l'ECR permettra de déterminer une prime commerciale.

4.1 Comparaison des modèles Parc et Flotte Ouverte

4.1.1 Description des variables et contexte de l'étude

Toutes les variables tarifaires du modèle actuellement en place concernant les parcs ne sont pas communes aux flottes ouvertes. Il a donc été nécessaire d'exclure certaines variables tarifaires afin d'avoir une comparaison exploitable des deux modèles. L'exclusion de ces variables n'a pas d'incidence sur la modélisation de la charge sinistre puisqu'il s'agit de variables impliquant l'utilisation d'un coefficient de majoration ou minoration tarifaire selon les cas.

Par ailleurs, le modèle sur le PARC étant déjà en place et les variables tarifaires déjà connues, nous ne nous intéresserons qu'à la modélisation de la prime pure afin de faciliter la lecture des résultats et de conserver le maximum de données dans une unique base. Nous avons donc modélisé la prime pure par un modèle de Tweedie de paramètre $p = 1,5$ afin de pouvoir modéliser la présence de nombreux sinistres avec une charge nulle. L'implémentation de ce modèle permet d'éviter la multiplication des variables explicatives induite par le modèle fréquence $\times$ coût moyen. Par ailleurs, le code INSEE de la ville du risque n'est pas pertinent sur les flottes ouvertes car il s'agit d'entreprises de taille importante et généralement multi-sites, pour lesquels le numéro SIRET et par conséquent le code INSEE, n'est pas connu. Nous avons donc exclu cette variable de la comparaison des deux modèles qui servaient à la construction d'un zonier pour le modèle PARC.

Les variables prises en compte dans le modèle utilisé sur le PARC et permettant d'estimer la prime pure sont :

- veh_ageveh : âge du véhicule, déduit de la variable veh_dtmcirc ;
- cnt_nafret : regroupement à partir de la variable cnt_naf qui correspond à l'activité de l'entreprise ;
- veh_cdgenrn : il s'agit du genre de véhicule qui dissocie les véhicules légers des véhicules utilitaires par exemple ;
- veh_carross : segmentation de la carrosserie du véhicule. On différencie par exemple les berlines, des breaks, des camionnettes ou encore des monospaces ;
- veh_pfisc : la puissance fiscale d'un véhicule, présente sur la carte grise, permet d'intégrer des informations sur le type de véhicule. En effet, un véhicule ayant une puissance fiscale importante peut être synonyme de voiture de sport par exemple ;
- veh_risqmarq : regroupement de la variable veh_libmar. Les marques telles que Audi, Porsche et Mercedes sont considérées comme à risque car les coûts de réparation peuvent être élevés.

Nous avons alors sélectionné les mêmes variables sur la base des données liées aux flottes ouvertes et sans distinction de leur facteur discriminant.

4.1.2 Comparaison des indicateurs

La première comparaison qu'il est possible de faire de ces deux modèles concerne le *spread* des variables. En effet, il est nécessaire de s'assurer que les variables utilisées dans le modèle PARC sont aussi discriminantes que pour le modèle FLOTTE.

C'est ce que nous pouvons constater sur la figure 4.1 ci-après où on remarque que tous les *spreads* 100/0% du modèle FLOTTE sont supérieurs à ceux du modèle PARC. Si on s'intéresse au *spread* 95/5%, la variable âge du véhicule semble être la variable la plus discriminante dans les deux modèles avec un spread de 310% pour les FLOTTES et de 288% pour les PARCS.

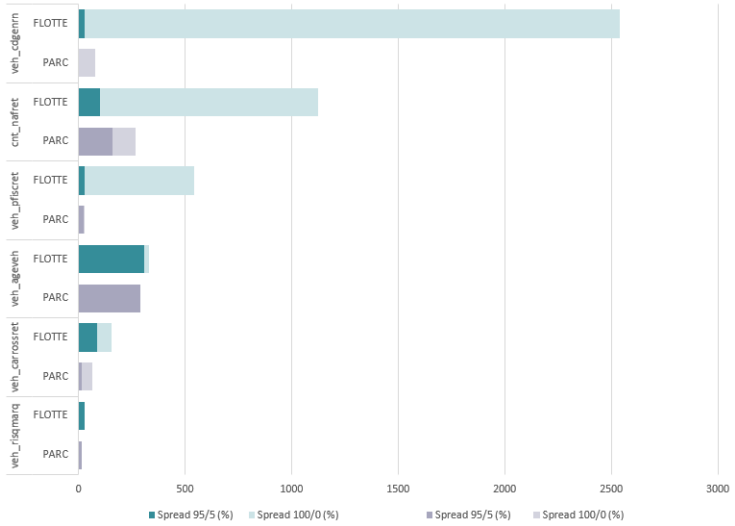


FIGURE 4.1 – Comparaison du *spread* PARC et du *spread* FLOTTE pour les variables du modèle

En revanche, toutes les variables ne contribuent pas de la même manière à chacun des modèles. En effet, même si le *spread* 95/5% de la variable représentant le code NAF de l'entreprise du modèle PARC est supérieur à celui des FLOTTES, le *spread* 100/0% pour cette même variable est bien inférieur à celui des FLOTTES. Ceci peut s'expliquer par la présence d'une plus grande hétérogénéité dans l'activité des entreprises ayant un contrat FLOTTE. En effet, les transports urbains de voyageurs (TRU) sont présents en grand nombre dans les flottes de grande taille, ce qui n'est pas le cas dans les flottes de petite taille et peut expliquer la différence entre le plus petit et le plus grand coefficient dans la modélisation de la prime pure. C'est ce que nous permet de constater le graphique 4.2 ci-dessous :

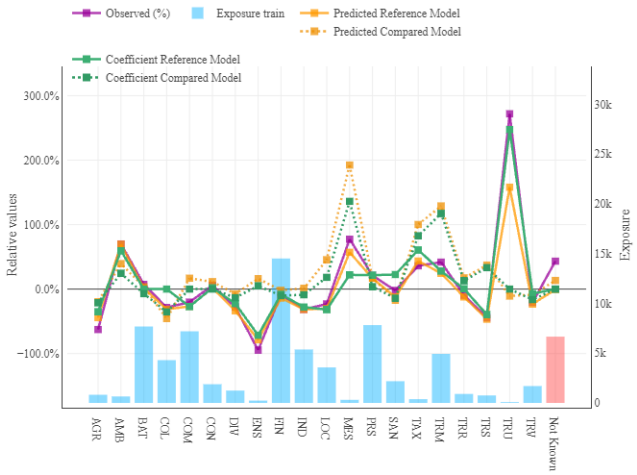


FIGURE 4.2 – Coefficients PARC (*Compared*) et FLOTTE (*Reference*) pour la variable activité de l'entreprise

Mise à part la variable représentant l'activité, l'ensemble des variables semblent apporter une bonne information au modèle FLOTTE puisque les *spreads* 95/5% sont supérieurs à ceux du modèle PARC.

Cette information est vérifiée par les courbes de Lorenz présentées sur la figure 4.3 ci-dessous où on constate que les courbes des modèles PARC et FLOTTE sont très proches. Cet indicateur graphique nous permet de calculer l'indice de Gini sur les bases tests qui est de 26,56% sur le modèle FLOTTE, légèrement supérieur à celui sur le PARC qui est de 24,46%.

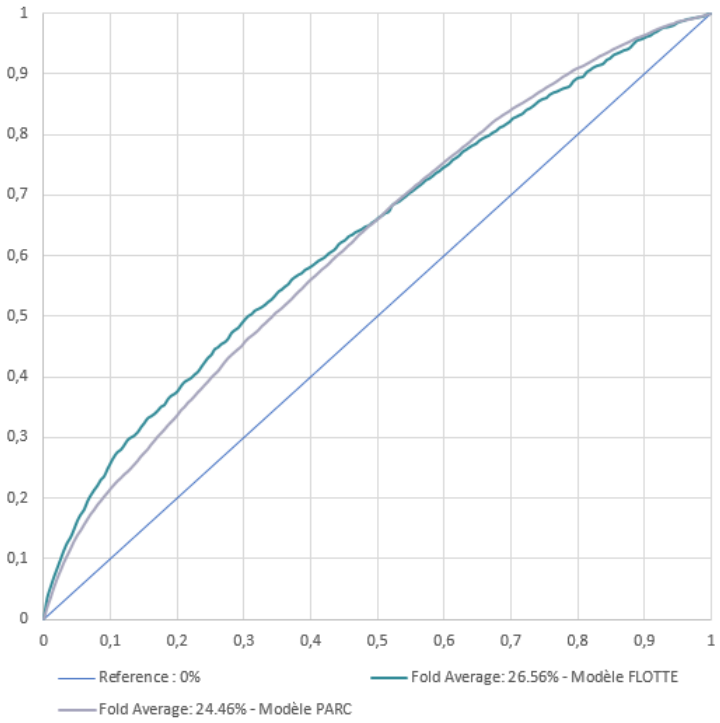


FIGURE 4.3 – Comparaison de la courbe de Lorenz des modèles PARC et FLOTTE

D'autres indicateurs statistiques, présentés dans le tableau 4.1 ci-après, permettent de confirmer que les modèles PARC et FLOTTE sont très proches, avec un léger avantage au modèle FLOTTE, que ce soit dans la base de modélisation ou de validation. L'écart entre la base de modélisation et la base de validation nuance malgré tout l'adéquation mais cela reste acceptable au regard de la quantité d'information :

TABLE 4.1 – Récapitulatif des indicateurs statistiques des modèles PARC et FLOTTE

Indicateurs	FLOTTE		PARC	
	Base de modélisation	Base de validation	Base de modélisation	Base de validation
Gini	30,54%	34,49%	25,09%	24,30%
RMSE	1 405	1 339	1 591	1 645
Déviance moyenne	72,71	72,58	77,98	78,01
MAE	273,6	273,7	306,9	304,4

Par ailleurs, en regardant l'histrogramme "3D" (*heatmap*) des résidus de déviance (figure 4.4), il est possible de confirmer que le modèle s'adapte relativement bien aux données des flottes ouvertes puisqu'on constate que la majorité des résidus sont centrés en 0. Les résidus négatifs s'expliquent par le fait de la présence d'un nombre important de risques sans sinistres, ce qui implique que leur prédiction soit supérieure à l'observation :

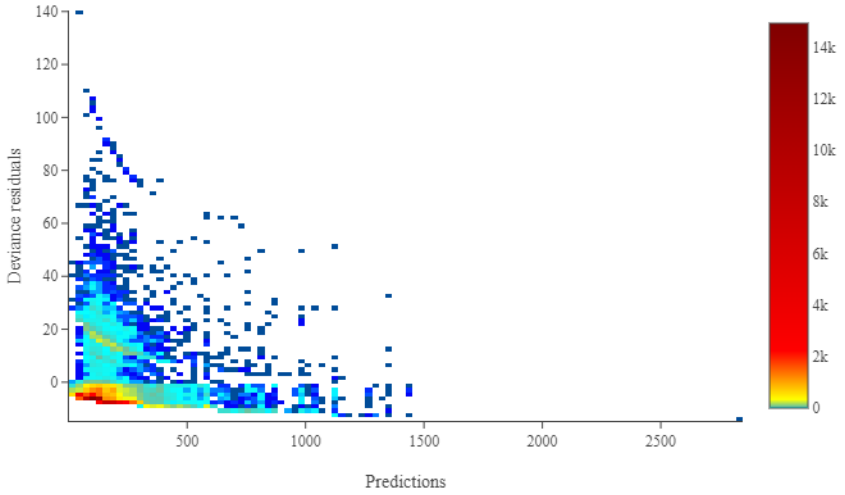


FIGURE 4.4 – Histogramme "3D" des résidus de la déviance

En effet, la *lift curve* présentée sur le graphique 4.5 met en exergue une possible sous-modélisation des contrats les moins sinistrés puisqu'on constate sur les premiers quantiles que la courbe jaune qui représente les prédictions est en dessous de la courbe violette qui représente les observations. On peut néanmoins voir que le modèle s'ajuste correctement aux données.

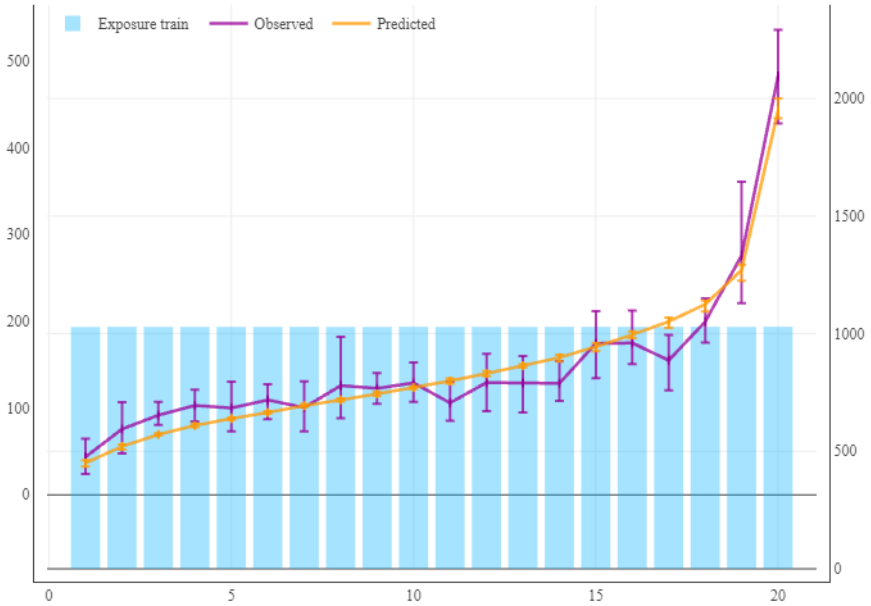


FIGURE 4.5 – *Lift curve* de la prime pure du modèle FLOTTE selon les variables du PARC

4.1.3 Résultats

On peut donc raisonnablement conclure que le modèle de prime pure actuellement en place sur le PARC pourrait être utilisé sur les FLOTTES, exception faite des variables exclues qui sont sans impact sur la modélisation puisqu'elles n'interviennent qu'après la modélisation effectuée sur le PARC en tant que coefficient de réduction ou de majoration tarifaire. Cette comparaison pourra peut-être permettre, sans évolutions majeures des outils de souscription du PARC, d'intégrer les FLOTTES et ainsi leur faire bénéficier d'un tarif plus précis, véhicule par véhicule.

4.2 Modèles propres aux Flottes Ouvertes

Dans l'éventualité où des budgets seraient alloués pour l'établissement d'un outil de souscription spécifique aux FLOTTES, il a été nécessaire d'étudier et de développer un modèle dédié en tenant compte de l'importance de l'ensemble des variables à notre disposition. Nous développerons dans cette partie, les étapes ayant mené à la sélection des variables discriminantes utilisées dans le modèle de fréquence $\times$ coût moyen et de prime pure que nous détaillerons dans un deuxième temps.

4.2.1 Analyse préliminaire

Variables à prédire

La première étape dans la construction d'un modèle GLM est l'étude des variables à prédire car, si pour la fréquence, l'étude du nombre de sinistres n'a pas grand intérêt, il n'en est pas de même pour la modélisation du coût moyen.

En effet, lors de leurs déclarations, de nombreux sinistres sont ouverts au forfait dans les outils de gestion et ne représentent pas réellement la charge effective pour l'assureur. Ces charges de sinistres forfaitaires peuvent induire une mauvaise adéquation aux lois utilisées lors de la modélisation du coût moyen du fait de la présence de pics de volumétrie engendrés par la sur-représentation de ces montants forfaitaires.

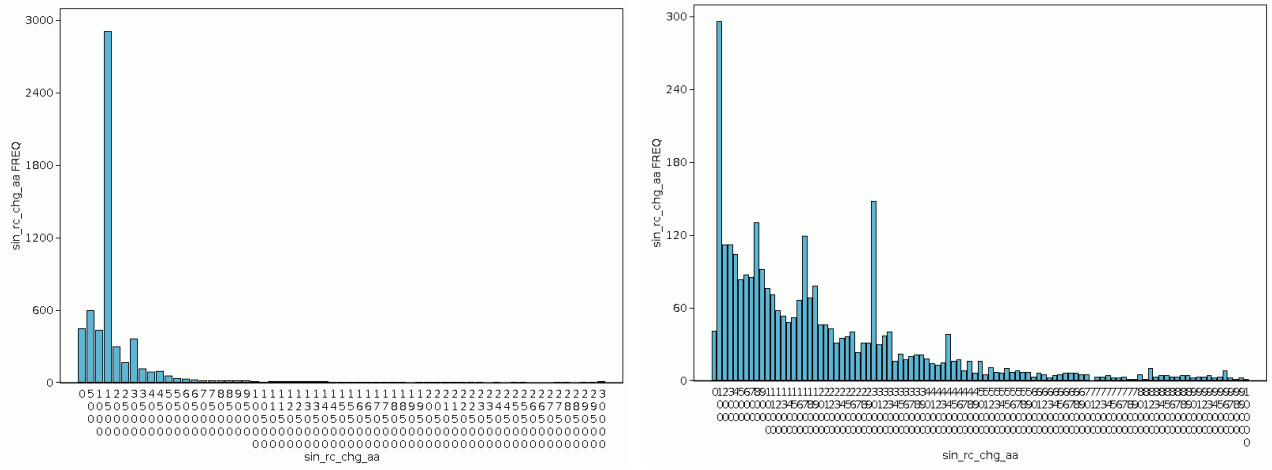
Afin d'éviter cette pollution dans l'étude de la charge, on effectue un premier filtre en utilisant une vision vieillie de 12 mois des sinistres. En plus de réduire le nombre des ouvertures au forfait, le vieillissement de la charge sinistre permet d'être au plus proche de la sinistralité réelle. Néanmoins, même si cette période peut suffire dans le règlement d'un sinistre, il arrive encore dans de nombreux cas, notamment lorsqu'il s'agit de sinistres corporels, que ces derniers ne soient pas clos et donc, toujours ouverts au forfait même un an après leur déclaration.

On distingue alors deux cas :

— Les ouvertures au forfait AXA. Il s'agit généralement des sinistres ouverts forfaitairement et en attente d'évaluation de la part d'un expert par exemple.

— Les ouvertures au forfait IRSA [1] et IRCA [2]. Un sinistre RC peut engendrer des recours entre assureurs selon la responsabilité conventionnelle de leur assuré (0%, 50% ou 100% responsable). En effet, l'application des conventions IRSA et IRCA permet aux assureurs qui l'ont signé, d'indemniser directement leur assuré sinistré puis d'effectuer un recours auprès de la compagnie de l'assuré responsable de l'accident. Les montants de recours sont forfaitaires et sont fixés chaque année par les conventions. À titre d'exemple, le montant forfaitaire pour un sinistre matériel 100% responsable à compter du 1^{er} janvier 2020 a été fixé à 1 568 €. Concernant les sinistres corporels sans invalidité, le montant forfaitaire IRCA a été fixé à 1 480 € au 1^{er} janvier 2020.

Un retraitement de ces charges forfaitaires a donc été effectué. C'est ce que nous montre le graphique 4.6 ci-dessous où l'on peut remarquer un pic de presque 3000 sinistres sur la figure 4.6a avant retraitements. Concernant la figure 4.6b après retraitements, on constate quelques pics de sinistres mais qui sont négligeables au regard de leur nombre :



(a) Avant retraitement des forfaits

(b) Après retraitement des forfaits

FIGURE 4.6 – Répartition des coûts moyens

Les ouvertures au forfait IRSA ou IRCA précédemment décrites peuvent entraîner la présence de sinistres ayant une charge négative une fois le recours exercé. Par ailleurs, la présence de contrats en conservation sur le périmètre des FLOTTES peut également apporter sa part de charges négatives. En effet, lors d'un montage en conservation, un montant de sinistre négatif est intégré dans les systèmes de gestion afin de représenter le montant de la conservation. Par la suite, chaque sinistre déclaré viendra en déduction de cette conservation, qui reste à la charge de l'assuré. Lorsque la charge des sinistres dépasse le montant en conservation, on récupère alors le montant de sinistres réellement à la charge de la compagnie d'assurance. En conséquence, il a été nécessaire d'exclure ces montants négatifs de la modélisation car les lois de Tweedie ou Gamma par exemple, ne peuvent modéliser de charges négatives. Il sera alors nécessaire de mener une étude propre à ces charges spécifiques et de les modéliser distinctement.

Enfin, dans le cadre de l'application des conventions IRSA et IRCA, nous avons tenté de modéliser les charges de sinistres responsables d'une part et les charges de sinistres non responsables d'autre

1. IRSA : Convention d'indemnisation directe de l'assuré et de recours entre sociétés d'assurance automobile
2. IRCA : Convention d'indemnisation et de recours corporel automobile

part. Cependant, au vue de la volumétrie de sinistres dont nous disposons après retraitements, la distinction des sinistres selon leur responsabilité n'a pas permis de rendre les résultats fiables et nous observions une très forte volatilité dans l'indice de Gini entre les bases de modelisation et de validation due à un sur-apprentissage des données de modélisation.

Étude des corrélations

La deuxième étape dans la construction des modèles consiste à la sélection des variables discriminantes par une étude des corrélations. Cette étape essentielle permet d'éviter la sélection de variables fortement corrélées entre elles et qui n'apporteraient pas suffisamment d'information complémentaire afin de décrire la charge hors graves. Il a donc été nécessaire d'étudier d'une part, les variables quantitatives et d'autre part, les variables qualitatives.

Variables quantitatives Le ρ de Spearman a été utilisé pour l'analyse des corrélations des variables quantitatives. Cette mesure à l'avantage de permettre l'étude de la dépendance entre deux variables sans que la relation qui les unit ne soit de type affine. En effet, plutôt que de s'intéresser aux corrélations des valeurs prises par les deux variables, la méthode du ρ de Spearman se concentre sur les rangs de ces variables. On définit alors le coefficient de corrélation de Spearman tel que :

$$\rho_{(X,Y)} = \frac{Cov[rg(X), rg(Y)]}{\sigma_{rg(X)}\sigma_{rg(Y)}}$$

Avec $Cov[rg(X), rg(Y)]$ la covariance des variables de rang X et Y et $\sigma_{rg(X)}$, $\sigma_{rg(Y)}$ les écarts-types associés.

La représentation sous forme de corrélogramme présenté sur la figure 4.7 permet alors de faciliter la sélection des variables quantitatives :

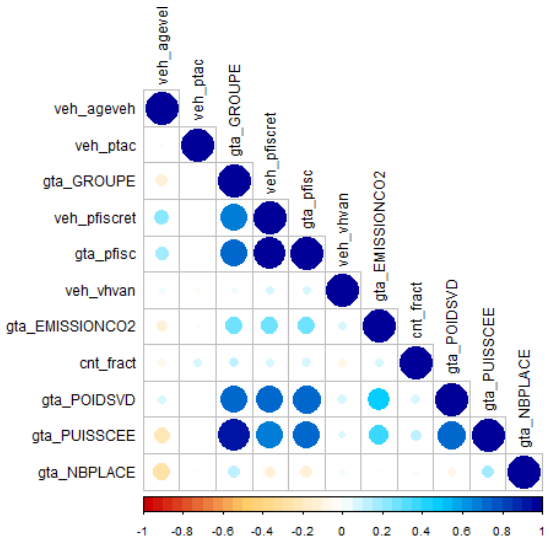


FIGURE 4.7 – Corrélogramme des variables quantitatives

En effet, l'étude de ces corrélations montre notamment la relation logique entre les puissances fiscales issues des données SRA d'une part et d'autre part, celles issues de la base véhicule. Nous avons donc sélectionné la variable ayant le meilleur taux de complétude pour cette étude. Par ailleurs, on constate également de fortes relations entre les variables "gta_GROUPE", "gta_pfisc", "veh_pfiscret", "gta_POIDSVD" et "gta_PUISSCEE". En effet, la variable "gta_GROUPE" a pour finalité d'aider les assureurs à tarifer la garantie RC. C'est un indicateur qui reflète la dangérosité intrinsèque des véhicules. Son calcul fait intervenir :

- La puissance réelle du véhicule ;
- Les masses du véhicule (poids à vide et PTAC) ;
- La vitesse maximale du véhicule ;
- Un indicateur estimant la sécurité globale du véhicule. Ce paramètre dépend de la conception et des équipements de sécurité du véhicule.

Il est donc tout à fait normal qu'elle soit corrélée au poids du véhicule ("gta_POIDSVD") mais également à la puissance d'un véhicule ("gta_PUISSCEE" ou "veh_pfiscret"). Il sera alors possible d'identifier les véhicules de type sport qui auront une puissance largement supérieure aux véhicules utilitaires par exemple. Ainsi, plus le groupe est élevé, plus la puissance de l'automobile sera élevée. En outre, le groupe permet également, en combinant les différents éléments cités précédemment, d'évaluer la dangérosité du véhicule en raison des dommages que peut infliger un véhicule de poids et de puissance supérieurs. La variable de groupe SRA permet donc a elle seule de tenir compte d'un maximum d'informations relatives au véhicule. C'est donc naturellement cette variable qui sera privilégiée dans l'élaboration d'un modèle propre aux FLOTTES, toujours selon son degré d'importance.

Variables qualitatives Concernant l'étude des corrélations des variables qualitatives, la méthode du V de Cramer a été retenue. Cette méthode permet la mesure d'association entre deux variables et indique la force avec laquelle elles sont liées. Ces mesures d'association sont calculées à partir d'une normalisation du test du χ^2 d'indépendance appliqué au tableau de contingence de deux variables qualitatives. La statistique d^2 du χ^2 calcule alors l'écart entre les valeurs observées et les valeurs attendues en cas d'indépendance. Elle est définie par :

$$d^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(N_{ij} - \frac{N_{i.}N_{.j}}{n}\right)^2}{\frac{N_{i.}N_{.j}}{n}}$$

Où :

- N_{ij} représente le nombre d'occurences communes aux modalités x_i et y_j des variables qualitatives X et Y respectivement ;
- $N_{i.} = \sum_{j=1}^{l} N_{ij}$ et $N_{.j} = \sum_{i=1}^{k} N_{ij}$;
- k et l sont respectivement le nombre de modalités de X et Y ;
- n est le nombre total d'observations.

Le test du χ^2 d'indépendance permet alors de tester :

$$\begin{cases} H_0 : \text{"La fréquence est identique sur les deux variables"} \\ H_1 : \text{"La fréquence diffère selon les modalités"} \end{cases}$$

Sous H_0, $d^2 \sim \chi^2_{(k-1)(l-1)}$. Ainsi, pour un test d'ordre $(1-\alpha)$, si $q_{(1-\alpha)}(\chi^2_{(k-1)(l-1)})$ représente le quantile d'ordre $(1-\alpha)$ de la loi du χ^2 à $(k-1)(l-1)$ degrés de liberté, on ne rejette pas H_0 lorsque

$d^2 < q_{(1-\alpha)}(\chi^2_{(k-1)(l-1)})$. Cela signifie que les variables sont considérées comme indépendantes. Dans le cas contraire, on peut considérer qu'il existe une relation significative entre les deux variables.

On peut alors définir le V de Cramer tel que :

$$V = \sqrt{\frac{d^2}{n \times [min(k,l) - 1]}}$$

La valeur de V varie entre -1 et 1. Les deux variables qualitatives étudiées seront considérées indépendantes lorsque V est proche de 0.

Cette méthode a donc été appliquée aux variables qualitatives susceptibles de convenir au modèle et nous obtenons ainsi les relations présentées dans le corrélogramme 4.8 ci-après :

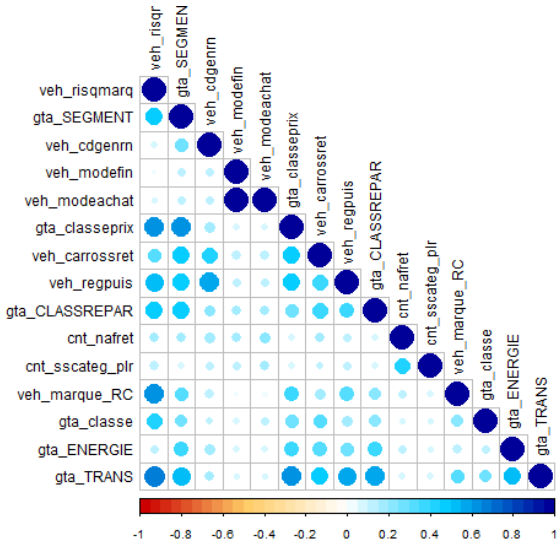


FIGURE 4.8 – Corrélogramme des variables qualitatives

On constate facilement la relation triviale entre les variables de mode d'achat ("veh_modeachat") et de mode de financement du véhicule ("veh_modefin") puisqu'elles sont issues de la même source de données mais incluent des segmentations différentes. Dans l'hypothèse où ces variables seraient discriminantes pour le modèle, il faudra n'en sélectionner qu'une seule afin de ne pas polluer le modèle d'informations redondantes.

Par ailleurs, on remarque également des corrélations relativement marquées concernant les variables "gta_classeprix", "veh_risqmarq" et "gta_segment". La logique est également respectée ici aussi car la classe de prix SRA est un indicateur en cas de perte totale ou de vol de véhicule. Elle prend en compte la valeur à neuf du véhicule lors de sa mise sur le marché. Des tranches de prix sont alors établies et actualisées chaque année en fonction de l'évolution de l'indice INSEE du prix des véhicules neufs. S'il s'agit d'un excellent indicateur pour les garanties liées aux dommages, incendie, vol et bris de glace, elle n'a que peu d'intérêt dans la modélisation de la garantie RC exception faite de la garantie RC matérielle. La valeur à neuf d'un véhicule peut renseigner sur le modèle du véhicule et ainsi influer sur la variable "veh_risqmarq" qui classe les véhicules selon les marques dans deux catégories, les véhicules dits "risqués" et les véhicules "standards". On peut donc logiquement

s'attendre à ce que les véhicules les plus chers soient catégorisés "véhicules risqués". Enfin, le segment du véhicule indique le positionnement de chaque modèle sur le marché automobile et catégorise les véhicules en fonction de leur taille ou de l'usage que l'on en fait. À ce titre, cette segmentation qui permet de distinguer les véhicules dits "urbains", des "citadines" ou encore des "berlines", est également liée à la classe de prix des véhicules.

Enfin, on constate également des relations positivement marquées entre les variables transmission du véhicule ("gta_TRANS"), classe de prix SRA ("gta_classeprix") et classe de réparation SRA ("gta_CLASSREPAR"). Par ailleurs, la variable évoquant la transmission classe les véhicules selon leur mode de transmission. On distingue alors les véhicules ayant quatre roues motrices ou quatre roues permanentes. Il s'agit généralement de véhicules "4x4". On distingue également les véhicules à propulsion des véhicules à traction. Dans le premier cas, il s'agit principalement de voitures sportives puissantes procurant une meilleure motricité à l'accélération. Dans le deuxième cas, il s'agit de la transmission qui équipe une grande majorité des véhicules présents sur le marché. Il est par conséquent logique d'avoir une certaine corrélation avec la classe de prix SRA du véhicule ou la classe de réparation SRA du véhicule. En effet, la classe de réparation SRA est un indicateur estimant le coût de la réparation d'un véhicule. Il fait intervenir le coût pondéré de remplacement du panier de pièces ainsi que le coût des chocs avant et arrière à 15km/h. Cet indicateur est réactualisé chaque année en fonction de l'évolution de l'indice SRA du coût de la réparation. On peut alors observer une corrélation positive entre ces variables due à la relation qui lie les véhicules dont le prix est élevé (4x4, voitures de sport) à leur mode de transmission particulier mais également au prix élevé des réparations pour ce type d'automobiles.

Cette étude préliminaire a donc permis la sélection des variables les plus discriminantes sans apporter d'informations redondantes qui pénaliseraient les résultats et la robustesse des modélisations, que ce soit dans le modèle de fréquence $\times$ coût moyen ou dans le modèle de prime pure, que nous allons décrire dans la suite de ce chapitre.

4.2.2 Modèle de Fréquence

Plusieurs modèles de fréquence ont été testés. En effet, nous avons tenté de modéliser la fréquence des sinistres RC selon leur degré de responsabilité. Cependant, la base des sinistres étant peu conséquente, des résultats peu fiables ont été obtenus. On observait effectivement de fortes disparités entre les bases de modélisation et de validation ne nous permettant pas de conclure sur la robustesse du modèle. De la même manière, il aurait été possible d'étudier la fréquence des sinistres selon une certaine garantie telle que la responsabilité civile corporelle ou la responsabilité civile matérielle mais là encore, le faible volume des sinistres dissociés selon la garantie impliquée, ne nous permettait pas de conclure et de valider ce modèle. Nous présenterons donc dans cette partie, les résultats du modèle de fréquence prenant en compte l'ensemble des sinistres RC et permettant d'obtenir de meilleurs résultat, sans distinction de leur degré de responsabilité. Il sera possible dans un futur proche et, lorsque les données seront suffisantes, d'effectuer une modélisation spécifique des sinistres en fonction de leur responsabilité.

Traditionnellement sur les risques IARD, la modélisation du nombre de sinistre s'effectue en appliquant une pondération par le temps de présence de l'assuré (ici le véhicule) pendant lequel il a bénéficié de la protection assurantielle. Le temps de présence est ainsi ramené en jour et permet de

pondérer chaque survenance de sinistre par "l'année-police" du véhicule associé. Une modélisation de la fréquence des sinistres par la loi de Poisson a donc été utilisée faisant intervenir la fonction de lien canonique "Log-Poisson".

Sélection des variables

En utilisant les méthodes de sélection des variables tarifaires décrites dans le chapitre 3, il a été possible, par le biais d'un outil de modélisation, de construire une *grid search* mettant en lumière les différents modèles créés à partir de la sélection des variables discriminantes. On observe sur la figure 4.9 que le nombre de variables correspond à l'axe des abscisses et que l'indice de Gini, indicateur de performance des modèles, est représenté sur l'axe des ordonnées. L'intervalle d'erreur représente les variations de performance entre les différents *folds*.

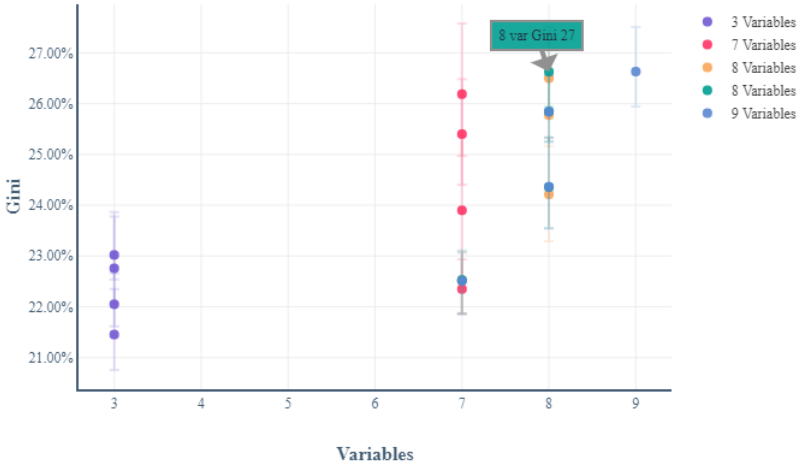


FIGURE 4.9 – *Grid search* des modèles de fréquence testés

Il s'agit donc de sélectionner le modèle ayant un indice de Gini élevé tout en tenant compte du nombre de variables afin d'éviter de créer un modèle avec un nombre de variables conséquent qui serait difficile à implémenter dans les outils de souscription d'une part et pour lequel la recherche d'informations auprès du client serait fastidieuse d'autre part. La facilité de souscription pour le client est gage de souscription souvent réussie. On remarque alors que les modèles à 7, 8 et 9 variables ont un indice de Gini supérieur aux modèles à 3 variables. Le modèle à 8 variables a donc été choisi car il possède un indice supérieur à celui à 7 variables et que le modèle à 9 variables n'apporte pas d'information complémentaire statisfaisante par rapport au modèle à 8 variables.

Les 8 variables utilisées dans la modélisation de la fréquence des sinistres RC sont présentées selon leur importance et leur *spread* sur la figure 4.10 ci-après :

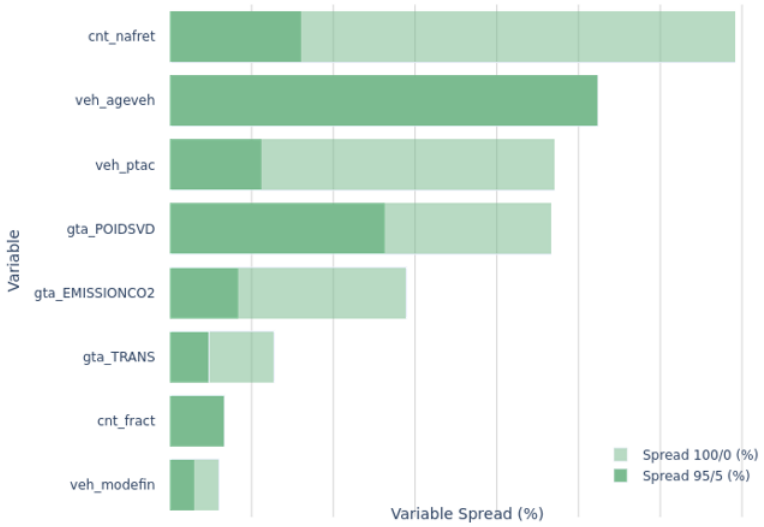


FIGURE 4.10 – Variables utilisées dans la prédiction des fréquences

La majorité de ces variables sont des informations présentes sur la carte grise du véhicule et ne nécessitent pas d'efforts particuliers de la part du client afin de les fournir. En effet, l'âge du véhicule peut se déduire à partir de la date de la première mise en circulation, le type de transmission peut se déduire du numéro CNIT [1], la puissance fiscale, le poids à vide et le rejet de CO_2 sont des informations présentes en lecture directe sur la carte grise. Seules les informations sur le fractionnement du contrat ainsi que sur le code NAF de l'entreprise seront à fournir par le client.

On peut noter ici que, même si la variable âge du véhicule est la variable la plus discriminante du modèle de fréquence, la variable qui renseigne sur le code NAF de l'entreprise possède la plus grande influence sur le modèle. Enfin, on retrouve dans ces variables, certaines variables présentes dans le modèle PARC décrit dans la première partie de ce chapitre.

Validation du modèle

Une fois la sélection des variables effectuée, il faut maintenant s'intéresser à la validation du modèle. Nous allons par conséquent analyser les différents indicateurs statistiques à notre disposition, à la fois sur la base de validation mais également sur la base de modélisation afin de s'assurer de l'homogénéité entre ces dernières.

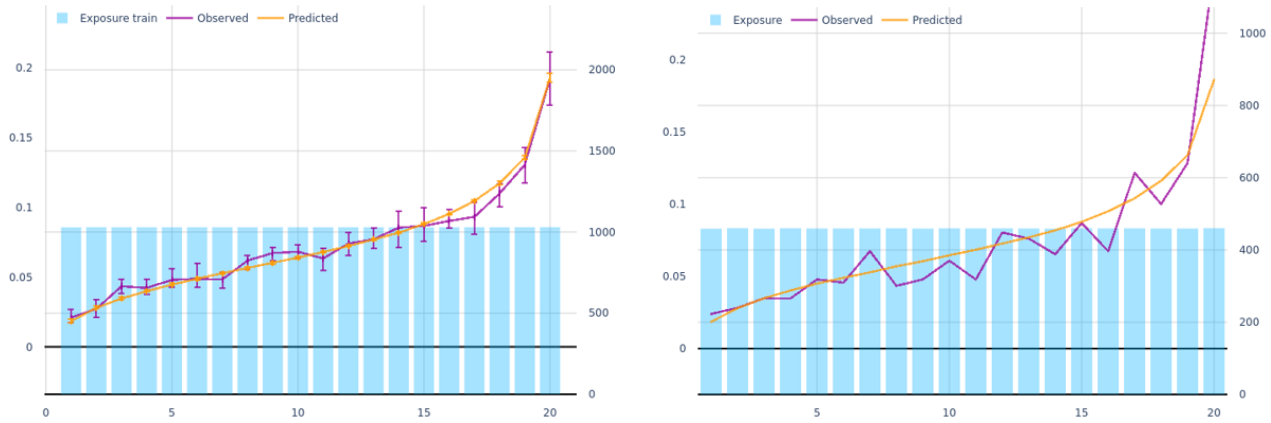
Indicateurs statistiques Le tableau 4.2 ci-après permet de constater que la fréquence moyenne observée sur la base modélisation est relativement proche de la fréquence observée sur la base de validation. Par ailleurs, même si l'indice de Gini de la base de validation est légèrement supérieur à celui de la base de modélisation, on remarque que les RMSE, déviances moyennes et MAE sont très proches. On peut donc considérer que le modèle est stable et l'appliquer à la base de validation.

1. CNIT : Code National d'Identification du Véhicule

TABLE 4.2 – Indicateurs statistiques sur les bases de modélisation et de validation du modèle de fréquence

Indicateurs	Base de modélisation	Base de validation
Fréquence moyenne observée	7,405%	7,366%
Fréquence moyenne prédite	7,405%	7,372%
Gini	29,9%	31,26%
RMSE	0,3562	0,3546
Déviance moyenne	0,4105	0,4074
MAE	0,1368	0,1360

Lift Curve Les *lift curves* présentées sur la figure 4.11 ci-dessous permettent de confirmer une bonne adéquation des prédictions sur la base de modélisation (4.11a). La *lift curve* s'ajuste un peu moins aux données de la base de validation (4.11b) mais ceci s'explique par le fait que la base de validation ne contient que 20% des données totales et la fréquence peut paraître très instable, en particulier sur le dernier quantile où on observe une sous-représentation du risque avec de nombreux sinistres. En revanche, on constate que la fréquence prédite suit bien la tendance sur la base de validation.



(a) Base de modélisation

(b) Base de validation

FIGURE 4.11 – *Lift curves* du modèle de fréquence

Analyse des résidus Afin de juger de l'adéquation au modèle de Poisson pour l'analyse de la fréquence, une étude des résidus doit être menée. Une méthode consiste à utiliser les résidus quantiles randomisés définis par Dunn & Smyth (1996) tels que :

$$r_i = \Phi^{-1}(F(y_i; \hat{\mu}_i, \hat{\phi}))$$

Où :

- Φ est la fonction de répartition de la distribution de la loi $\mathcal{N}(0,1)$;
- F est la fonction de répartition d'une distribution continue quelconque ;
- $\hat{\mu}$ est l'estimateur du maximum de vraisemblance du paramètre de la moyenne ;
- ϕ est le paramètre de dispersion d'une famille exponentielle qui est estimé par la méthode du maximum de vraisemblance.

Par ailleurs, l'hypothèse de normalité des résidus inclut les distributions appartenant à la famille exponentielle. En particulier, les résidus quantiles randomisés transformés de la loi de Poisson, qui appartient à la famille exponentielle, sont normalement distribués. C'est ce que nous pouvons remarquer sur l'histogramme "3D" sur la figure 4.12 ci-dessous qui représente la distribution des résidus quantiles randomisés de la modélisation de la fréquence par une loi de Poisson. La dispersion des résidus est bien centrée en 0 et comprise entre -2 et 2 selon l'axe des ordonnées.

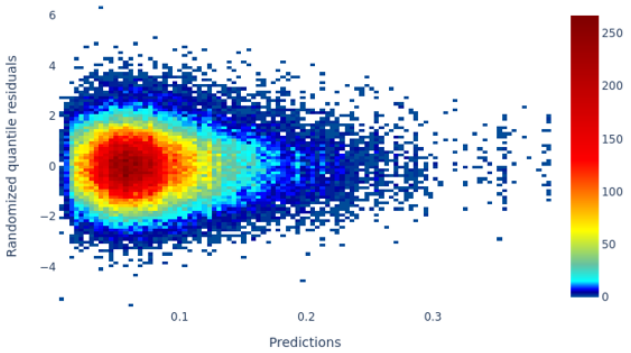


FIGURE 4.12 – Histogramme "3D" des résidus quantiles randomisés du modèle de fréquence

D'autre part, l'analyse des résidus agrégés est régulièrement utilisée pour les modèles de fréquence ou de prime pure car ils permettent de prendre en compte la présence de nombreuses valeurs nulles dans la variable à prédire. La présence de ces valeurs peut en effet biaiser l'analyse des résidus classiques car on peut alors constater que les résidus sont négatifs pour les valeurs nulles de la variable à prédire. Néanmoins, cela ne permet pas de dire que le modèle ne s'ajuste pas correctement aux données. En effet, si on s'intéresse à la modélisation d'une prime pure par exemple, les résidus d'un individu n'ayant aucun sinistre seraient négatifs. Or, le seul fait de l'absence de sinistres ne permet pas de justifier une prime nulle. Afin de corriger cette contradiction, on détermine les résidus agrégés qui correspondent aux résidus de chaque groupe de données ayant une prédiction similaire. Ils sont définis par :

$$r_{A_{groupe}} = \frac{\sum_{i \in groupe} (\text{valeur observée}_i - \text{valeur prédite}_i)}{\sqrt{Var\left(\sum_{i \in groupe} \text{valeur prédite}_i\right)}}$$

En particulier, il est possible d'observer les résidus de déviance agrégés sur la figure 4.13 ci-après qui nous permet de conclure sur l'absence de bimodalité de la fréquence et nous conforte dans l'idée que le modèle s'adapte correctement aux données :

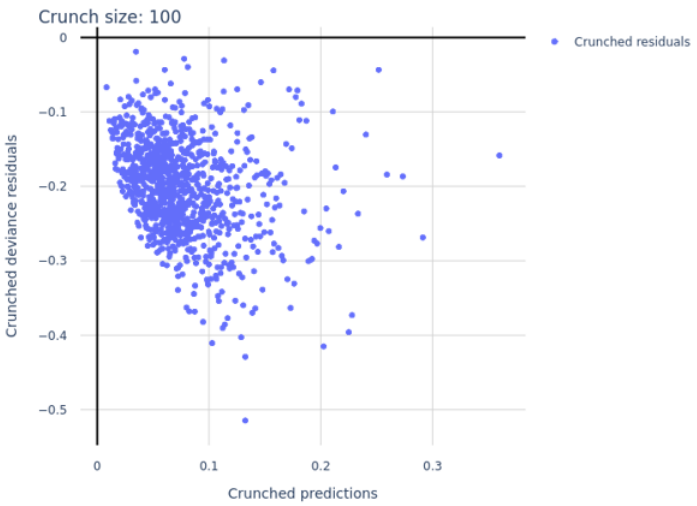


FIGURE 4.13 – Résidus de déviance agrégés pour le modèle de fréquence

4.2.3 Modèle de Coût Moyen

De la même manière que pour le modèle de fréquence, nous avons tenté de dissocier les sinistres responsables des sinistres non responsables. Cependant, la base des sinistres se trouvant réduite après le retraitement des différentes charges spécifiques décrites dans le paragraphe 4.2.1, il n'a pas été possible d'obtenir de résultats fiables et robustes. Il a donc été nécessaire de considérer l'ensemble des sinistres sans distinction du caractère responsable de chacun d'entre eux. De plus, les retraitements des ouvertures forfaitaires ont été fait en tenant compte de la responsabilité de chaque sinistre. En effet, pour chaque forfait IRSA ou IRCA 100% responsable, l'équivalent 50% responsable a également été retraité, ce qui a considérablement diminué le volume de données. Néanmoins, cette action permettra de modéliser uniquement la charge propre à AXA. Une étude sur la propension des sinistres forfaitaires devra être menée par la suite. Concernant l'analyse du coût moyen, nous avons donc utilisé une loi Gaussienne Inverse comme cela se fait usuellement, afin de modéliser la charge de sinistres hors graves strictement positive.

Sélection des variables

En utilisant la même méthode de sélection de variables décrite précédemment, la *grid search* présentée sur le graphique 4.14 ci-après nous permet de constater que l'ensemble des modèles testés ont un indice de Gini relativement faible par rapport au modèle de fréquence. Par ailleurs, on remarque également que les modèles testés les plus efficients ne comportent que peu de variables explicatives. En effet, la majorité des modèles testés avec un plus grand nombre de variables ne nous permettait pas de conclure sur la bonne adéquation aux données car on observait alors un sur-apprentissage de la base de modélisation par rapport à la base de validation. Nous avons donc choisi le modèle à 2 variables dont l'indice de Gini est le plus élevé et pour lequel nous observions une certaine stabilité dans les indicateurs statistiques.

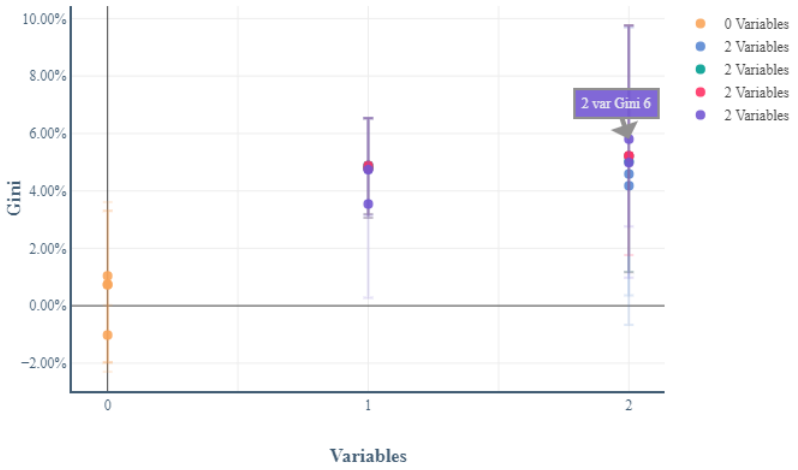


FIGURE 4.14 – *Grid search* des modèles de coût moyen testés

Les variables sélectionnées sont présentées sur la figure 4.15 ci-dessous qui permet d'apprécier l'importance de chaque variable. On remarque que les deux variables ayant le plus d'importance ne sont pas des variables utilisées dans le modèle de fréquence. En effet, la puissance fiscale permet de différencier le type de véhicule. On peut s'attendre à ce qu'un véhicule ayant une puissance fiscale plus élevée soit plutôt de type véhicule de sport ou de type 4x4 et par conséquent, aura des coûts de réparation plus élevés par exemple. Par ailleurs, le genre SRA du véhicule permet de dissocier les véhicules dits particuliers des véhicules utilitaires. On peut alors s'attendre à ce que les coûts des sinistres des véhicules utilitaires soient plus élevés que ceux des véhicules de particuliers. Ces deux variables semblent donc pouvoir expliquer la distribution du coût moyen des sinistres. D'autre part, il est généralement plus difficile de modéliser le coût moyen par rapport à la fréquence des sinistres car nous disposons de moins de données après les retraitements des charges spécifiques. De plus, le coût moyen étant une variable continue, il est plus compliqué de s'ajuster précisément à l'ensemble des coûts qui est très volatile. Ceci explique le fait qu'il y a souvent moins de variables explicatives dans les modèles de coût moyen.

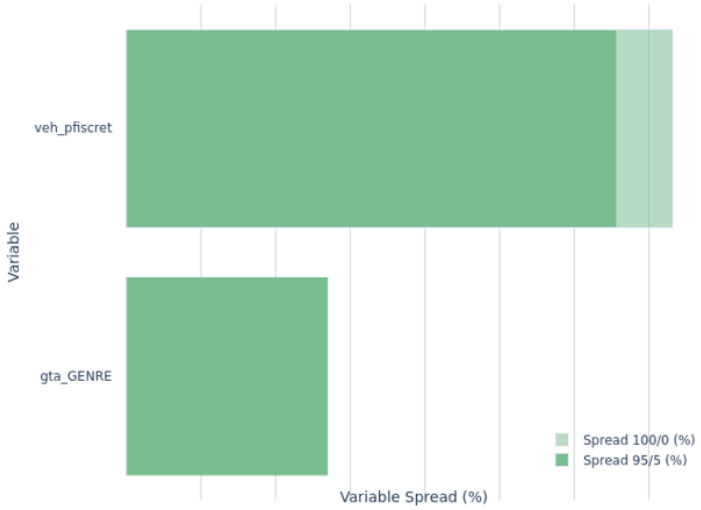


FIGURE 4.15 – Variables utilisées dans la prédiction des coût moyens

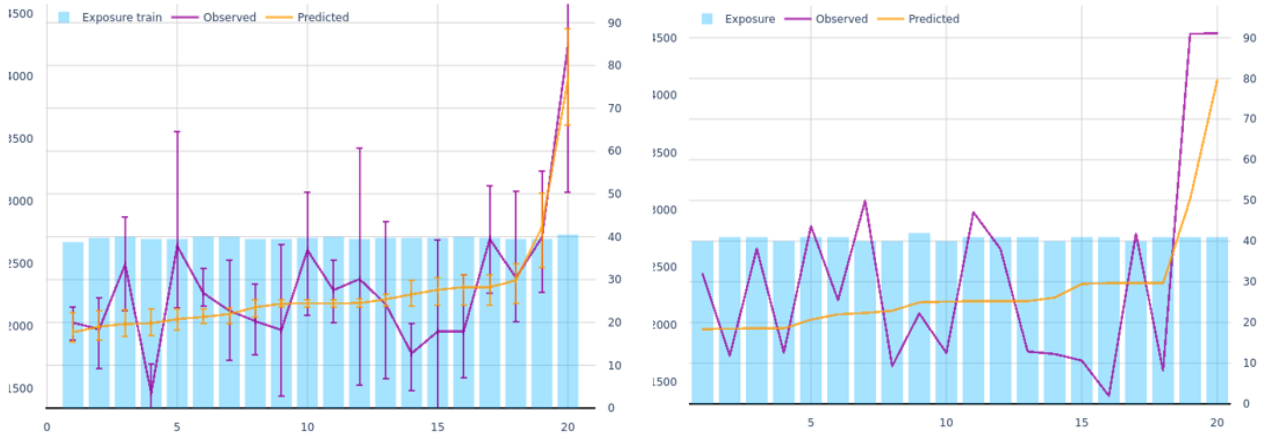
Validation du modèle

Indicateurs statistiques Au delà du fait que l'indice de Gini n'est pas d'une très grande qualité concernant le modèle de coût moyen, il est quand même nécessaire d'étudier les autres indicateurs statistiques sur les bases de modélisation et de validation afin de s'assurer de la stabilité du modèle. Le tableau 4.3 ci-dessous permet de s'assurer que les données sont homogènes entre les bases de modélisation et de validation car les coûts moyens observés sont du même ordre de grandeur dans les deux bases. Par ailleurs, nous ne constatons pas d'écarts significatifs entre les autres indicateurs statistiques. L'écart constaté sur l'indice de Gini entre les bases de modélisation et de validation est tout à fait acceptable.

TABLE 4.3 – Indicateurs statistiques sur les bases de modélisation et de validation du modèle de coût moyen

Indicateurs	Base de modélisation	Base de validation
Coût moyen observé	2 309	2 397
Coût moyen prédit	2 285	2 308
Gini	9,76%	7,13%
RMSE	3 224	3 461
Déviance moyenne	0,003565	0,006805
MAE	1 888	1 970

Lift curve L'analyse des *lift curves* présentes sur la figure 4.16 permet de constater l'extrême volatilité des coûts moyens induite par le faible nombre de données présentes dans la base retraitée des charges forfaitaires et des sinistres négatifs. Par ailleurs, même si les coût moyens observés sont similaires entre la base de modélisation et de validation, on constate que la comparaison entre les valeurs observées et les valeurs prédites est erratique.



(a) Base de modélisation

(b) Base de validation

FIGURE 4.16 – *Lift curves* du modèle de coût moyen

Analyse des résidus L'analyse des résidus de quantiles randomisés représentés sur l'histogramme "3D" ci-dessous (figure 4.17) confirme la mauvaise adéquation du modèle de coût moyen aux données. En effet, nous constatons que la distribution des résidus n'est pas caractéristique d'une distribution de la loi normale par rapport à l'axe des ordonnées.

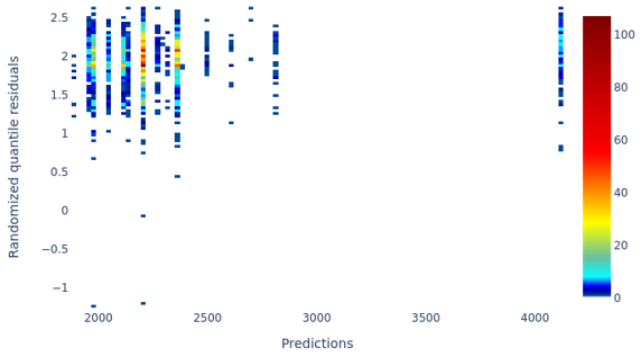


FIGURE 4.17 – Histogramme "3D" des résidus quantiles randomisés du modèle de coût moyen

4.2.4 Aggrégation des modèles de fréquence et de coût moyen

L'approche fréquence × coût moyen repose sur les modélisations distinctes de la fréquence de survenance d'un sinistre d'une part et d'autre part, du coût moyen d'un sinistre lorsqu'il y a survenance. Cette distinction met en lumière plusieurs inconvénients.

En effet, cette approche n'est permise qu'en effectuant une hypothèse très forte d'indépendance entre les fréquences et le coût moyen des sinistres. Par ailleurs, le modèle multiplicatif d'estimation de la prime pure par le biais de la fréquence et du coût moyen peut engendrer une multiplication de variables discriminantes. C'est ce que nous avons pu constater dans les deux modèles décrits précédemment. Les variables utilisées dans le modèle de coût moyen ne sont pas contenues dans le modèle de fréquence. De plus, afin de ne pas biaiser la segmentation, il faut ajouter aux contraintes les retraitements nécessaires et obligatoires dans le cadre d'une modélisation indépendante du coût moyen tels que les charges nulles, négatives ou encore les ouvertures aux différents forfaits AXA, IRSA et IRCA. Ces retraitements ont considérablement réduit la base de données disponibles, d'autant plus que nous n'avions qu'un historique d'un an de sinistralité consécutif à la mise en place récente du FVA et de la non prise en compte de la sinistralité atypique de 2020. Enfin, les valeurs à prédire pour le coût moyen étant très hétérogènes, même dans une base de données peu conséquentes, il est difficile de pouvoir trouver un modèle s'ajustant parfaitement aux données. Dans notre cas, le montant minimum d'un sinistre était de 0,38 € et variait jusqu'à un montant maximum de 30 000 €.

Pour ces raisons particulières, l'approche en méthode fréquence × coût moyen n'est peut être pas la plus adaptée. Néanmoins, cette approche aurait été plus fiable si nous disposions de plus de données concernant la sinistralité. C'est ce que nous confirme la *lift curve* du modèle agrégé de fréquence et de coût moyen présentée sur la figure 4.18 ci-après où l'on constate un écart important sur la majorité des quantiles entre la charge prédite et la charge observée. On observe en effet que la prime pure prédite est supérieure à l'observé :

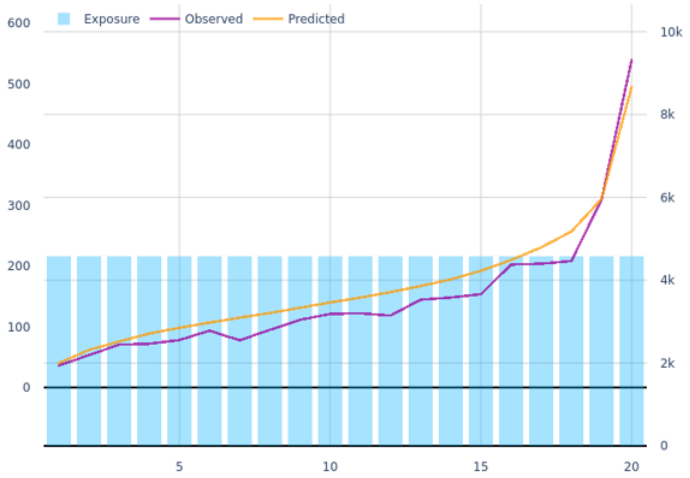


FIGURE 4.18 – *Lift curve* du modèle de fréquence × coût moyen

4.2.5 Modèle de Prime Pure

Une approche permettant d'éviter de segmenter les bases de données en base de fréquence d'une part, et base de charges sinistres d'autre part, consiste à utiliser un modèle de prime pure. En effet, cette approche a l'avantage de permettre de modéliser une distribution continue pour des valeurs supérieures à 0 tout en tenant compte d'une éventuelle masse en 0 que pourraient représenter les contrats ayant une fréquence nulle ou une charge nulle de sinistre par exemple.

Par ailleurs, le modèle de fréquence × coût moyen étant multiplicatif, il peut arriver que le nombre de variables soit important. En effet, dans l'estimation de la prime pure par l'approche utilisant la fréquence et le côut moyen, il faut considérer à la fois les variables issues du modèle de fréquence d'une part et les variables discriminantes du modèle de coût moyen d'autre part, tout en sachant que ces deux ensembles de variables peuvent tout à fait être disjoints. En modélisant directement la prime pure, on peut alors réduire le nombre de variables utilisées dans la tarification mais également s'affranchir du retraitement des charges nulles opéré dans la base utilisée pour les coûts moyens.

Dans cette partie, nous allons présenter l'approche prime pure faisant intervenir la distribution de Tweedie décrite dans le chapitre 3 et utilisée afin de modéliser la charge hors graves des sinistres flottes ouvertes.

Sélection des variables

De la même manière que sur les modèles de fréquence ou de coût moyen, nous procédons à une sélection des variables ayant la plus grande influence dans la modélisation de la prime pure tout en tenant compte de leurs différentes corrélations.

La *grid search* de la figure 4.19 ci-après permet de comparer les différents modèles testés en fonction du nombre de variables discriminantes et de leur indice de Gini respectif. Les différents intervalles d'erreurs qui représentent les performances entre les *folds* permet d'apprécier de la robustesse des

modèles associés. Notre sélection se porte donc naturellement vers les modèles à 10 et 12 variables qui sont relativement proches en terme de performance par rapport aux modèles à 3, 5 ou 11 variables. Le modèle à 10 variables ayant l'indice de Gini le plus élevé a notre préférence car les deux variables supplémentaires du modèle à 12 variables ne semblent pas apporter beaucoup plus d'information. On peut également remarquer que le modèle en place sur les PARCS est, lui aussi, un modèle à 10 variables, hormis la variable de zonier. Dans le cadre d'une future implémentation dans les outils de souscription, il n'y aurait pas de nécessité à développer de nouvelles cases tarifaires, ce qui peut s'avérer très couteux.

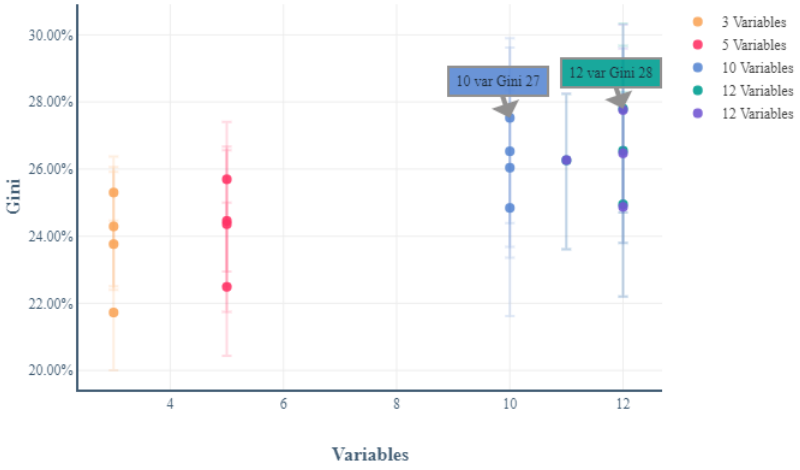


FIGURE 4.19 – *Grid search* des modèles de prime pure testés

Les 10 variables issues du modèle sélectionné sont présentées sur la figure 4.20 ci-après représentant le *spread* de chacune d'entre elles. On remarque sans surprise que la variable âge du véhicule est la variable la plus discriminante avec un *spread* de 300% que ce soit dans le cas 100/0% ou 95/5%. La deuxième variable la plus discriminante est la variable "cnt_nafret" qui renseigne sur l'activité principale de l'entreprise ayant souscrit un contrat FLOTTE. Les activités étant variées, cela peut entraîner une variation de *spread* 100/0% par rapport au *spread* 95/5%. En revanche, une telle différence ne peut trouver sa source que dans une valeur atypique sur un code NAF donné. La classe de réparation SRA est également un excellent indicateur de la prime pure. En effet, l'association SRA définit un classement des véhicules en fonction des tranches de coût de réparation. Ces coûts de réparation sont ensuite réactualisés chaque année en fonction de l'évolution de l'indice du coût de la réparation. On peut donc imaginer qu'un véhicule ayant un coût de réparation élevé aura mécaniquement une prime pure plus élevée. Par ailleurs, comme pour la modélisation de la fréquence, nous retrouvons également les variables renseignant sur le taux d'émission de CO_2, sur le mode de financement et sur le fractionnement du contrat. Enfin, de nouvelles variables sont ici intégrées au modèle de prime pure et semblent apporter une bonne information. Il s'agit de la variable de classement de prix SRA qui regroupe les 8 dernières classes de prix SRA pour lesquelles il y avait peu de données (gta_classeprix_ret") et également les variables : genre du véhicule ("veh_cdgenrn") qui permet de savoir s'il s'agit d'un véhicule léger ou utilitaire, type de carrosserie ("veh_carrossret") qui regroupe les véhicules selon leur type de carrosserie et segment SRA que nous avons décrit précédemment. Dans l'ensemble, nous constatons que de nombreuses variables ont un très grand *spread* 100/0%. Par la suite, une analyse détaillée de la cause sera nécessaire.

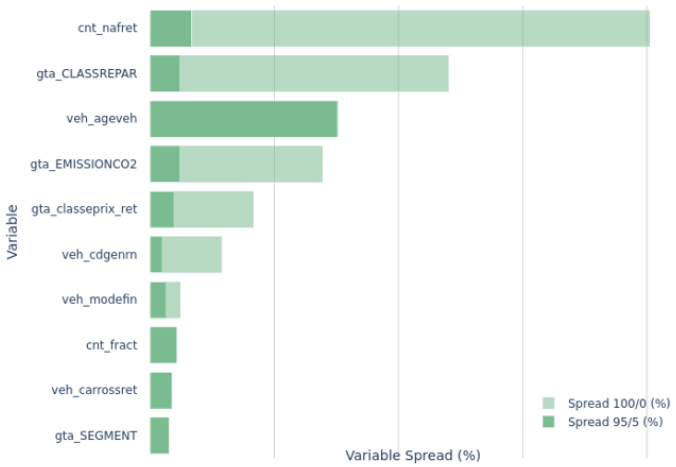


FIGURE 4.20 – Variables utilisées dans la prédiction de la prime pure

Validation du modèle

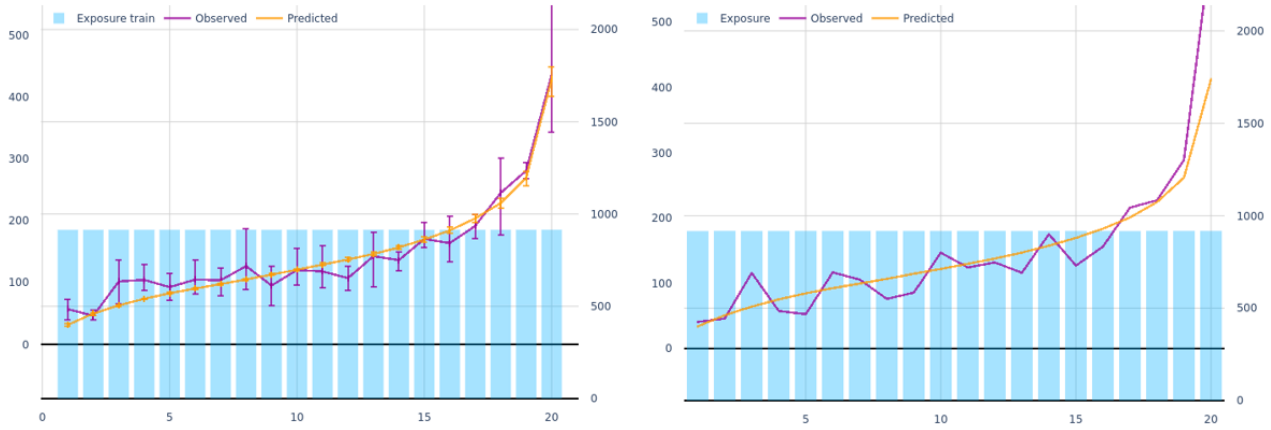
Les variables ayant été décrites succinctement, il est utile de s'intéresser aux différents outils nous permettant de valider ou d'invalider le modèle de prime pure. Comme pour les modèles de fréquence et de coût moyen, nous allons étudier les différents indicateurs statistiques afin de juger de la robustesse du modèle puis nous allons nous intéresser aux *lift curves* sur les bases de modélisation et de validation puis nous analyserons les résidus.

Indicateurs statistiques Afin de comprendre l'importance de la volumétrie des données, le nombre d'observations ainsi que l'exposition totale, des bases de modélisation et de validation ont été ajoutées au tableau 4.4 des indicateurs statistiques. L'exposition totale correspond à la somme des années-polices de chaque véhicule. Dans l'ensemble, on remarque que tous les indicateurs statistiques sont relativement proches entre les bases de modélisation et de validation. En particulier, les primes pures moyennes prédites s'approchent bien des primes pures moyennes observées. Par ailleurs, l'indice de Gini, le RMSE, la déviance moyenne ainsi que le MAE sont stables. On peut donc considérer que les données sont homogènes entre les bases de modélisation et de validation. On conclut donc que le modèle est stable et que nous pouvons l'appliquer sur la base de validation.

TABLE 4.4 – Indicateurs statistiques sur les bases de modélisation et de validation du modèle de prime pure

Indicateurs	Base de modélisation	Base de validation
Nombre d'observations	87 251	21 943
Exposition totale	73 154	18 356
Prime pure moyenne observée	147,30	151,12
Prime pure moyenne prédite	144,30	143,52
Gini	33,99%	34,70%
RMSE	1 410	1 350
Déviance moyenne	71,84	73,14
MAE	270,6	273,9

Lift curve L'analyse des *Lift curves* à partir de la figure 4.21 nuance cependant les bons indicateurs statistiques. En effet, même si on voit que les prédictions s'ajustent correctement aux observations de la base de modélisation, on remarque quelques écarts de prédiction par rapport aux données observées de la base de validation sur certains quantiles. Ceci peut s'expliquer soit par le manque de données dans la base de validation qui ne permet pas d'obtenir la stablité de la charge, soit par l'oubli de retraitement de certaines charges ouvertes forfaitairement. On retrouve par exemple, certains pics de sinistres décrit dans l'analyse préliminaire. Néanmoins, les courbes de prédictions suivent malgré tout la tendance des observations.



(a) Base de modélisation

(b) Base de validation

FIGURE 4.21 – *Lift curves* du modèle de prime pure

Analyse des résidus L'analyse des résidus de déviance agrégés représentés sur la figure 4.22 ci-après permet valider de l'hypothèse d'indépendance des résidus. Par ailleurs, ils permettent également de vérifier qu'il n'y a pas de biais dans la modélisation et que les hypothèses des GLM sont également respectées. Comme attendu, les résidus sont proches de 0 et n'ont pas de forme spécifique.

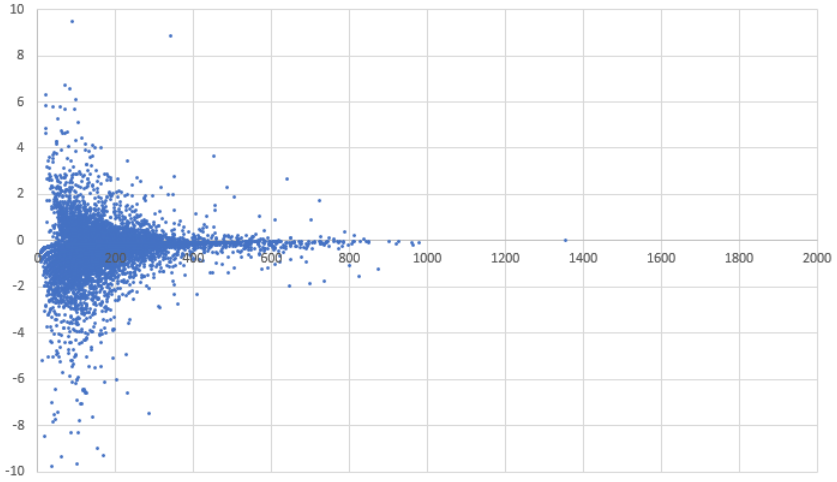


FIGURE 4.22 – Résidus de déviance agrégés pour le modèle de prime pure

On peut donc raisonnablement conclure que le modèle de prime pure approché par la loi de Tweedie s'adapte bien aux données. Une base de données de sinistralité plus conséquente aurait certainement permis une meilleure stabilité entre les bases de modélisation et de validation mais cela reste tout à fait acceptable au regard des autres modèles testés et des différents indicateurs étudiés.

Description des variables

Les différents critères de sélection du modèle ayant été étudiés, nous pouvons maintenant nous intéresser à l'étude des variables utilisées dans le modèle de prime pure. En effet, il est important de comprendre l'apport de chacune des variables dans le modèle mais également de s'assurer qu'il n'y a pas de sur-dispersion des coefficients dans le GLM. Afin de comprendre l'influence de chaque modalité, nous présenterons les valeurs relatives des coefficients du modèle (*Coefficient value*), la prime pure observée (*Observed Average*) ainsi que la prime pure moyenne prédite (*Fitted Average*) pour chacune des variables présentes dans le modèle.

Code NAF La variable "cnt_nafret" permet de segmenter le risque selon le code NAF de l'entreprise souscrivant un contrat FLOTTE. Il permet de renseigner sur l'activité de l'entreprise comme par exemple les entreprises agricoles (AGR), les entreprises de transport médicalisé telles que les ambulances (AMB) ou encore les entreprise de VTC ou de taxis (TAX). C'est la variable pour laquelle on constate le plus grand *spread* 100/0%. En effet, on peut remarquer sur la figure 4.23 que, sans surprise, les auto-écoles (ENS) ont le coefficient le plus bas. À l'inverse, les entreprises liées au transport urbain (TRU) ont le coefficient le plus élevé. Il sera peut-être nécessaire de corriger ce *spread* en lissant les coefficients car il peut y avoir un sinistre important sur ce segment qui a pénalisé la modélisation.

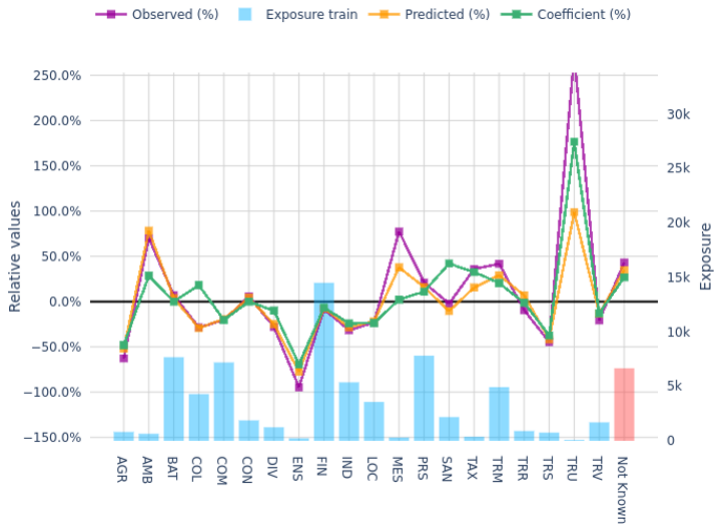


FIGURE 4.23 – Coefficient de la variable "cnt_nafret"

Classe de réparation SRA Comme il a été mentionné précédemment, la classe de réparation SRA est un indicateur estimant le coût de la réparation d'un véhicule. Elle est définie par une lettre représentant des montants de réparation et varie de "A" à "Z". La classe "HC" représente les véhicules hors classe et les classes "ZA" à "ZE" ont été ajoutées récemment afin de prendre en compte les indices de réparation de plus en plus élevés avec les nouvelles technologies embarquées sur les véhicules récents. On constate sur la figure 4.24 que le coefficient le plus élevé concerne la catégorie H. Cela signifie que les véhicules de classe de réparation "H" auront une prime 3 fois plus élevée que les véhicules de classe de réparation "F" ou "G". Afin de ne pas créer de disparité entre les classes voisines, il conviendra de lisser ce coefficient. En effet, une explication possible à un coefficient aussi élevé peut provenir de la sur-sinistralité engendrée par une flotte de véhicules légers présents en nombre dans les entreprises de messageries de type La Poste.

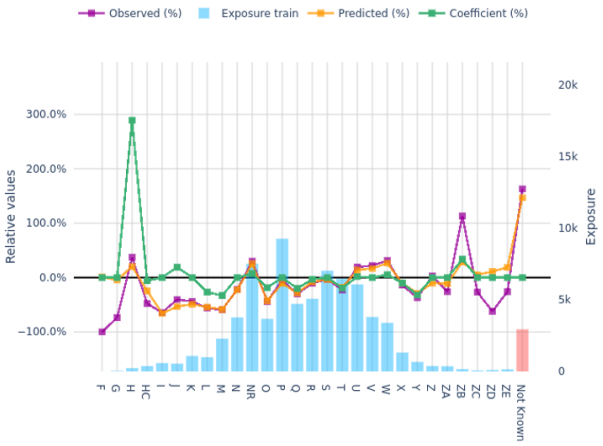


FIGURE 4.24 – Coefficient de la variable "gta_CLASSREPAR"

Âge du véhicule La variable âge du véhicule est la variable la plus discriminante du modèle. C'est ce qu'on peut voir sur la figure 4.25 car les coefficients du modèle prédisent relativement bien l'observé. La charge des sinistres a tendance à diminuer en fonction de l'âge du véhicule, soit parce que les sinistres ne sont plus déclarés à cause d'un véhicule trop ancien pour être réparé, soit parce qu'ils sont moins utilisés.

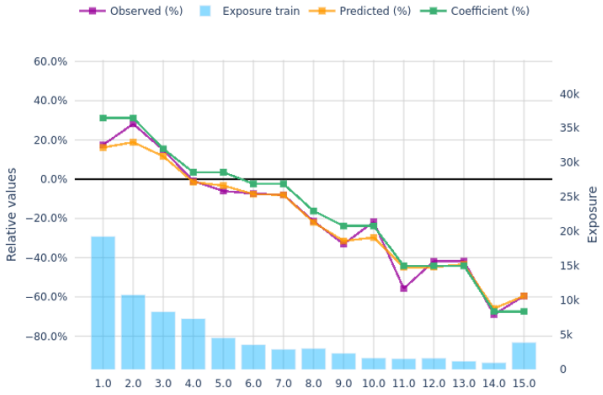


FIGURE 4.25 – Coefficients de la variable "veh_ageveh"

Émission de CO_2 La variable permettant d'apprécier le niveau d'émission de CO_2 est une variable très segmentante. C'est ce que nous pouvons remarquer lorsqu'on observe la répartition des charges observées sur la figure 4.26. En revanche, on distingue 3 groupes qui semblent bénéficier des mêmes coefficients. En effet, les véhicules émettant le moins de CO_2 ainsi que ceux émettant plus de 130g/km de CO_2 semblent avoir un coefficient plus élevé que les véhicules émettant entre 90g/km et 130g/km de CO_2. Ceci peut s'expliquer par le fait que les véhicules électriques, n'émettant mécaniquement pas de CO_2, ainsi que les véhicules émettant le plus de CO_2 tels que les SUV ou les 4x4 peuvent avoir des charges de sinistres plus élevées que les voitures "classiques" notamment à cause du coût de réparation de ces véhicules particuliers. Il serait intéressant de regrouper ces modalités afin de ne pas intégrer trop de volatilité dans les coefficients.

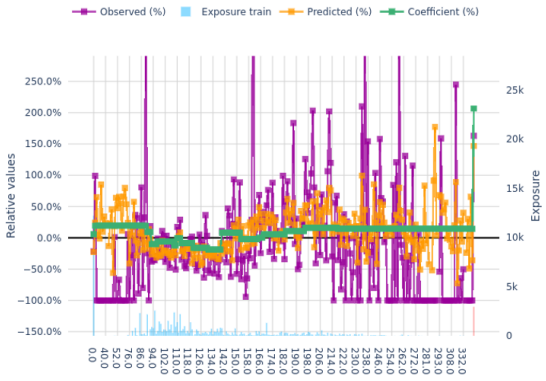


FIGURE 4.26 – Coefficients de la variable "gta_EMISSIONCO2"

Classe de prix SRA La classe de prix SRA est définie par la valeur à neuf du véhicule. Actualisée chaque année en fonction de l'indice INSEE du prix des véhicules neufs, elle est définie par une lettre variant de la même manière que la classe de réparation SRA. Cependant, étant donné le peu de volume de véhicules présents dans les dernières classes de prix, nous avons fait le choix de regrouper les véhicules des classes "W" à "ZE". Ce regroupement n'est peut être pas suffisant car on constate sur la figure 4.27 que le coefficient de la classe T est le deuxième coefficient le plus bas alors que sa classe de prix associée est élevée. Il conviendra là aussi de regrouper certaines classes présentant le moins d'effectifs afin de réduire le *spread* des coefficients.

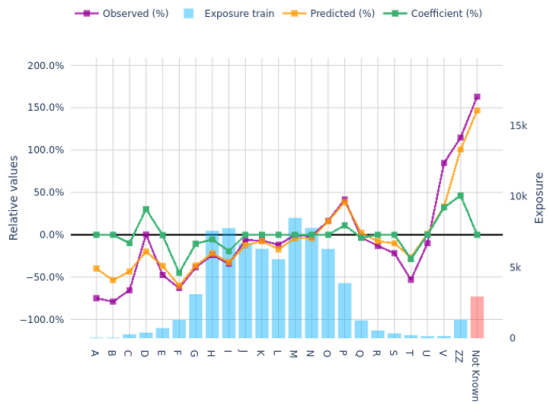


FIGURE 4.27 – Coefficients de la variable "gta_classeprix_ret"

Code genre Le code genre du véhicule est issu des bases véhicules AXA. Il permet de dissocier les véhicules légers (VL), des véhicules spécialisés (VS), des véhicules de société (VT) ou encore des véhicules utilitaires (VU). On constate sur la figure 4.28 que les VS et VT sont sous-représentés et par conséquent, sans charge sinistre suffisante pour pouvoir modéliser ces modalités. Les coefficients associés sont donc naturellement bas. Afin de ne pas introduire de biais dans le modèle, une solution consisterait à lisser ces coefficient au niveau du coefficient des VL par prudence et afin de ne pas appliquer de rabais tarifaire sur ces segments méconnus.

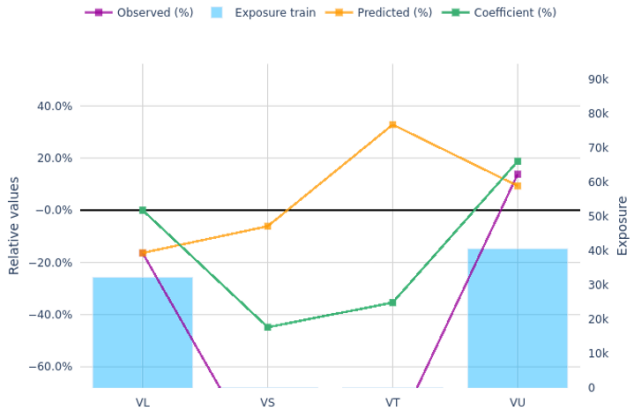


FIGURE 4.28 – Coefficients de la variable "veh_cdgenrn"

Mode de financement Le mode de financement du véhicule est également une variable issue des bases véhicules internes d'AXA. Cette variable permet de savoir si le véhicule a été acheté à crédit, en *leasing* ou location avec option d'achat, en crédit bail ou en financement particulier. La modalité autre regroupe les véhicules dont le mode financement n'est pas connu. On remarque sur la figure 4.29, que les véhicules achetés à crédit ainsi que ceux achetés en crédit-bail ont les coefficients les plus élevés. On peut alors penser que les véhicules n'étant pas la propriété de l'entreprise sont les véhicules les plus sinistrés.

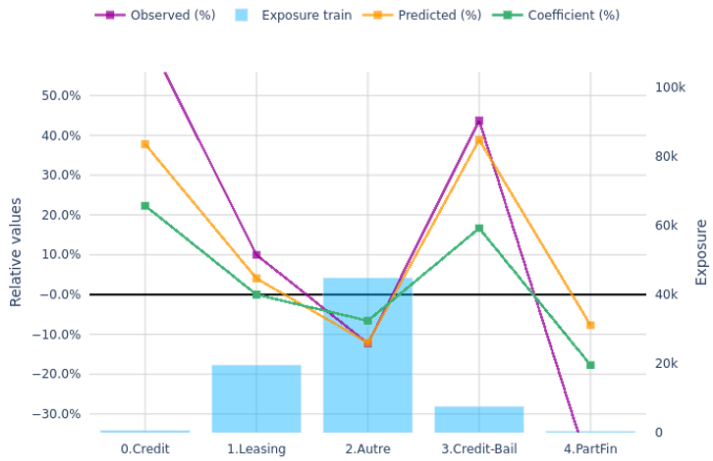


FIGURE 4.29 – Coefficients de la variable "veh_modefin"

Fractionnement du contrat Le fractionnement du contrat a été intégré au modèle car il apportait une information complémentaire par rapport aux autres variables. En effet, on constate sur la figure 4.30 que les contrats annualisés ont environ un coefficient de -20% par rapport à la moyenne ce qui laisse à penser que les clients qui paient leurs cotisations en une fois peuvent être considérés comme les "bons" clients pour lesquels on peut penser qu'ils ont les moyens de mieux entretenir leurs véhicules. Par ailleurs, il est tout à fait logique de faire bénéficier à l'assuré qui paie sa cotisation en une fois, d'une réduction tarifaire au titre d'actes de gestion diminués.

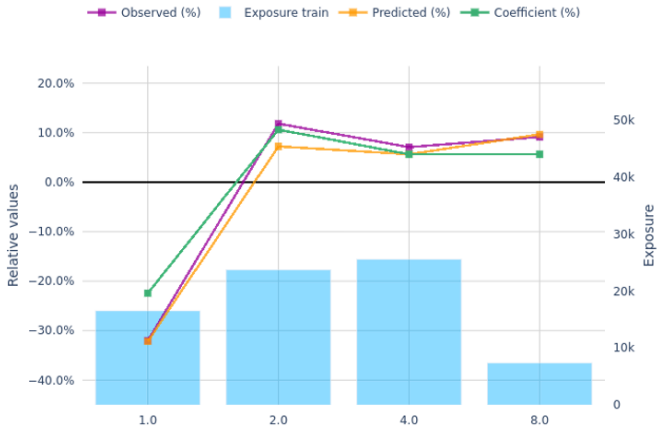


FIGURE 4.30 – Coefficients de la variable "cnt_fract"

Carrosserie La variable carrosserie est issue des bases véhicules d'AXA. Cette variable renseigne sur le type de carrosserie du véhicule. Il peut s'agir d'un 4x4 (44), d'une berline (BE), d'un cabriolet (CB) ou encore d'un ludospace (LS) ou d'un monospace (MS). Les coefficients de cette variable étant relativement proches, ils ne permettent pas de distinguer les carrosseries qui auront un coefficient de majoration. En revanche, on peut voir que les charges de sinistres liées aux ambulances (AM) et aux cabriolets (CB) sont les plus élevées. Il serait intéressant de lisser les coefficients afin de s'approcher d'avantage de l'observé.

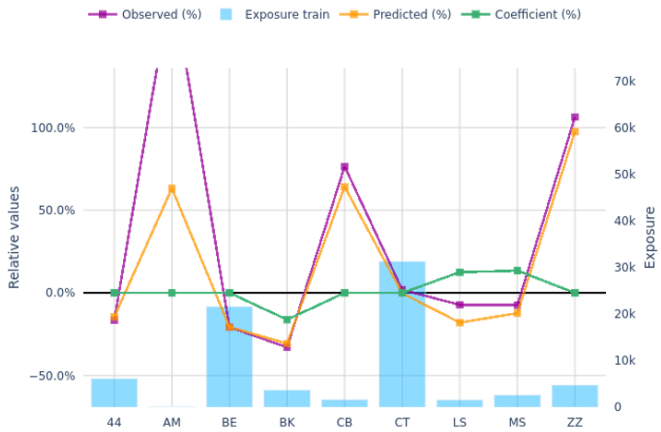


FIGURE 4.31 – Coefficients de la variable veh_carrossret

Segment SRA La segmentation SRA reprend en partie la classification européenne des véhicules particuliers auxquels s'ajoute, les véhicules utilitaires. Concernant les véhicules particuliers, le segment "A" correspond aux véhicules urbains ou "petites citadines", le segment "B" correspond aux citadines polyvalentes, le segment "H" correspond aux grandes et très grandes berlines, le segment "M1" inclut les compactes, il s'agit du segment le plus représenté en France et le segment "M2" représente les berlines de taille moyenne. Du côté des véhicules utilitaires, les segments "K1" et "K2" représentent les fourgonnettes. On retrouve bien sur la figure 4.32, les segments des véhicules utilitaires qui sont pénalisés par un coefficient supérieur à celui des autres segments.

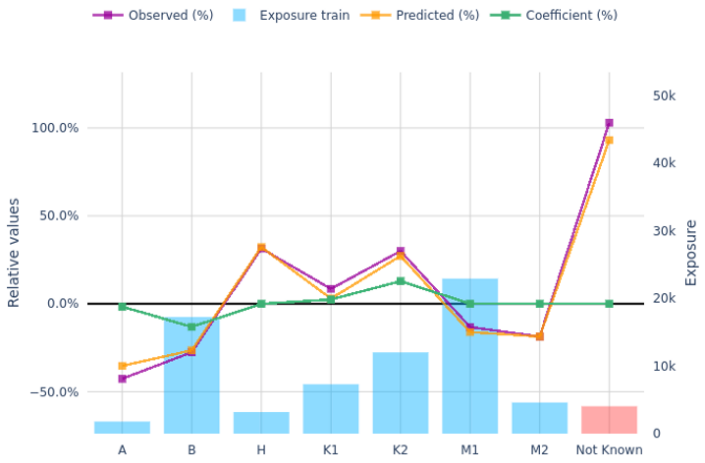


FIGURE 4.32 – Coefficients de la variable "gta_SEGMENT"

Zonier

Le modèle de prime pure ayant été validé, nous avons tenté de mettre en place un zonier à 13 zones afin de respecter la structure actuellement en place concernant le modèle PARC. En effet, un zonier géographique est usuellement construit après avoir prédit le risque en fonction des variables discriminantes. Nous cherchons alors à savoir s'il existe une information complémentaire qui permet de modéliser la part de risque non modélisée par les différentes variables explicatives choisies. La variable zonier est construite indépendamment des autres variables de sorte qu'elle ne diminuera pas les résidus de la modélisation. En effet, elle ne peut qu'améliorer le modèle dans l'éventualité où la dimension géographique aurait une influence sur la fréquence des sinistres. En outre, on peut s'attendre à ce que la fréquence des sinistres soit plus importante dans les grandes agglomérations que dans les provinces où la circulation est moins dense.

La théorie de la mise en place d'un zonier repose sur les mesures d'autocorrélations spatiales selon l'emplacement des sinistres ainsi que leurs charges associées par le biais de l'utilisation de la distance de Moran ou de Geary par exemple. En effet, on constate très souvent, comme c'est le cas dans les agglomérations, que les variables ayant une dimension géographique sont soumises à des dépendances ou à des interactions spatiales. Afin de mettre en évidence ces dépendances, les mesures d'autocorrélations spatiales prennent généralement en compte deux critères que sont la proximité spatiale et la proximité de la valeur prise par la variable. La distance de Moran, qui a la particularité de mesurer le niveau d'autocorrélation spatiale mais également de tester la significativité de cette

mesure, a donc été utilisée dans la suite de ce travail. Pour plus de détails, le lecteur est invité à se référer à Oliveau (2010).

La première étape consiste donc à affecter à chaque observation, une donnée spatiale. Si pour le périmètre PARC, le code INSEE de la ville du risque est connu car il s'agit en général de petites flottes dont les véhicules sont utilisés dans une zone souvent restreinte, ce n'est pas le cas pour les FLOTTES. En effet, les FLOTTES sont composées d'un grand nombre de véhicules qui sont généralement amenés à parcourir des distances plus ou moins importantes.

Dans le but d'étudier si une donnée géographique pouvait amener une information complémentaire au modèle de prime pure, nous avons utilisé la seule donnée à notre disposition, le code postal de souscription. Cette donnée, bien que ne reflétant pas intrinsèquement le lieu du risque, permet néanmoins d'avoir une estimation de la zone où pourrait se trouver le risque et ainsi amener la finesse nécessaire à l'utilisation de la distance de Moran dans l'étude de la dépendance entre "voisins". En effet, nous avons fait l'hypothèse que les véhicules d'une entreprise qui souscrirait un contrat FLOTTE "naviguerait" autour de la zone où a été conclue l'affaire. Nous avons donc construit une base de codes postaux et répertorié leur latitude et leur longitude respectives. Par ailleurs, associer un code postal à une ligne de risque ne suffit pas dans l'élaboration d'un zonier. Il conviendra en effet d'appliquer plusieurs degrés de lissage afin de ne pas affecter une zone à un unique code postal ou encore deux zones différentes entre deux villes géographiquement proches. Ce niveau de lissage est lisible sur la *grid search* 4.33 ci-dessous, plus la distance de Moran est élevée, plus grand est le lissage. Le modèle initial est représenté en orange à droite du graphique, dans la seule perspective de pouvoir comparer les indices de Gini avec les modèles de zonier. Nous avons donc réalisé une comparaison de quelques modèles ayant une distance de Moran différente.

FIGURE 4.33 – *Grid search* des modèles de prime pure avec zonier

En premier lieu, nous avons analysé les indicateurs statistiques du modèle ayant l'indice de Gini le plus élevé. Les modèles réprésentés à gauche sur la *grid search* 4.33 correspondent aux modèles ayant les indices de Gini les plus élevés car leur distance de Moran est la plus faible. Cela signifie que le niveau de lissage est faible et que l'on accorde plus d'importance à la variable de zonier ce qui peut impliquer un sur-apprentissage des données géographiques. C'est que nous pouvons voir sur le zonier représenté sur la figure 4.34 où l'on constate la présence de zones très différentes sur un périmètre géographique très restreint. Par ailleurs, l'indice de Gini de la base de modélisation (48,32%) et celui de la base de validation (38,06%) sont très éloignés du à ce sur-apprentissage, ce qui nous permet de conclure que le modèle n'est pas stable.

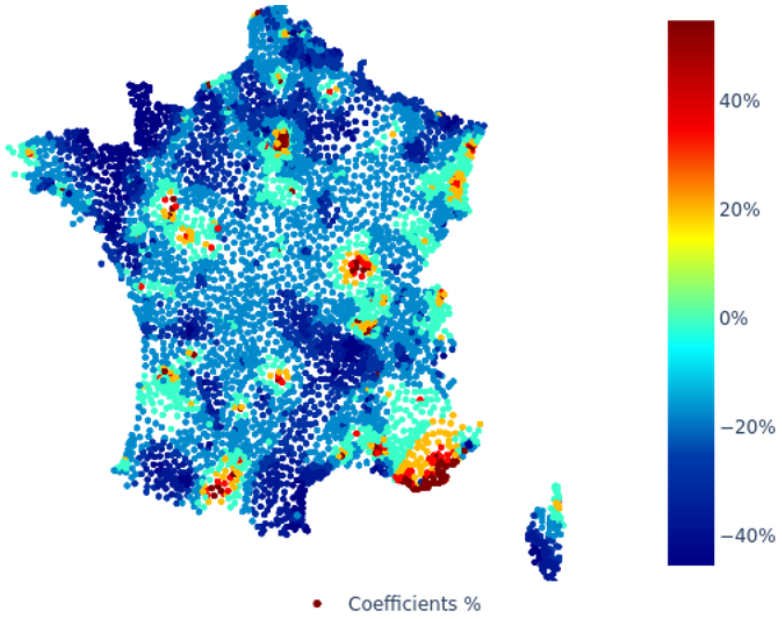


FIGURE 4.34 – Zonier avec un faible lissage

Étant donné la non stabilité des modèles dont la distance de Moran est la plus faible, nous avons étudié les modèles ayant des distances de Moran moyennes qui permettent de ne pas trop segmenter le risque en fonction de la géographie tout en essayant de conserver une certaine disparité dans les zones afin de ne pas trop lisser le modèle, ce qui reviendrait à ne différencier aucune zone. Les modèles répondant à ces critères ont une distance de Moran comprise entre 100 000 et 200 000 et leur indice de Gini est compris entre 28% et 30%. Parmis ces modèles, nous avons conservé le modèle ayant le plus de stabilité entre les bases tests et les bases d'entrainement. Cela permettra d'avoir une meilleure robustesse entre la base de validation et la base de modélisation comme on peut le voir dans la table 9 en annexe A.4. En revanche, cela à un coût comme on peut le constater sur la figure 4.35 ci-après où le lissage est moins segmentant. Ainsi, les "voisins" les plus proches des villes où sont présents les mauvais risques sont pénalisés.

Enfin, on constate également que la variable de code postal n'est pas suffisante dans le cadre de l'application d'un zonier sur les flottes ouvertes. Même si elle donne une bonne indication géographique du risque, la carte 4.35 ci-après semble indiquer une sur-sinistralité dans la région centre. En réalité, il s'agit d'une importante entreprise dont le siège social est domicilié dans cette région mais pour laquelle les véhicules parcourent la France entière. Néanmoins on remarque que, mise à part l'agglomération lilloise, le zonier modélisé fait apparaître les grandes métropoles dans lesquelles on observe généralement une sur-sinistralité par rapport aux zones rurales.

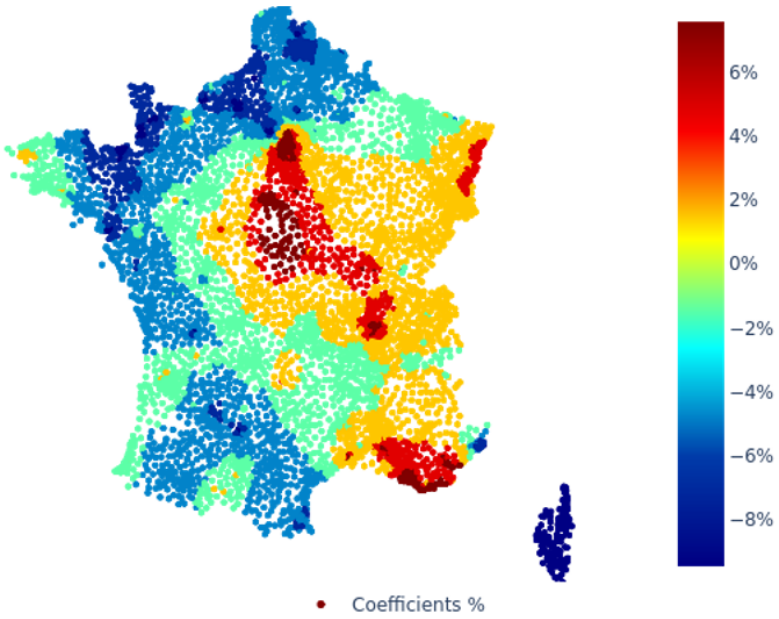


FIGURE 4.35 – Zonier avec un lissage modéré

4.3 Récapitulatif des modèles - Conclusion

Dans un environnement où la simplicité de la souscription est louée par le client, il devient de plus en plus urgent de simplifier les modèles de tarification tout en permettant la récupération d'importantes informations permettant d'établir un tarif juste, précis et compétitif. Les informations demandées au client sont donc d'une importance capitale desquelles il faut pouvoir en extraire le plus de données possibles.

Grâce à l'immatriculation de chaque véhicule, il est possible de récupérer des informations de la carte grise ainsi que des variables tarifaires par le biais de sources de données externes et libres d'accès (*open data*). Dès lors, il est possible de faciliter la souscription en ne demandant qu'une quantité minimale d'information à l'assuré.

Par ailleurs, en utilisant les variables prises en compte dans la tarification des parcs, nous avons montré qu'il était possible d'appliquer ce modèle aux flottes ouvertes, qui a notamment fait ses preuves puisque le segment des parcs bénéficie d'une meilleure rentabilité que celui des FLOTTES. Cette démonstration permettrait un gain économique certain dans la nouvelle tarification des flottes ouvertes en minimisant les coûts de développement informatiques. Les FLOTTES seraient alors traîtées de la même manière que les PARCS.

Pour autant, nous avons tenté de décrire la charge de sinistres par un modèle propre aux flottes ouvertes en utilisant la décomposition fréquence $\times$ coût moyen afin de produire un modèle plus fiable. Cependant, nous nous sommes heurtés à un manque de données entraînant une absence de robustesse dans la modélisation du coût moyen qui ne nous a pas permis de conclure sur l'efficacité de ce modèle. En revanche, l'approche prime pure en utilisant la distribution de Tweedie a permis de surmonter ce manque de données en intégrant dans les modélisations, les charges nulles et l'absence

de sinistralité sur les véhicules concernés ce qui a mécaniquement entrainé une certaine stabilité.

Enfin, le tableau récapitulatif 4.5 des indicateurs statistiques des principaux modèles étudiés nous donne un bon aperçu de la puissance de chacun d'entre eux. Une connaissance enrichie des sinistres permettra, à terme, d'apporter une réponse concrête au modèle de fréquence × coût moyen qui, comme nous avons pu le voir, a tendance à sur-estimer la charge sinistre mais apportera également une meilleure stabilité des observations entre les bases de modélisation et de validation.

TABLE 4.5 – Indicateurs statistiques sur les bases de validation des principaux modèles testés

Indicateurs	Modèle Prime Pure "PARC"	Modèle Fréquence × Coût moyen "FLOTTE"	Modèle Prime Pure "FLOTTE"	Modèle Zonier "FLOTTE"
Prime pure moyenne observée	151,12	148,03	151,12	151,12
Prime pure moyenne prédite	145,08	166,39	143,52	143,52
Gini	34,49%	34,71%	34,70%	36,09%
RMSE	1 339	1 397	1 350	1 349
Déviance moyenne	72,58	72,18	73,14	72,67
MAE	273,7	289,6	273,9	272,0

Conclusion

Ce mémoire a pour objectif principal de redresser la rentabilité de la branche automobile en permettant de nouveau la croissance au travers de l'amélioration et de la sophistication du tarif des flottes ouvertes par l'acquisition de nouvelles données. La mise en place de cette nouvelle tarification intervient dans un environnement concurrentiel au sein duquel les assureurs cherchent, par le biais d'innovations, à segmenter leur tarif afin de le rendre le plus précis et le plus compétitif possible. Par ailleurs, ce travail s'incrit dans l'ambition stratégique d'AXA France de redresser ce segment déficitaire par la connaissance approfondie de ses flottes ouvertes pour lesquelles l'entreprise n'avait que peu d'information. L'étude s'est donc concentrée sur la tarification de la garantie principale et obligatoire de responsabilité civile des véhicules majoritaires de catégorie 1 dits véhicules légers.

La création du Fichier des Véhicules Assurés (FVA) ainsi que l'obligation légale pour les assureurs de déclarer les véhicules qu'ils assurent, a nécessité la mise en place d'outils spécifiques permettant la récupération, la saisie ainsi que la déclaration des immatriculations de l'ensemble des véhicules assurés par AXA. Ce projet, qui a débuté bien avant l'application de la directive en janvier 2019, a été l'occasion pour la branche automobile entreprise d'accroître sa connaissance du segment des flottes ouvertes par le biais de l'acquisition du numéro d'immatriculation de chaque véhicule. En effet, véritable numéro d'identité, l'immatriculation permet d'avoir accès à de nombreuses informations sur le véhicule qui peuvent s'avérer déterminantes dans un tarif. Ces données, provenant aussi bien de sources internes que de sources externes, ont permis de tester différents modèles permettant de modéliser la charge des sinistres. Un travail de recherche a donc été nécessaire afin de comprendre et de choisir les modèles les plus adaptés à la problématique.

Cependant, avant de pouvoir s'intéresser à un modèle propre au segment des flottes ouvertes, il a été nécessaire d'étudier le modèle de tarification utilisé sur le segment des parcs qui sont de plus petite taille et qui ont l'avantage d'avoir plus de données. En effet, dans un soucis de simplification et de réduction des coûts, il a paru évident de savoir s'il était possible de proposer un tarif "parc" aux flottes ouvertes afin de faciliter leur intégration dans les outils de souscription actuellement en fonction. Les modèles GLM ont donc été favorisés car ils ont l'avantage d'être implémentables rapidement, à moindres coûts et sont ceux utilisés pour la tarification du parc. L'approche prime pure par une loi de Tweedie a été privilégiée afin de conserver un maximum de données de sinistralité dans le but d'apporter une certaine stabilité aux résultats dont une attention particulière a été portée tout au long de ce travail. Par ailleurs, des retraitements sur les données de sinistres ont dû être effectués afin de ne conserver que la charge attritionnelle effective et supprimer le biais que pouvaient apporter les sinistres ouverts aux forfaits (AXA, IRSA et IRCA) ainsi que les charges négatives (recours et conservations) dans les modélisations. L'estimation de la prime pure par les variables utilisées sur le parc a alors montré de bonnes performances qui sont nuancées par un écart relativement faible de l'indice de Gini entre les bases de modélisation et de validation. Par la suite,

une approche fréquence × coût moyen a été utilisée afin de construire un modèle propre aux flottes ouvertes, utilisant les variables les plus discriminantes du segment. Des performances prometteuses ont été constatées sur le modèle de fréquence mais neutralisées par la mauvaise adéquation du modèle de coût moyen à la charge des sinistres. L'approche prime pure pour la création d'un modèle propre aux flottes ouvertes a donc de nouveau été privilégiée afin de tenir compte de la présence importante de données sans sinistralité. La modélisation a montré de très bonnes performances en terme de stabilité et de robustesse, à la fois sur les bases de modélisation mais également sur les bases de validation. L'efficacité du modèle ayant été prouvée, un zonier a été mis en place comme c'est le cas pour les parcs. L'intégration de données géographiques a permis l'optimisation du modèle en améliorant la performance mesurée par l'indice de Gini, tempérée cependant par la mauvaise information apportée parfois, par le code postal de souscription. En effet, le code postal de souscription peut correspondre dans certain cas au siège social de l'entreprise et non au lieu de garage des véhicules. Ces différents travaux ont permis d'une part, de constater que le modèle parc pouvait être répliqué sur les flottes mais également de connaître les variables les plus discriminantes sur ce segment en vue d'une éventuelle implémentatation d'un tarif spécifique aux flottes ouvertes.

Néanmoins, la quantité des données liées, notamment à la sinistralité, fut la principale limite de cette étude. En effet, l'une des problématiques majeures concernant les produits IARD Entreprises est la différence de volume de données par rapport aux produits destinés aux particuliers. L'absence de détails sur le client ou d'informations concernant le produit assuré ne permettent pas d'obtenir un tarif aussi précis que ceux proposés dans le cadre d'une assurance automobile du particulier par exemple. Par ailleurs, les retraitements de sinistres forfaitaires ou de charges négatives effectués ont considérablement réduit la base de données sinistres. De plus, la récente mise en place du FVA n'a pas permis de prendre en compte plus d'une année de vision car les véhicules ne sont recensés que depuis le 1^{er} janvier 2019. Une meilleure consistance des données permettra sans doute l'amélioration de la fiabilité des modèles testés.

Indépendamment de la prise en compte future de données supplémentaires qui permettront par la suite, d'appliquer ce travail en segmentant la fréquence en fonction de la responsabilité du conducteur du véhicule mais également d'apporter de la stabilité aux résultats, il sera nécessaire d'effectuer ce travail de modélisation sur les autres catégories de véhicules. En effet, si l'obligation de déclaration au FVA ne concerne, pour l'heure, que les véhicules de moins de 3,5 tonnes, AXA France a fait le choix de recenser l'ensemble de ses véhicules qu'elle assure afin d'anticiper l'obligation de déclaration pour les véhicules de plus de 3,5 tonnes. Les évolutions futures consisteront donc à la prise en compte de ces données afin de proposer une tarification de la garantie RC aux véhicules des autres catégories. Par ailleurs, un futur mémoire pourrait faire, par exemple, l'objet de la tarification des garanties optionnelles telles que le dommage, l'incendie, le vol et le bris de glace, pour lesquels il sera nécessaire d'implémenter un modèle particulier car la connaissance de la mise en place de ces garanties par véhicule n'est pas acquise pour le périmètre des flottes ouvertes.

Enfin, au-delà du fait que la mise en place du FVA permettra à terme, de détecter automatiquement les véhicules non assurés, à la fois par les forces de l'ordre mais également par les lecteurs automatiques de plaques d'immatriculation, et ainsi participer à la lutte menée par le Fonds de Garantie des Assurances Obligatoires de dommages (FGAO) contre la conduite sans assurance, elle a surtout permis à la branche automobile entreprise de bénéficier d'un atout considérable dans la refonte de sa gamme tarifaire sur le segment des flottes automobiles. Ces évolutions permettront assurément de reconquérir ce marché tendu.

Table des figures

Liste des tableaux

Bibliographie

Boyer-Chammard, R. (2008), 'Processus de surveillance et de majorations des contrats flottes d'entreprise d'axa france', *Mémoire d'actuariat* .

Denuit, M. & Charpentier, A. (2005), *Mathématiques de l'assurance non-vie, Tome 2 : tarification et provisionnement*, Econometrica.

Dunn, P. K. & Smyth, G. K. (1996), 'Randomized quantile residuals', *Journal of Computational and Graphical Statistics* .

Gonnet, G. (2010), 'Étude de la tarification et de la segmentation en assurance automobile', *Mémoire d'actuariat* .

McCullagh, P. & Nelder, J. (1989), *Generalized Linear Models*, 2nd edn, Chapman & Hall/CRC.

Nelder, J. A. & Wedderburn, R. W. M. (1972), 'Generalized linear models', *Journal of the Royal Statistical Society* .

Nguyen, T. T.-V. (2008), 'Flottes automobiles : Un nouveau modèle de tarification. impact de la conservation sur la distribution du ratio sinistres à primes', *Mémoire d'actuariat* .

Oliveau, S. (2010), *Autocorrélation spatiale : leçons du changement d'échelle*, Vol. 39, L'espace géographique 2010/1.

Rebadj, A. (2016), 'Estimation de la "valeur contrat" des assurés automobiles dans le cadre de la mise en place d'un indicateur "valeur client"', *Mémoire d'actuariat* .

Annexes

A.1 Référentiels RC de fréquence et de coût moyen

TABLE 6 – Tableau des fréquences de la caculette flottes ouvertes

Segmentation	RCMAT 100%	RCMAT 0%	RCCORP 100%	RCCORP 0%	RC Total
VL	6,4%	4,0%	0,6%	0,2%	11,1%
TPM M3T5	21,9%	7,1%	1,9%	0,5%	31,4%
TPM P3T5	22,0%	7,6%	1,1%	0,4%	31,1%
TPV M3T5	8,1%	8,0%	1,1%	0,6%	17,8%
TPV P3T5	13,8%	6,6%	1,0%	0,4%	21,7%
TPC	10,7%	3,1%	0,5%	0,2%	14,6%
Engins de chantier	10,1%	2,7%	0,4%	0,1%	13,4%
Engins agricoles	3,3%	1,1%	0,1%	0,1%	4,5%

TABLE 7 – Tableau des coûts moyens en euros de la caculette flottes ouvertes

Segmentation	RCMAT 100%	RCMAT 0%	RCCORP 100%	RCCORP 0%	RC Total
VL	1 182	355	8 863	1 841	1 278
TPM M3T5	1 259	398	9 770	485	1 572
TPM P3T5	1 435	300	13 903	3 123	1 605
TPV M3T5	1 112	320	8 927	3 618	1 318
TPV P3T5	1 020	283	14 609	9 520	1 570
TPC	1 508	445	13 409	2 580	1 719
Engins de chantier	1 322	681	27 584	2 946	2 097
Engins agricoles	1 938	616	20 073	5 137	2 293

A.2 Famille exponentielle - Transformation des distributions

Dans cette annexe sont présentés des exemples d'utilisation de la formule 3.1 concernant la transformation des principales distributions utilisées sous forme exponentielle.

Loi Normale $\mathcal{N}(\mu, \sigma^2)$

La densité d'une famille de loi Normale s'écrit :

$$\begin{aligned} f_Y(y_i|\theta_i, \mu_i) &= \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{(y_i-\mu_i)^2}{2\sigma^2}\right\} \\ &= exp\left\{-\frac{\mu_i^2}{2\sigma^2}\right\} exp\left\{-\frac{y_i^2}{2\sigma^2} - \frac{1}{2}ln(2\pi\sigma^2)\right\} exp\left\{y_i\frac{\mu_i}{\sigma^2}\right\} \end{aligned}$$

D'où en posant :

$$\begin{aligned} Q(\theta_i) &= \frac{\theta_i}{\phi} = \frac{\mu_i}{\sigma^2} \\ a(\theta_i) &= exp\left\{-\frac{\mu_i^2}{2\sigma^2}\right\} \\ b(y_i) &= exp\left\{-\frac{y_i^2}{2\sigma^2} - \frac{ln(2\pi\sigma^2)}{2}\right\} \end{aligned}$$

La famille de loi Normale se met sous la forme canonique 3.2 et peut donc être considérée comme une famille exponentielle.

Loi de Bernoulli $\mathcal{B}(1, p)$

Soient $(Y_1, \ldots, Y_n)$ n variables indépendantes de loi de Bernoulli tel que $\mathbb{E}(Y_i) = p_i$. On a alors pour tout $i = 1, \ldots, n$:

$$\begin{aligned} \mathbb{P}(Y_i = 1) &= p_i \\ \mathbb{P}(Y_i = 0) &= 1 - p_i \end{aligned}$$

Que l'on peut réécrire :

$$\begin{aligned} f_Y(y_i|\theta_i, p_i) &= p_i^{y_i}(1-p_i)^{1-y_i} \\ &= (1-p_i)exp\left\{y_i ln\left(\frac{p_i}{1-p_i}\right)\right\} \end{aligned}$$

D'où en posant :

$$\begin{aligned} Q(\theta_i) &= ln\left(\frac{p_i}{1-p_i}\right) \\ a(\theta_i) &= 1 - p_i \\ b(y_i) &= 1 \end{aligned}$$

La famille de loi de Bernoulli se met sous la forme canonique 3.2 et peut donc être considérée comme une famille exponentielle. En considérant la somme de n_i variables de Bernoulli, on déduit aisément que la famille de loi Binomiale fait également partie de la famille exponentielle.

Loi de Poisson $\mathcal{P}(\lambda)$

Soient n variables indépendantes de loi de Poisson $\mathcal{P}(\lambda_\rangle)$. Leur densité est donnée par :

$$\begin{aligned} f_Y(y_i|\theta_i,\lambda_i) &= \frac{\lambda_i^{y_i}}{y_i!}e^{(-\lambda_i)} \\ &= exp\{-\mu_i\}exp\{y_i ln(\lambda_i)\}ln\left(\frac{1}{y_i!}\right) \end{aligned}$$

D'où en posant :

$$\begin{array}{rcl} Q(\theta_i) & = & ln(\lambda_i) \\ a(\theta_i) & = & exp\{-\mu_i\} \\ b(y_i) & = & ln\left(\frac{1}{y_i!}\right) \end{array}$$

La famille de loi de Poisson se met sous la forme canonique 3.2 et peut donc être considérée comme une famille exponentielle.

Loi Gamma $\mathcal{G}(p,\lambda)$

Si on considère une variable aléatoire Y de loi Gamma de vecteur paramètres $\theta = (p,\lambda)$, sa densité s'écrit alors :

$$\begin{aligned} f_Y(y|p,\lambda) &= \frac{y^{p-1}exp(-\lambda y)\lambda^p}{\Gamma(p)} \\ &= \frac{\lambda^p}{\Gamma(p)}y^{p-1}exp(-\lambda y) \end{aligned}$$

D'où en posant :

$$\begin{array}{rcl} Q(\theta) & = & -\lambda \\ a(\theta) & = & \frac{\lambda^p}{\Gamma(p)} \\ b(y) & = & y^{p-1} \end{array}$$

La famille de loi Gamma se met sous la forme canonique 3.2 et peut donc être considérée comme une famille exponentielle.

A.3 Principales distributions de la famille exponentielle

TABLE 8 – Tableau des principales distributions de la famille exponentielle

Distribution	**Probabilité / Densité**	**Espérance**	**Variance**
Normale $\mathcal{N}(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathbb{1}_{\mathbb{R}}(x)$	μ	σ^2
Bernoulli $\mathcal{B}(1, p)$	$\mathbb{P}(X = x) = p^x (1-p)^{1-x}$, $x \in \{0, 1\}$	p	$p(1-p)$
Binomiale $\mathcal{B}(n, p)$	$\mathbb{P}(X = k) = \binom{n}{k} p^n (1-p)^{n-k} \mathbb{1}_{\{0,\dots,n\}}(k)$	np	$np(1-p)$
Poisson $\mathcal{P}(\lambda)$	$\mathbb{P}(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} \mathbb{1}_{\mathbb{N}}(k)$	λ	λ
Gamma $\mathcal{G}(p, \lambda)$	$\frac{\lambda^p}{\Gamma(p)} e^{-\lambda x} x^{p-1} \mathbb{1}_{\mathbb{R}_+^*}(x)$	$\frac{p}{\lambda}$	$\frac{p}{\lambda^2}$

A.4 Indicateurs statistiques des modèles de zonier

TABLE 9 – Récapitulatif des indicateurs statistiques des modèles de zonier

Indicateurs	Zonier lissage faible		Zonier lissage modéré	
	Base de modélisation	Base de validation	Base de modélisation	Base de validation
Gini	48,32%	38,06%	36,30%	36,09%
RMSE	1 399	1 350	1 410	1 349
Déviance moyenne	67,32	72,53	71,19	72,67
MAE	261,7	264,6	268,9	272,1