**Mémoire présenté le :**
**pour l'obtention du diplôme**
**de Statisticien Mention Actuariat**
**et l'admission à l'Institut des Actuaires**

Par : ~~Monsieur~~/Madame BERSON Elise

***Titre du mémoire :***

Refonte de la garantie Responsabilité Civile Automobile du produit Garages

Confidentialité : ☐ NON ☒ OUI (Durée : ☐ 1 an ☒ 2 ans)

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

*Membres présents du jury de l'Institut des Actuaires* signature

*Entreprise :*

Nom : AXA France

AXA France I.A.R.D.
Société Anonyme au Capital de 214 799 030 €
Entreprise régie par le Code des Assurances
Siège social : 313, Terrasses de l'Arche
92727 NANTERRE CEDEX
722 057 460 RCS Nanterre

*Signature :*

*Directeur de mémoire en entreprise :*

*Nom :* DA SILVA Laura

*Signature :*

*Invité :*

*Nom :*

*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)*

Signature du responsable entreprise

AXA France I.A.R.D.
Société Anonyme au Capital de 214 799 030 €
Entreprise régie par le Code des Assurances
Siège social : 313, Terrasses de l'Arche
92727 NANTERRE CEDEX
722 057 460 RCS Nanterre

Signature du candidat

*Membres présents du jury de la filière*

Années 2019-2020

**Mémoire de fin d'étude**

M2 Actuariat

Élise Berson

# Refonte de la garantie Responsabilité Civile Automobile du produit Garages

Tutrices :

Entreprise : Laura Da Silva

Académique : Maud Thomas

# Résumé

L'objectif de ce mémoire est la refonte de la garantie Responsabilité Civile Automobile du produit Garages. Cette évolution devra répondre aux problématiques de dégradation de rentabilité et de simplification en tant que fil conducteur du groupe AXA.

Dans le cadre d'une refonte de tarification, les étapes de collecte et traitement des données sont primordiales afin de s'assurer de la qualité de la modélisation. C'est pour cela qu'une part non négligeable des travaux ont été menés en ce sens, à la fois pour la pré-sélection de variables adaptées au risque à modéliser, pour fiabiliser les données existantes, et pour traiter les données manquantes.

Une fois les données collectées et fiabilisées, les étapes de modélisation vont permettre d'expliquer et de comprendre le risque étudié. La garantie Responsabilité Civile Automobile étant ici l'objet de cette étude, le but est d'appréhender au mieux ces spécificités. Dans la mesure où l'indemnisation concerne un tiers, la notion de responsabilité de l'assuré et du (ou des) tiers entre en jeu. Par conséquent, cela se reflète dans la charge des sinistres et peut créer des biais dans sa distribution. L'objectif est alors de choisir une modélisation qui s'adaptera au mieux aux données à expliquer. Les travaux se baseront sur les modèles linéaires généralisés, en tant que pilier de la modélisation non vie, mais les conclusions seront également appuyées par l'apprentissage statistique pour valider les résultats. Enfin, la composante géographique ne sera pas négligée et amènera à la création d'un zonier dédié au risque étudié.

# Abstract

This memorandum aims to overhaul the Motor third party liability coverage of the Garages product. The overhaul of this product will have to answer to the issues of deteriorating profitability and simplification as a guiding principle of the AXA Group.

In the context of a tariff overhaul, the stages of data collection and processing are essential to ensure the quality of the modelling. This is why a significant amount of work has been carried out in this area, both for the pre-selection of variables adapted to the risk to be modelled, to make existing data more reliable, and to process missing data.

Once the data has been collected and made reliable, the modelling stages will enable the risk being studied to be explained and understood. As Motor Third Party Liability cover is the subject of this overhaul, the aim is to gain a better understanding of these specificities. Insofar as the indemnity involves a third party, the notion of the liability of the insured and the third party (or parties) comes into play. Consequently, this is reflected in the burden of claims and may create biases in its distribution. The objective is then to choose a model that will best fit the data to be explained. Modelling will be based on generalized linear models, as a pillar of non-life modelling, but conclusions will be supported by statistical learning to validate the results. Finally, the geographical component will not be neglected and will lead to the creation of a zone dedicated to the risk studied.

# Remerciements

Je tiens avant tout à remercier Laura Da Silva, ma tutrice en entreprise et conseiller d'études actuarielles dans l'équipe Actuariat Produit des risques de fréquence de la Direction Actuariat Pilotage Entreprises (DAPE) chez AXA France pour son encadrement, son implication et son aide compétente tout au long de la réalisation de ce mémoire, sans oublier ses précieux encouragements.

Je remercie également Gérald Lucas, manager de cette équipe, pour son expertise et ses conseils à toutes les étapes de ce projet.

J'aimerais remercier Véronique Marpillat, responsable Actuariat Produit, pour son soutien et pour m'avoir accueilli dans le service.

J'adresse aussi mes remerciements à tous ceux qui ont participé à la réalisation et à la relecture de ce mémoire.

Merci à toute l'équipe de la DAPE pour leur accueil et leur bienveillance..

Pour finir, je tenais à remercier les enseignants de l'ISUP pour leurs apports théoriques, et plus particulièrement ma tutrice académique Maud Thomas pour ses nombreux conseils.

# Table des matières

# Introduction

Le marché de l'assurance se divise en deux grandes catégories. La première comprend les assurances relatives à la durée de vie humaine, à savoir l'Assurance Vie. La deuxième concerne les assurances qui n'ont pas pour objet la durée de vie humaine. Elle regroupe notamment les assurances de responsabilité, de personnes et les assurances dommages, couvrant les assurés en cas de sinistres. Elle est appelée Assurance Non Vie. Ce type de contrat d'assurance se définit comme un contrat établi entre une compagnie d'assurance et un assuré le protégeant contre un événement aléatoire non relatif à sa vie. Cette couverture est assurée contre le versement d'une prime destinée à couvrir la charge des sinistres selon le principe de mutualisation des risques. Le but étant de collecter suffisamment de primes sur l'ensemble du portefeuille, tout en adaptant le montant des primes à la proportion du risque pour chaque contrat. Cette part est estimée lors de la modélisation d'un tarif.

Ce mémoire s'attardera particulièrement sur la garantie principale, en termes de charges comme de primes, du produit Garages. Il s'agit de la garantie Responsabilité Civile Automobile, obligatoire pour les contrats de ce produit mais également pour tous les contrats d'assurance Automobile en France. Cette obligation d'assurance a pour but d'indemniser tous les dommages causés à un tiers, l'objet des prestations versées par l'assurance n'est donc pas l'assuré, mais les victimes d'un sinistre dont l'assuré est responsable, totalement ou partiellement. Plusieurs spécificités sont alors à prendre en compte pour modéliser cette garantie. La première est la définition de la part de responsabilité de l'assuré, dont la valeur va impacter le montant à indemniser. Le seconde est la présence d'une convention, appelée "IRSA", ayant pour but de faciliter et accélérer le paiement de l'indemnisation aux assurés dans ces circonstances, en demandant à chaque assureur d'indemniser totalement les frais de son assuré, et ensuite faire un recours auprès de l'assureur de la personne responsable du sinistre pour récupérer au nom de l'assuré victime les sommes dues. Cependant, cela crée des biais dans la distribution de la charge sinistre de part ces recours qui sont effectués entre assureurs.

Par ailleurs, des contraintes au niveau du produit seront également à prendre en compte, à savoir un objectif de simplification du produit et de la structure de tarification pour répondre aux exigences d'AXA France à ce sujet. En effet, la simplification est un fil conducteur des transformations de la compagnie. De plus, un élément déclen-

cheur de la refonte de ce produit est la dégradation de la rentabilité sur les dernières années. Une analyse des éléments qui ont pu impacter cette dégradation sera détaillée permettra d'orienter le tarif en fonction de ces conclusions.

Afin de répondre à l'ensemble de ces problématiques, ce mémoire s'articulera en plusieurs parties. Dans un premier temps, une présentation plus détaillée du contexte et des enjeux de la refonte de ce produit permettra de mieux comprendre l'intérêt de ce projet. Dans un deuxième temps, une analyse des données disponibles et des retraitements effectués sera donnée. Puis, une fois ces données collectées et fiabilisées, les différentes modélisations retenues pour la garantie Responsabilité Civile Automobile du produit Garages seront présentées.

*NB :* Pour des raisons de confidentialité, les résultats chiffrés présentés ont été modifiés mais les ordres de grandeur ont été conservés pour maintenir le sens des conclusions.

# Chapitre 1

# Présentation et mise en contexte

## 1.1 Enjeux de l'entité

L'entité IARD (Incendie, Accidents, Risques Divers) Entreprises d'AXA France est leader sur le marché des entreprises : elle constitue 15% de la part de marché. Cela représente par ailleurs 2,7 milliards d'euros de chiffre d'affaires (CA) à fin 2019. Sur les dernières années, le groupe AXA s'est focalisé sur le développement de ses entités IARD et notamment sur le marché des Entreprises. Cette étude s'inscrit donc dans cette ambition pour mieux répondre aux besoins de cette clientèle, que ce soit en termes d'adéquation des offres à leurs demandes spécifiques, ou en termes de simplification à toutes les échelles des produits. L'entité AXA IARD Entreprises tend alors à répondre à ces problématiques, et se décompose ainsi :

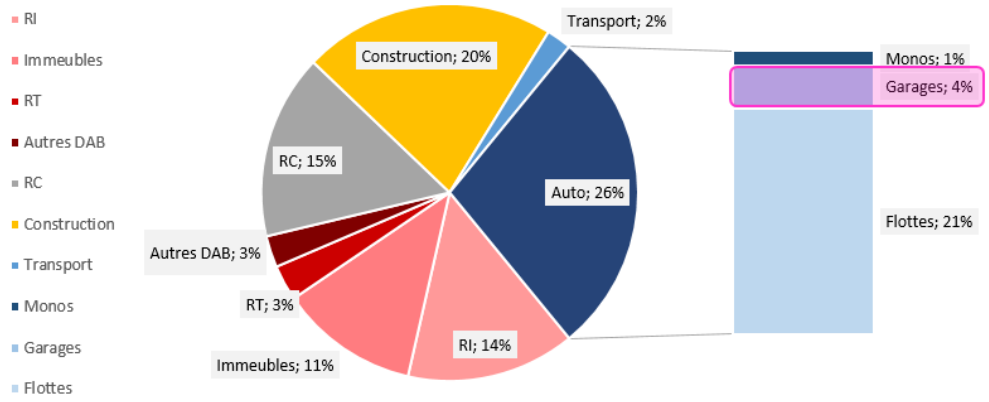



FIGURE 1.1 – Répartition du Chiffre d'Affaires du portefeuille IARD Entreprises à fin 2019

Cette répartition permet de mieux comprendre le positionnement de la branche Automobile et plus spécifiquement du produit Garages qui représente 4% du Chiffre d'Affaires

d'AXA IARD Entreprises. Une présentation plus détaillée de la branche Automobile et plus spécifiquement du produit Garages sera apportée dans la suite pour appréhender leurs problématiques respectives.

## 1.2 Descriptif de la branche Automobile

### 1.2.1 Positionnement sur le marché

Le produit Garages appartient à la branche Automobile de l'entité IARD Entreprises au sein d'AXA France. Une présentation des problématiques liées à cette branche sera donnée, pour ensuite entrer plus en détails sur les enjeux de ce produit.
La répartition du CA montre que la branche Automobile est la plus importante d'AXA IARD Entreprises avec 26% du Chiffre d'Affaires à fin 2019. Cependant, elle doit faire face à des difficultés de rentabilité du fait de la forte concurrence sur le marché. Les produits Automobile correspondent à des produits dit "d'appel" qui visent à proposer du multi-équipement aux clients sur d'autres branches. Le but est alors d'avoir des tarifs très compétitifs qui permettent d'attirer la clientèle, en gardant pour objectif une rentabilité globale sur l'ensemble des produits souscrits par le client.

### 1.2.2 Les produits proposés

L'Automobile Entreprise propose divers produits destinés à couvrir les besoins des professionnels sur toutes leurs problématiques liées à l'automobile. Il existe différents produits pour assurer les véhicules légers au même titre que les poids lourds qui se distinguent principalement par le nombre de véhicules qui composent la matière à assurer du client :

— entre 1 et 4 véhicules de plus de 3,5 Tonnes à assurer, le client sera dirigé vers un contrat dit "Mono", souscrit pour chaque véhicule. La cible correspond à des artisans ou petites sociétés, et il est possible de proposer une tarification proche de celle proposée par l'assurance des particuliers en se basant sur les informations spécifiques à chaque véhicule et à chaque conducteur

— entre 5 et 50 véhicules à assurer, il s'agira de flottes de véhicules. Cette taille de flotte correspond au produit "Parc" pour lequel la tarification se fait véhicule par véhicule grâce aux caractéristiques fournies pour chaque véhicule. Il n'y a pas de tarification selon les caractéristiques du conducteur pour les flottes de véhicules.

— au-delà de 50 véhicules à assurer, le produit concerné est le produit "Flotte". Il correspond à la couverture automobile pour de grandes entreprises ou pour les collectivités locales. La problématique de ce produit est l'absence de détail à ce jour sur les véhicules présents dans la flotte, il n'est alors pas possible de faire une tarification véhicule par véhicule. L'introduction du FVA [1]

1. Fichier des Véhicules Assurés : Fichier à remplir par les assureurs pour recenser l'ensemble des véhicules assurés dans leur portefeuille, notamment dans le but de lutter contre la non-assurance et faciliter l'identification en cas de délits - ce fichier permettra d'avoir le détail des véhicules contrat par

Les véhicules sont distingués selon leur usage, à savoir le transport de marchandises, le transport de voyageurs, ou simplement en tant que moyen de déplacement dans le cadre de l'activité professionnelle.

Un autre produit proposé par la branche Automobile ne dépend pas du nombre de véhicules présents au contrat, et c'est celui au coeur de cette étude : l'assurance des professionnels de l'automobile, c'est à dire le produit des Garages et Concessions. La refonte de ce produit va permettre d'approfondir la connaissance des spécificités de ce dernier par rapport à d'autres produits Automobile dont la tarification est centrée sur les véhicules à assurer.
Le découpage du portefeuille de ce produit pourra s'effectuer grâce aux sous-produits ou à partir d'une segmentation selon les données risques :

| Sous produit | Garages | | Concessions |
|---|---|---|---|
| Segmentation | Bas de segment | Moyen de segment | Haut de segment |
| Règle de découpage | Critères :<br>— une superficie inférieure à $1500m^2$<br>— un chiffre d'affaires annuel inférieur à 2,3M€<br>— la valeur du bâtiment et du contenu cumulée ne doit pas dépasser 1100 fois la valeur de l'indice FFB [1] au moment de la souscription. | Si l'un de ces critères n'est pas respecté | Pas de distinction au sein du sous-produit Concessions |

TABLE 1.1 – Segmentation du produit

## 1.3 Descriptif du produit

Le produit des Garages et Concessions est destiné aux professionnels de l'automobile, à savoir ceux exerçant les activités de : garagistes, de concessionnaires, ou encore de contrôle technique. Le contrat peut couvrir diverses garanties liées à la fois aux dommages aux véhicules de l'entreprise, mais également à ceux des véhicules confiés par la clientèle pendant les réparations ou lors de toute autre intervention sur le véhicule, et en supplément de la garantie Responsabilité Civile Automobile obligatoire pour toute assurance de véhicule. Par ailleurs, et c'est une spécificité de ce produit : il existe des

contrat, et ainsi avoir une tarification proche de celle du produit Parc.

garanties qui couvrent également les dommages aux biens immobiliers et matériels qui composent les locaux de l'entreprise, mais aussi sa responsabilité civile professionnelle. Il s'agit donc d'un produit très complet qui propose des garanties de dommages aux biens, en plus des garanties automobiles présentes dans tous les autres produits de la branche. Ces garanties peuvent être souscrites "à la carte" en respectant l'obligation de souscription de certaines garanties comme la RC automobile. Cependant, il existe également un produit "tout-en-un" pour faciliter la souscription des garages de bas de segment et qui regroupe ces garanties, à la fois automobile, de dommages aux biens et de responsabilité civile professionnelle. Il s'adresse uniquement aux garages qui respectent les critères du bas de segment.

Le portefeuille peut se distinguer aussi en deux sous-produits :

— les Garages

— les Concessions

Cette distinction s'explique par les caractéristiques différentes au niveau du type de risque : l'activité pratiquée dans l'un et dans l'autre n'aura pas des niveaux de risques équivalents. En effet, les concessions ont plutôt une activité orientée vers la vente alors que les garages sont plutôt orientés vers des activités de réparation. Il ne s'agit alors pas du même type de sinistres : les activités orientées vers la vente ont généralement un plus grand nombre de véhicules neufs ou récents sujets aux incendies et catastrophes naturelles qui sont stockés sur une certaine période en fonction des délais pour la mise en vente de ces véhicules. Les activités de réparation sont plus sujettes aux sinistres de RC professionnelle du fait du risque d'erreur croissant avec le nombre de réparations : les véhicules peuvent être plus nombreux que sur les activités de vente mais la durée de stockage dans les locaux est généralement bien plus faible, la typologie des sinistres est donc différente. Par conséquent, dans les contrats cela se traduit actuellement par des garanties proposées différentes. Parmi ces différences il faut par exemple noter que les concessions, étant généralement d'une superficie plus élevée, notamment pour permettre le stockage des véhicules, souscrivent des garanties dommages aux biens de manière indépendante, auprès de produits distribués par la branche des risques industriels, que ce soit chez Axa France ou ailleurs. Cela leur permet d'avoir une couverture plus importante que celle offerte par le produit "tout-en-un" proposé par la branche automobile. L'objectif de cette étude est alors également de prendre en considération ce type de problématiques opérationnelles afin de développer la couverture du produit, voire d'élargir la clientèle potentielle en l'attirant par un produit simplifié et pour une cible plus importante.
Il est possible de résumer ces spécificités dans le tableau suivant :

| Sous-produit | Garages | | Concessions | |
|---|---|---|---|---|
| Garanties | Obligatoire | Facultative | Obligatoire | Facultative |
| RC Automobile (RCAU) | × (véhicules confiés et en vente uniquement) | × (véhicules propriété de l'entreprise) | × (tous véhicules) | |
| Défense-Recours (DR) | × | | × | |
| Dommages accidentels (DOMA) | × (véhicules confiés et en vente uniquement) | × (véhicules propriété de l'entreprise) | × (tous véhicules) | |
| Incendie et Vol (VOLINCAU) | × (véhicules confiés et en vente uniquement) | × (véhicules propriété de l'entreprise) | × (tous véhicules) | |
| Bris de glace Automobile (BDGAU) | | × | × | |
| Sécurité du conducteur (SDC) | | × | × | |
| Assistance | | × | × | |
| Protection juridique | | × | × | |

TABLE 1.2 – Tableau des garanties Automobile

| Sous-produit | Garages | | Concessions | |
|---|---|---|---|---|
| Garanties | Obligatoire | Facultative | Obligatoire | Facultative |
| RC professionnelle (RCNA) | × | | × | |
| Dommages aux biens (locaux et contenu, hors véhicules) (INCNA, VOLNA, DDE, BDGBDMNA) [1] | | × | × | |
| Pertes d'exploitation (PEVV) | | × | × | |

TABLE 1.3 – Tableau des garanties Non Automobile

Une présentation succincte de ces garanties va permettre de mieux comprendre l'offre proposée pour ce produit :

— La garantie RC Automobile est une garantie obligatoire pour tous les contrats d'assurance Automobile en France pour indemniser les dommages causés aux tiers lors d'un accident de circulation ; cette garantie étant l'objet de ce mémoire, une présentation plus détaillée sera fournie

— La garantie Défense-Recours va de pair avec la garantie RC Automobile car elle permet de prendre en charge et les frais judiciaires liés à la défense ou au recours de l'assuré lorsqu'il y a un litige entre les parties sur les circonstances de l'accident

— La garantie Dommages Accidentels permet de prendre en charge les dommages causés à un véhicule assuré sans qu'il n'y ait de tiers impliqué

— Les garanties Incendie et Vol Automobile couvrent les dommages ou le remplacement du véhicule assuré, en cas d'incendie ou de vol de celui-ci

— La garantie Bris de glace Automobile prend en charge les frais liés à la réparation ou au remplacement du pare-brise du véhicule assuré

— La sécurité du conducteur permet de couvrir tous les frais et, le cas échéant, les pertes de salaires du conducteur du véhicule assuré induites par un accident de circulation pour les personnes désignées au contrat

— L'assistance est un service d'aide matérielle, logistique (dépannage) et parfois sociale (soutien psychologique) en cas de sinistre avec le véhicule assuré

— La protection juridique fonctionne comme la garantie Défense-Recours, mais s'applique aux litiges qui opposent l'assuré à un tiers concernant le véhicule assuré en dehors de circonstances d'un accident de circulation impliquant ce tiers. Cela peut être par exemple le cas lors de l'achat, de la vente ou encore de la location du véhicule assuré, mais aussi lors de la réparation ou de l'entretien de celui-ci.

— La RC professionnelle couvre les dommages liés à l'exercice de l'activité professionnelle, généralement lors de la réparation ou la vente de véhicules par exemple, mais également aux dommages causés par l'entreprise à des tiers n'étant pas liés contractuellement.

— La garantie Dommages aux biens concerne l'ensemble des biens (hors véhicules) de l'entreprise et sont couverts pour tous les dommages causés par l'incendie, le vol, les dégâts des eaux, y compris le bris des vitrines ou des machines

— La garantie de pertes d'exploitation permet de couvrir la baisse de chiffre d'affaires de l'entreprise, selon des situations définies au contrat, notamment en cas de fermeture due aux dommages aux locaux

La répartition de ces différentes garanties est la suivante :

| Prime 6 ans | | RCDRAU | DOMA | VOLINCAU | BDGAU | SDC | RCNA | VOLNA | INCNA | DDE | BDGBDMNA | PEVV | CLIM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Garages | Axapac | 18% | 28% | 13% | 4% | 1% | 10% | 8% | 7% | 2% | 6% | 4% | 0% |
| | XF | 21% | 32% | 15% | 5% | 3% | 16% | 2% | 1% | 1% | 2% | 2% | 0% |
| | OSE | 20% | 21% | 14% | 5% | 4% | 14% | 4% | 9% | 3% | 4% | 4% | 0% |
| Concessions | | 19% | 31% | 13% | 4% | 3% | 20% | 2% | 2% | 2% | 2% | 3% | 0% |

| Charge 6 ans | | RCDRAU | DOMA | VOLINCAU | BDGAU | SDC | RCNA | VOLNA | INCNA | DDE | BDGBDMNA | PEVV | CLIM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Garages | Axapac | **24%** | **15%** | **14%** | 4% | 2% | 12% | 5% | 19% | 1% | 1% | 2% | 2% |
| | XF | **29%** | **19%** | **23%** | 4% | 2% | 15% | 1% | 6% | 0% | 0% | 1% | 1% |
| | OSE | **19%** | **18%** | **32%** | 4% | 1% | 10% | 2% | 12% | 0% | 0% | 0% | 2% |
| Concessions | | **21%** | **15%** | **36%** | 3% | 1% | 15% | 2% | 3% | 1% | 0% | 1% | 2% |

FIGURE 1.2 – Répartition des garanties -par sous-produit et outil de souscription

La principale garantie de ce produit, quel que soit l'outil de souscription utilisé (Axapac, XF ou OSE, qui seront détaillés par la suite) et pour les deux sous-produits, est la garantie Responsabilité Civile AUtomobile (RCDRAU), et cela que ce soit en termes de sinistralité comme de Chiffre d'Affaires apporté. Cette étude sera donc basée sur la modélisation de cette garantie.

## 1.4 Descriptif de la garantie étudiée

### 1.4.1 La garantie Responsabilité Civile automobile

#### Responsabilité Civile Automobile

La garantie Responsabilité Civile automobile (RC automobile) est présente et obligatoire dans tous les contrats d'assurance automobile afin de pouvoir circuler en France. D'après la définition de l'administration française, elle doit couvrir la réparation de tous les dommages corporels et matériels causés à autrui et causés lors d'un accident de la route. Il existe cependant des exclusions de prise en charge qui sont indiquées dans les conditions générales de chaque contrat, à savoir que les exclusions générales sont les suivantes :

— les dommages causés intentionnellement par l'assuré

— les dommages causés par le transport de matières dangereuses

— tous les comportements qui relèvent de l'illégalité (ex. conduire sans permis, en état d'ivresse)

— ce qui est du fait des sous-traitants

— les dommages causés par la guerre (civile ou étrangère)

En assurance automobile il existe également des franchises et des limites de prise en charge mais elles ne sont pas applicables pour la prise en charge des dommages corporels dans le cadre de la RC Automobile du fait du principe de réparation intégrale inhérente aux dommages causés aux tiers. Il n'y a pas non plus de franchise sur les dommages matériels dans ce cas-là mais il existe cependant une limite d'indemnisation pour ces dommages qui est fixée à 10 millions d'euros par sinistre et dans la limite de 7,6 millions par véhicule. Il s'agit de limites d'indemnisations propres au contrat mais dont le montant doit être supérieur à celui imposé par la législation qui est de 1,22M€ [1].

#### Volumes modélisés sur le portefeuille

Les volumes de cette modélisation vont être détaillés afin de mieux comprendre les ordres de grandeurs, notamment pour la suite des résultats et analyses. Il s'agit des volumes vus à fin 2018 afin de prendre en compte des problématiques de vieillissement de la charge sinistre qui seront détaillées dans la suite de ce paragraphe.

1. Ce montant est fixé par l'État et revu tous les 5 ans en fonction de l'inflation au niveau européen.

**Production** La garantie RC Automobile du produit Garages représente 15 000 contrats en cours à fin 2018 et 100 000 véhicules assurés dans le cadre de l'activité du professionnel. Les véhicules confiés pour réparation par la clientèle ne sont pas comptabilisés ici : ces derniers ne sont assurés que dans le cadre d'un dommage lié à cette prise en charge et ne sont alors pas déclarés ligne à ligne dans le contrat du professionnel contrairement aux autres véhicules, dits "véhicules propriétaires". Ces contrats représentent 20M€ de primes acquises pour la garantie RC Automobile à cette même date. Pour rappel, la prime acquise correspond à la part, sur l'année civile, de la prime payée par l'assuré, qui est à retenir pour la période de temps durant laquelle le contrat est effectif. Par exemple, pour un contrat dont la date de prise d'effet est au 1er juillet, la valeur de la prime acquise sera de 50% de la valeur annuelle de la prime émise (cotisation couvrant les primes jusqu'à l'échéance du contrat).

**Sinistralité** En termes de sinistralité, le nombre de sinistres déclaré en 2018 au titre de cette garantie s'élève à 9 000 en 2018, ce qui correspond à 27M€ de charges sinistres. Le rapport Sinistres à Primes, ou Sinistres à Cotisations (S/C) est particulièrement dégradé sur cette année-là : il est de 94,7% ; mais parmi ces sinistres, 200 sont considérés comme "graves", c'est-à-dire lorsqu'ils dépassent un montant de 30 000€. Ce montant est fixé de manière commune sur la branche Automobile afin de faciliter le pilotage de l'ensemble des produits. Ainsi, si ces sinistres sont exclus de la charge, le produit voit son ratio diminuer pour atteindre ratio S/C de 89,2%. De plus, parmi ces sinistres il existe des sinistres dits "atypiques", c'est-à-dire qui s'élèvent à plus de 2M€ de charge, ils sont au nombre de 2 sur l'année 2018 mais représentent 37% de la charge totale. L'inclusion de ces sinistres pourra faire l'objet de retraitements dans la modélisation du fait de leur faible nombre avec une charge très importante qui biaise la sinistralité observée.
La garantie RC Automobile est par ailleurs impactée par des montants de sinistres négatifs qui sont liés aux recours effectués entre assureurs dans le cadre de la Convention IRSA. Cette convention et les conséquences sur la modélisation feront l'objet d'explications plus détaillées par la suite.
De plus, il est important de noter que dans le cadre de cette étude, la charge sinistre considérée est évaluée 12 mois après la survenance pour avoir une charge la plus proche de la charge réelle du sinistre, et non pas un montant forfaitaire attribué lors de l'ouverture du sinistre par le gestionnaire. En effet, cela est le cas si la valeur réelle n'est pas connue à la date de déclaration du sinistre. Ce délai de 12 mois n'est pas toujours suffisant pour les sinistres corporels qui ne sont pas systématiquement stabilisés au bout d'un an mais cela concerne généralement les sinistres atypiques qui seront ici retraités avant modélisation. Cette durée retenue pour l'évaluation des sinistres permet de concilier deux critères importants dans le cadre d'une modélisation :

— Avoir une vision récente de la sinistralité : l'historique concerne les années de survenance 2016 à 2018, avec les sinistres à vision respective de 2017 à 2019. La période retenue correspond ainsi aux dernières données connues, en prenant en compte le vieillissement de 12 mois de la charge

— et considérer une vision vieillie pour être au plus proche de la sinistralité réelle.

Le descriptif de la garantie étudiée peut maintenant se poursuivre avec les éléments de tarification utilisés dans le produit distribué actuellement.

### Modélisation dans le tarif actuel

**Garages** A ce jour, plusieurs formules de tarifications sont utilisées selon s'il s'agit d'une entreprise exerçant une activité de garagiste ou une activité de concessionnaire. Cette distinction s'est effectuée par la différence de typologie de risque entre les garages qui ont majoritairement une activité de réparation et les concessionnaires qui ont une activité orientée vers la vente de véhicules.
La sous-catégorie des Garages concerne 90% du portefeuille du produit en nombre de contrats et 70% des primes, une présentation plus en détails du tarif appliqué pour ce segment va être donnée. Sur ce segment, la prime pure de cette garantie est donnée par :

$$PP = (RCOT + (NVDIR \times RCOTDIR)) \times ANTE \times ZONAU \times RCLS$$

avec

— RCOT : la cotisation de base dont la valeur dépend de l'effectif de l'entreprise

— NVDIR : le nombre de véhicules qui appartiennent au dirigeant pour son utilisation propre, il s'agit du, ou des seuls véhicules pour lesquels le conducteur est connu puisqu'il s'agit de l'assuré.

— RCOTDIR une valeur fixe strictement positive dont le but est de modéliser le risque lié aux véhicules du dirigeant

— ANTE : une variable calculée à partir des sinistres déclarés au moment de la souscription à travers le relevé d'information édité par le précédent assureur. Elle correspond au ratio du nombre de sinistres automobiles et non automobiles (RC professionnelle notamment) sur l'effectif et pondéré par la durée de sinistralité fournie (entre 1 et 3 ans). De plus, lorsque l'entreprise a été assurée depuis moins d'un an, comme pour une entreprise en création, la valeur de la variable sera fixée à sa valeur la plus élevée telle que définie dans la tarification, à savoir 1,5

— ZONAU : une variable définie à partir d'un zonier élaboré par le marché des Particuliers et Professionnels pour la garantie RC Automobile de ce marché. Ce zonier est à la maille du Code INSEE du risque et est composé de 13 zones

— RCLS : une variable dépendant du type d'activité de l'entreprise, elle se décompose en 26 classes d'activités (cf A.3)

La formule présentée ici n'est pas sous un format uniquement multiplicatif ou additif, mais selon une combinaison des deux. Par conséquent, une mise à jour des coefficients associés à chaque variable semble délicate : la modélisation ne pourra pas être reproduite selon le même schéma de formule. De plus, il n'y a pas de constante pour le tarif qui serait

commune à tous les contrats. Le but est donc également d'obtenir une formulation plus standard pour faciliter la mise à jour du tarif à l'avenir et ce en considérant la modélisation sous un format multiplicatif. Ce format permet de mesurer l'impact de chaque variable indépendamment des autres, et de procéder aux modifications nécessaires de l'un ou plusieurs des coefficients à partir des informations recueillies grâce au suivi du produit : le pilotage est ainsi simplifié. Par ailleurs, l'implémentation informatique des nouveaux tarifs sera facilitée par une structure multiplicative. Il s'agit donc d'une amélioration profitable à tous les niveaux. De plus, il est important de rappeler qu'une ligne directrice portée par AXA France est la simplification, cela s'applique sur cette refonte et pour toutes les étapes :

— au moment de la souscription : avec une compréhension facilitée pour les distributeurs et par voie de conséquence pour les clients potentiels

— tout au long du suivi du produit avec la possibilité de cibler des typologies de contrats ou de garanties qui dégraderaient la rentabilité globale

**Concessions** Le tarif présenté est celui dédié principalement aux Garages, il existe un tarif dédié pour le sous-produit Concessions, dont la structure pour la garantie RC Automobile est la suivante :

$$
\begin{aligned}
\text{PP Concessions} \;=\; & (\text{cotisation RC fixe} \\
& + \;(\text{montant fixe} \times \text{effectif}) \\
& + \;\text{cotisation associée à l'assurance de véhicules supplémentaires } {}^{1}) \\
& \times \;\text{coefficient d'activité principale} \\
& \times \;\text{coefficient de la zone du risque} \\
& \times \;\text{coefficient des antécédents de sinistralité}
\end{aligned}
$$

La structure n'est donc pas identique mais les données risques utilisées sont similaires, ce qui laisse la possibilité de faire une analyse combinée des deux sous-produits sur une même base d'étude.
Le tableau suivant va permettre de résumer les éléments propres à chacun de ces sous-produits :

| | Garages | Concessions (hors Besse et Cetri) |
|---|---|---|
| Volumes | 14 000 contrats pour 14,4M€ de primes acquises à la garantie | 1000 contrats pour 3,3M€ de primes acquises à la garantie |
| Rentabilité y compris graves (S/C) | 90,6% | 109,4% |
| Outils de souscription | AXAPAC, XF et OSE TPC | XF et OSE TPC |
| Variables tarifaires | — Effectif<br>— Activité principale<br>— Nombre de véhicules du dirigeant<br>— Code postal du risque<br>— Antécédents de l'entreprise | — Effectif<br>— Activité principale<br>— Nombre de véhicules total<br>— Code postal du risque<br>— Antécédents de l'entreprise |
| Structure du tarif actuel | $(A + (B \times C)) \times D \times E \times F$ | $(Z + (Y \times X) + W) \times (V \times U \times T)$ |

TABLE 1.4 – Récapitulatif de la structure du produit

Ce tableau permet de noter la différence structurelle et volumétrique entre les deux sous-produits : les Concessions ayant des primes bien plus élevées avec une sinistralité plus dégradée que les Garages.

La suite de la présentation du produit est suivie par quelques éléments de contexte afin de comprendre dans quel cadre et selon quels objectifs l'étude se place pour ainsi mieux appréhender les enjeux de la refonte du produit.

## 1.5 Objectifs de l'étude

### 1.5.1 Évolution de l'offre proposée

La refonte du tarif Garages a plusieurs objectifs dont le but final est d'améliorer la compétitivité, la compréhension et la facilité de souscription du produit afin de répondre aux différentes problématiques présentées plus tôt.

#### Remontées issues des distributeurs

L'une des initiatives à l'origine de cette refonte provient des remontées des distributeurs et souscripteurs concernant ce produit. Elles sont diverses, mais le manque de compétitivité du produit face à la concurrence ou encore le manque de lissage du tarif se retrouvent parmi les plus récurrentes. Cette notion concerne notamment la variable d'effectif, qui est au coeur du tarif : l'ajout d'un employé aurait un impact très fort sur

le tarif alors que le risque n'en serait pas affecté d'autant. Le montant des limites des garanties est également remis en cause car jugé insuffisant, cela est notamment valable pour l'offre "tout-en-un" qui regroupe toutes les garanties utiles à la cible des professionnels de l'automobile. En effet, les garanties de type "Non Automobile", c'est-à-dire les garanties de dommages aux locaux et au contenu seraient trop faibles pour une partie de notre portefeuille, qui oblige certains clients à prendre deux contrats pour leur entreprise :

— un contrat Automobile pour toutes les garanties de type RC et de dommages aux véhicules

— un contrat Risques Industriels pour toutes les garanties de dommages aux biens avec des limites plus élevées

La refonte du produit vise donc également à augmenter ces limites, c'est pourquoi une analyse en ce sens sera présentée dans cette étude.

### Étude de l'extension du produit "tout-en-un"

**Chiffrage de l'extension** Dans le cadre de la refonte du produit, l'objectif est de standardiser et simplifier la structure globale pour s'adapter au mieux à la demande des distributeurs et du marché. Plusieurs possibilités pourraient permettre d'étendre la clientèle du produit, et d'améliorer son attractivité. L'une d'elle est l'extension de l'offre "tout-en-un" qui est aujourd'hui soumise aux conditions suivantes :

— la surface des locaux doit être inférieure à $1500m^2$

— le Chiffre d'Affaires annuel doit être inférieur à $2,3M€$

— la valeur du bâtiment et de son contenu doit être inférieur à 1100 fois la valeur de l'indice FFB, de l'ordre de 980€.

Une analyse a ainsi été menée pour estimer les gains potentiels d'une extension de ce périmètre en termes de surface et de chiffre d'affaires. A ce jour, les clients dont l'entreprise n'entre pas dans ces critères peuvent souscrire un contrat avec des garanties Automobile à travers le produit de la branche concernée, et des garanties de dommages aux biens, grâce au produit Garages de la branche Risques Industriels, d'AXA France ou d'une autre compagnie. Cependant, il est également possible pour les clients qui entrent dans les critères du produit "tout-en-un" de souscrire uniquement un contrat avec des garanties Automobile, et de souscrire les garanties dites "non automobiles" par ailleurs, chez Axa France ou tout autre compagnie.
Le but est alors de chiffrer à la fois le gain associé aux clients qui pourraient potentiellement être récupérés sur le produit global de la branche Automobile grâce à cette extension, mais également d'établir des marges de précaution sur ces chiffrages liées au fait que certains clients décident de ne pas souscrire le produit "tout-en-un" malgré le fait qu'ils entrent dans les critères de cette offre. Les raisons de ce choix peuvent être liées aux plafonds de garanties proposés par la branche Automobile, qui sont inférieurs à ceux proposés par la branche Risques industriels. Cependant, le but n'est pas ici d'augmenter les plafonds de garantie et faire concurrence au produit cette branche, mais bien de faciliter le parcours client en permettant à davantage d'assurés potentiels de souscrire un unique produit qui couvre tous les risques liés à leur activité.

**Variables cibles de l'extension** L'étude de l'extension concerne deux des trois critères de l'offre actuelle, à savoir celui de surface et celui de CA :

— la surface pourrait être étendue à $3000m^2$ (au lieu de $1500m^2$)

— le chiffre d'affaires pourrait être étendu à $3M€$ voire $5M€$

L'une des principales problématiques a été de récupérer les informations de CA et de surface des contrats souscrits par la branche Risques Industriels. En effet, la tarification ne s'établit pas selon les mêmes critères, de ce fait ces informations étaient bien moins renseignées, le chiffrage est donc partiel. : seuls 40% des clients Garages du portefeuille RI, ayant également un contrat Auto, ont ces informations renseignées.
La cible principale de cette extension, et qui peut être chiffrable est l'ensemble des Garages ayant des garanties Auto, dont la surface est comprise entre $1500m^2$ et $3000m^2$ et/ou dont le Chiffre d'affaires est compris entre $2,3M€$ et $5M€$, sachant qu'elle ne doit dépasser aucun de ces deux critères pour être dans la cible.
Les résultats obtenus permettent ainsi de montrer que cette extension ne serait potentiellement intéressante que pour une extension de la surface à $3000m^2$ et de CA à $3M€$ car trop peu d'entreprises ayant un chiffre d'affaire supérieur entrent dans cette catégorie de surface. De plus, ce potentiel nouveau périmètre pourrait apporter un gain de l'ordre de 1% des primes Garages de la branche Automobile. La décision doit encore être prise auprès des personnes concernées mais le gain, même s'il est mesuré uniquement sur 40% du portefeuille RI, semble insuffisant pour négocier cette extension.

**Impact du mode de gestion du contrat** Une autre extension possible et qui a été étudiée est l'extension de la limite forfaitaire. Cette limite concerne la gestion des contrats :

— les contrats à gestion indexée ont une prime qui est automatiquement revue à chaque échéance en fonction de la valeur d'un indice donné, cela concerne tous les contrats ayant des garanties de dommages aux biens (locaux et contenu), l'indexation se base sur l'indice de la FFB (Fédération Française du Bâtiment).

— les contrats à gestion révisable sont des contrats dont la prime peut être révisée à chaque échéance selon l'évolution soit de la valeur du CA le plus souvent, soit de celle de l'effectif, cette évolution peut être à la hausse comme à la baisse

— les contrats à gestion forfaitaires sont des contrats dont l'évolution de la prime n'est pas soumise à celle d'un indice ou de la valeur du CA ou de l'effectif à l'échéance. L'assureur comme l'assuré ne peuvent alors pas bénéficier de l'évolution de ces données clients, que ce soit à la hausse comme à la baisse. Cependant, cela peut avoir plusieurs avantages pour les deux parties : cela permet de diminuer les frais liés à la gestion de la révision des contrats pour l'assureur, et permet à l'assuré d'avoir une prime plus stable, et n'étant pas impactée par des hausses de chiffres d'affaires ou d'effectif le cas échéant.

Cependant, la gestion forfaitaire est limitée à des assurés donc le CA annuel à la souscription est inférieur à 1M€, car cette gestion permet de faciliter l'expérience client sans créer trop de pertes pour l'assureur, mais plus le CA augmente plus les évolutions

de ces données peuvent être importantes et très impactantes sur le risque et donc sur la prime associée.
La proposition serait alors de donner la possibilité à plus de clients d'avoir ce type de gestion, à savoir augmenter la limite de 1M€ à 1,5M€. Cependant, avant d'étendre cette limite, il est important de chiffrer l'impact que cela pourrait avoir sur les potentiels gains et pertes liées aux évolutions des données
Sur la part des contrats qui peuvent être concernés par cette extension, une analyse sur les dernières années a été menée pour mesurer les évolutions à la hausse ou à la baisse de ces contrats en gestion révisable. Les résultats ont montré que sur cette tranche de CA, il y avait autant de contrats avec un CA à la hausse (évolution supérieure à 5%) que de contrats avec un CA à la baisse (évolution inférieure à $-5\%$). Dans ce contexte, il serait alors intéressant de faire migrer ces contrats vers une gestion forfaitaire étant donné qu'il n'y a pas plus de contrats à la hausse qu'à la baisse, ce qui aurait permis une évolution globale des primes à la hausse sur cette tranche.
Cependant, il faut également analyser les résultats en termes de montants : en effet, si les hausses sont plus importantes que les baisses en montants, alors le gain côté assureur pourrait être positif si cela couvre et dépasse les frais liés à cette gestion. Sous cette vision, la conclusion n'est alors pas tout à fait identique à celle en termes de nombre de contrats : la différence entre les primes des contrats à la hausse et celles des contrats à la baisse est positive, ce qui signifie que le gain pour l'assureur sur cette tranche de contrats est positif. Néanmoins, ce delta représente en moyenne sur trois ans $1,5\%$ de l'assiette de primes. Il n'est pas aisé de chiffrer la part des frais liés à la gestion des contrats concernés, mais le gain en gestion révisable semble faible au regard des potentiels gains apportés par un passage à une gestion forfaitaire. La préconisation pour la mise à jour de quelques éléments du tarif et par la suite pour sa refonte totale serait alors d'étendre cette limite forfaitaire à $1,5M€$ de CA annuel, au lieu de $1M€$.
En conclusion, l'extension du produit serait mise en place à travers l'extension de la limite du mode de gestion forfaitaire qui permettrait de gagner en frais généraux, poste non négligeable.

### 1.5.2 Simplification du tarif actuel

#### Format du tarif

Une autre problématique est le fonctionnement du tarif qui, comme cela peut se remarquer sur les formules par garantie, peut avoir un format peu compréhensible : la mise à jour du tarif au fil des évolutions du risque n'est donc pas aisée. Cela explique effectivement en partie pourquoi le tarif n'a pas été mis à jour depuis 2006, hormis quelques correctifs en 2014 sur les coefficients d'antécédents les plus élevés qui ont été augmentés, et sur le coefficient de l'une des activités principales. Une autre explication de cette version assez datée du tarif est la relative bonne rentabilité du produit par rapport aux autres produits de la branche Automobile. La priorité n'a alors pas été donnée à ce produit, mais les remontées de plus en plus fréquentes des souscripteurs, la dégradation de la sinistralité et la gestion des Garages souscrits sur le marché des Professionnels chez

AXA France par l'entité Entreprises ont ré-ordonné les priorités depuis.

#### Création d'un nouvel outil de souscription dédié

La refonte totale du tarif Garages et Concessions entre également dans le plan de migration du nouvel outil de souscription OSE, qui se développe au fur et à mesure au sein de l'entité IARD Entreprises. En effet, à ce jour l'outil de souscription sur la majeure partie du portefeuille est AXAPAC, un outil qui a été mis en place il y a de nombreuses années et toujours opérationnel, qui se révèle être une plateforme peu ergonomique et accueillante, qui tend à disparaître. D'autres outils de souscription sont également disponibles pour les contrats du haut de segment mais ils sont peu adaptés aux spécificités du produit et dégradent ainsi les remontées de données risques sur les contrats. La création de ce nouvel outil permettra de regrouper la souscription de l'ensemble des contrats Garages et Concessions en un seul outil dédié.
Cependant, la création de cette nouvelle plateforme ayant été reportée au vu de certaines contraintes, la refonte devrait dans un premier temps permettre de procéder à la mise à jour de certains éléments du tarif actuel malgré sa complexité. La refonte globale aurait alors lieu dans un second temps et se baserait sur cette étude pour mettre en place le tarif lorsque le nouvel outil sera disponible, et ce en respectant les objectifs de simplification de rentabilité fixés. Cela permettra en effet de pouvoir implémenter rapidement ce produit le moment venu, en mettant à jour si besoin les coefficients, ce qui sera alors facilité avec un tarif par garantie au format multiplicatif.

### 1.5.3 Problématique liée à la rentabilité du produit

La refonte du produit s'explique également par des raisons plus quantitatives avec une rentabilité qui serait en déclin. Une analyse plus poussée dans le cadre de cette étude a permis de cibler la source de ce déclin et de mieux comprendre l'origine de ces évolutions en termes de production de chiffre d'affaires comme de sinistralité. Les résultats ci-dessous permettront de mieux comprendre le contexte quantitatif de cette étude.

#### Contexte de production

Une analyse sur 4 ans d'historique a permis de montrer que la tendance de l'évolution de la croissance active est à la hausse depuis 2017, avec un chiffre d'affaires (CA) des affaires nouvelles nettement supérieur au CA des résiliations en 2019.

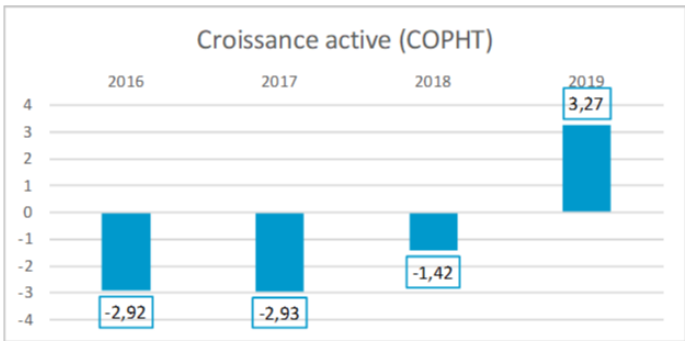



FIGURE 1.3 – Évolution de la croissance active du produit sur tous les marchés, sur 4 ans d'historique

Cet apport de production est porté par le segment des activités de garagistes du bas de segment, ce type de contrats est principalement apporté par le réseau des agents généraux AXA, et ils représentent la plus grande part du portefeuille (46%). La hausse de la production s'explique également par deux chaînes de Concessions qui ont un accord spécifique avec AXA France : Besse et Cetri, ce qui permet d'avoir de très bons chiffres de production sur le haut de segment, mais sans ce partenariat, les conclusions sur ce segment seraient inverses. Le moyen de segment a en revanche une tendance stable avec la présence de certains pics à la hausse ou à la baisse sur la croissance. La refonte du tarif pourrait ainsi permettre d'améliorer la compétitivité sur certains segments du produit, notamment sur les concessions, en poursuivant la tendance sur les garages du bas de segment.

### Contexte de sinistralité

L'évolution positive de la production de manière globale est à appréhender avec précautions au regard des résultats de sinistralité qui se dégradent sur tous les marchés et segments du produit. La croissance active positive et cette dégradation de la rentabilité pourraient laisser supposer que les affaires nouvelles souscrites qui ont permis la croissance active positive ne sont pas de "bons risques" et expliqueraient la hausse de la sinistralité. Une approche de la sinistralité par génération d'affaires nouvelles a pu permettre de mieux comprendre ce phénomène.

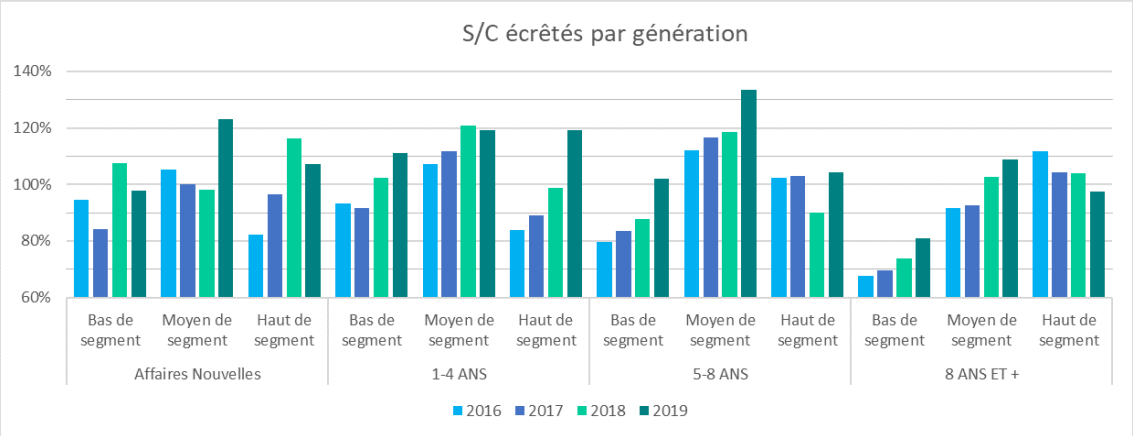



FIGURE 1.4 – Ratios de sinistralité par génération d'ancienneté - pour chaque segment, sur 4 ans d'historique

Il est important de remarquer que la hausse de la sinistralité concerne l'ensemble des générations, ce qui signifie que ce n'est pas l'apport de nouveaux contrats qui seraient de "mauvais risques" et qui expliquerait la dégradation de la rentabilité mais bien un effet global sur le portefeuille. La refonte du tarif va donc permettre de prendre en compte cette hausse globale de la sinistralité pour améliorer à terme sa rentabilité de façon générale. Une analyse de l'évolution par garantie a en outre permis de mieux comprendre l'origine de la sinistralité, afin d'affiner l'orientation du tarif par garantie.

Une baisse de la sinistralité s'observe depuis 2016 sur la garantie RC Automobile du bas de segment, qui représente la majorité du portefeuille, mais elle ne compense pas la dégradation d'autres garanties à savoir les garanties dites "Non Auto" comme la RC Professionnelle ou l'Incendie et le Vol des locaux, comme cela s'observe clairement sur le moyen de segment, mais également sur le bas de segment même si les garanties non auto ont ici été regroupées. Il faut noter que la garantie RC Auto qui sera l'objet de cette étude est prépondérante en termes de sinistralité sur tous les segments, ce qui permet d'avoir une base de sinistralité importante par rapport à d'autres garanties comme la Perte d'Exploitation.

# Chapitre 2

# Analyse des données du portefeuille

L'analyse des données va se découper en deux parties :

— un recensement des données à disposition et des données retenues pour l'étude

— une analyse statistique de ces données par des tris à plat et une analyse des corrélations

## 2.1 Périmètre de l'étude

**Sources des données**

Le produit Garages se décompose en plusieurs sources d'outils de souscription. Ces derniers n'ont pas toujours les mêmes données risques renseignées, ce qui crée des asymétries de complétude des données par outil. Cependant le but est de considérer une base de données qui comprend toutes les sources pour prendre en compte les spécificités de tous les contrats. En effet, les différentes sources correspondent généralement à des typologies de contrats différentes.
L'outil de souscription majoritairement utilisé est AXAPAC, il représente 70% du portefeuille du produit. Il s'agit d'un outil avec un visuel plutôt daté mais qui reste actuellement le plus adapté et le plus complet en attendant la création du nouvel outil dédié : OSE Garages (Outil de Souscription Entreprises). Deux autres outils sont également utilisés pour ce produit : XF et OSE TPC (Tous Produits Complet) : le but de ces outils est de s'adapter à plusieurs produits sur différentes branches afin de faciliter la souscription des contrats haut de segment de manière générale, alors qu'AXAPAC est utilisé pour les contrats de bas et moyen de segment. Ces derniers ont des visuels plus accueillants mais la récupération des données risques n'est pas aussi fournie, or comme il s'agit de contrats plutôt haut de segment, ils ne sont donc pas à négliger, et ce malgré la proportion de données manquantes plus importante.

**Activités concernées**

La refonte du tarif concerne l'ensemble des activités des professionnels de l'automobile. Aujourd'hui une tarification différente est faite entre les Garages et les Concessions mais dans le cadre de la simplification de l'ensemble des procédures, le nouveau tarif aura la même base de tarification, et les différences entre ces deux catégories s'opéreront à travers les coefficients appliqués aux données risques, notamment à travers l'activité pour distinguer, le cas échéant, les différences de risques qui peuvent exister. Cette simplification se justifie d'autant plus que les données risques sont similaires : une différenciation par le montant du tarif et non par l'offre proposée serait appropriée.
Parmi les activités de garagistes il est possible de distinguer notamment :

— les garages qui réparent des véhicules de moins de 3,5T,

— les garages qui réparent des véhicules de plus de 3,5T,

— les garages qui font de la réparation rapide uniquement

— les entreprises qui exercent l'activité de Contrôle Technique. Il faut noter que cette activité doit être la seule activité exercée par un professionnel de l'automobile. Afin d'éviter les conflits d'intérêts et dans le cadre de la sécurité des automobilistes, l'activité de Contrôle technique doit être totalement indépendante des activités de vente et de réparation, ce qui implique une obligation d'activité unique alors que les autres activités peuvent exercer des activités secondaires comme la vente pour un réparateur par exemple.

## 2.2 Variables disponibles

### 2.2.1 Agrégation des différentes sources disponibles

Les serveurs qui hébergent les bases de données d'AXA France permettent d'accéder à de nombreuses informations, le but est alors de rassembler pour chaque numéro de contrat toutes les informations disponibles à son sujet afin d'avoir une modélisation la plus complète possible.

**Bases contrats**

La première étape de la constitution de la base de données pour la tarification est d'isoler tous les contrats qui concernent le produit étudié, sur un historique fixé à cinq ans. Cet intervalle de temps a été défini de telle sorte qu'on puisse obtenir trois années de vision de contrats en cours, en ayant pour chaque année de vision un historique de données de deux ans pour permettre l'étude de l'évolution de certaines données du contrat comme le chiffre d'affaires ou autres modifications récentes apportées au contrat.
Le but étant d'obtenir le plus d'informations pertinentes sur chaque contrat, plusieurs bases sont utilisées en pratique :

— celle des contrats encore en cours actuellement,

— celle des contrats ayant été résiliés à ce jour, afin de récupérer les contrats qui étaient en cours sur la période de l'étude,

— celle qui recense les informations plus générales de chaque client et non spécifiques aux informations dédiées au contrat comme son SIRET, ou encore la région et le distributeur dont il est issu

L'ensemble de ces bases permettent ainsi l'obtention des variables suivantes :

— les dates de création du contrat, et de résiliation le cas échéant,

— pour chaque année civile étudiée, la période pendant laquelle le contrat a été présent,

— la prime annuelle associée au contrat, en notant que plusieurs types de primes sont associées : prime acquise, prime émise comptable, Cotisation potentielle annuelle HT (CoPHT) ; la prime acquise sera celle retenue dans le cadre de cette étude pour représenter au mieux la prime associée au risque en fonction de la présence effective du contrat. Pour la validation de la modélisation, la variable CoPHT sera aussi utilisée, afin de la comparer avec les prédictions.

**Bases SIREN**

Les données relatives à l'activité de l'entreprise sont des données clients qui ne sont pas toujours correctement renseignées ni mises à jour régulièrement dans les bases de contrats. En effet, les données des bases contrats ne sont généralement mises à jour que lorsque le contrat est dit "révisable". Cela signifie que sa prime va dépendre de l'évolution soit de son CA soit de l'effectif de son entreprise. Il s'agit d'éléments dont l'évolution peut être très variable et dont l'impact est significatif sur le risque associé au contrat, et par conséquent sur la couverture et les primes associées. Les contrats de ce type représentent 25% du portefeuille en termes de nombre de contrats. De même, l'assuré peut déclarer une information sur l'évolution de son risque, ce qui aboutit alors à un remplacement du contrat, ainsi d'autres données peuvent être mises à jour à cette occasion. En revanche, dans les autres situations, les données risques du contrat qui sont renseignées dans les bases citées ci-dessus sont des informations fournies à la souscription.
Afin d'améliorer la précision de données impactantes comme le CA ou l'effectif de l'entreprise assurée, d'autres sources sont utilisées et disponibles : il s'agit des données issues de l'INSEE. Elles permettent de récupérer ces deux informations avec une version plus récente. Néanmoins, cette donnée ne peut être récupérée uniquement pour les assurés dont l'information du SIRET est correctement complétée dans les bases initiales.

**Bases des données risques**

Les données liées au risque assuré ne sont pas stockées avec les bases relatives aux informations administratives du contrat. Ces données sont les informations déclarées à la souscription pour la majeure partie des variables, hormis le CA ou l'effectif pour certains contrats comme précisé plus tôt, du fait du caractère révisable de la prime selon l'évolution de ces données.

Les contrats étant issus de plusieurs outils de souscription différents, les sources des données risques sont également multiples. Pour la majeure partie des contrats, c'est à dire des contrats souscrits sur AXAPAC, les données sont disponibles sur des bases créées annuellement pour tous les contrats en cours durant l'année considérée. Il faut alors historiser ces données pour obtenir les valeurs des variables risques pour chaque année d'observation et ainsi réunir ces informations annualisées dans une unique base.
En ce qui concerne les contrats issus des autres outils de souscription comme XF et OSE plus récemment, les tarifs n'étant pas totalement identiques, les données déclarées ne sont pas exactement les mêmes que celles sous AXAPAC : parmi les différences les plus notables il faut noter l'absence de renseignement des variables de CA, d'effectif, et de la forme juridiques qui sont des variables impactantes dans le tarif. Par ailleurs, de nombreuses données risques liées aux garanties non auto ne sont pas renseignées, ce qui est lié au fait que le produit "tout-en-un" proposé concerne principalement des contrats de bas et moyen de segment, alors que ces deux outils sont plutôt dédiés aux contrats de haut de segment.
Ces données sont récupérées sur une source appelée "Données libres". Cette dénomination est liée au fait qu'il n'y aucun formatage ni obligation d'entrer les données déclarées à la souscription. Il faut alors associer à chaque variable la donnée correspondante afin de récupérer le plus d'information possible sur ces contrats parmi les données effectivement complétées.

**Bases véhicules**

Le produit Garages étant un produit de la branche Automobile, une autre source majeure dans la construction de cette base est la création d'une base véhicules propre à cette étude. L'objectif est de pouvoir associer à chaque contrat le nombre et le type des véhicules appartenant à l'entreprise. En effet, ce sont des véhicules pris en charge par l'assurance et concernés par toutes les garanties Automobile du contrat. Pour rappel, les véhicules confiés par la clientèle pour réparation ne sont ici pas désignés spécifiquement au contrat même s'ils sont également assurés par l'assurance Automobile de l'entreprise le temps de leur prise en charge.
Il existe une base véhicules qui recense tous les véhicules désignés aux contrats d'AXA IARD Entreprises. De cette base, les véhicules renseignés ont ainsi pu être récupérés et associés à chaque contrat. De nombreuses informations sont également à disposition pour chaque véhicule mais seules les variables suivantes ont été retenues :

— leur date d'entrée et sortie au contrat afin d'obtenir leur taux de présence sur chaque année de cette étude, et d'avoir une historisation sur le nombre de véhicules présent par année

— des informations relatives à la valeur de ces véhicules comme leur valeur assurée et/ou leur valeur à neuf afin de prendre en compte les montants assurés

Par ailleurs, à la souscription le détail du nombre de véhicules par type d'utilisation est demandé, ainsi lorsque cela est correctement renseigné, il est possible d'associer à

chaque véhicule sa fonction à savoir, s'il s'agit d'un véhicule de courtoisie, d'un véhicule du dirigeant, d'un véhicule utilisé dans le cadre de l'activité (par exemple un véhicule de dépannage) ou encore d'un véhicule de démonstration pour des activités de vente. Cependant, ces informations ne sont pas mises à jour chaque année ni lors de la déclaration d'entrée ou sortie de véhicules assurés au contrat. Ce type d'information étant non négligeable sur le niveau de risque : par exemple, un véhicule de courtoisie sera conduit par des personnes toujours différentes sans avoir l'information du type de conducteur, alors que les véhicules du dirigeant seront exclusivement conduits par le dirigeant. Or des informations plus détaillées sont disponibles à son sujet du fait qu'il est généralement le souscripteur du contrat.
Par conséquent, cette information détaillée a été retenue car elle est déclarée à la souscription, ce qui permet alors de s'en servir comme clé de répartition pour chaque catégorie de véhicule au fil des années du contrat. En effet, la déclaration d'un nouveau véhicule ou de la suppression d'un véhicule au contrat étant obligatoire, la donnée du nombre de véhicule total associé au contrat est disponible. Ainsi en utilisant la répartition du type de véhicule déclarée à la souscription, il est possible de l'appliquer à chaque assiette du nombre de véhicules, pour chaque année de présence du contrat afin d'estimer le nombre de véhicules de chaque catégorie sur cette période étudiée. Il est important de noter que ces données restent des estimations et non des valeurs observées, mais elles pourront améliorer la précision par rapport aux bases véhicules sources.

### Bases sinistres

La tarification d'une garantie se base sur la modélisation de la charge sinistre observée sur cette garantie. Ainsi, une fois toutes les données risques relatives au contrat récupérées, il faut ajouter les données de sinistralité par contrat et par année d'observation sur la période de l'étude. Trois années de vision sont considérées ici avec, pour chaque année de vision, deux années d'historique en amont. Cependant, dans le cadre de la modélisation des antécédents de sinistralité d'un contrat et étant donné qu'il s'agit d'une tarification à l'Affaire Nouvelle, trois années d'historique seront considérées pour chaque vision. Par conséquent, la sinistralité des contrats concernés sera retenue sur six ans : de 2012 à 2018.
La création d'une base de sinistralité associée à la modélisation de l'étude s'opère en plusieurs étapes. Les données sont disponibles par numéro de sinistre et chacun des sinistres est rattaché au numéro de contrat, ce qui permettra ensuite la fusion des bases de sinistralité et des bases risques. Par ailleurs, pour chaque sinistre, plusieurs informations sont retenues :

- le montant du sinistre, qui sera utile pour la modélisation de la charge ; la charge retenue sera la charge vieillie de 12 mois pour chaque sinistre pour éviter au maximum les montants forfaitaires appliqués au moment de la déclaration,
- la date de survenance du sinistre,
- la garantie impactée par ce sinistre,

— la part de responsabilité de l'assuré pour les sinistres de type RC Automobile, étant donné qu'ils impactent une tierce personne. Il y aura alors une intervention différente dans les recours, ce qui permettra d'expliquer la charge finale du sinistre.

— le fait que le sinistre entre ou non dans le cadre de la convention IRSA (cf. 2.3.2) est également indiqué.

Ces données ont ensuite été formatées pour obtenir une base de sinistralité en concaténant l'ensemble des sinistres d'un contrat par année de survenance. La base ainsi obtenue contient une ligne par contrat et par année de survenance, à savoir ici les années de vision considérées. Pour chaque ligne de la base, se trouvent : le nombre de sinistres, la charge globale correspondante, ainsi qu'un découpage de cette charge et de ce nombre de sinistres par garantie. De plus, afin d'intégrer la notion de responsabilité dans la modélisation, une distinction entre sinistres responsables (part de responsabilité supérieure à 50%) et non responsables (part de responsabilité inférieure à 50% a été effectuée en charge et en nombre de sinistres. Par ailleurs, pour chaque sinistre, la charge aura été découpée selon le montant du sinistre, à savoir :

— la charge hors graves, qui comprend tous les sinistres dont le montant n'excède pas 30 000€ : cette valeur correspond au seuil à partir duquel un sinistre est considéré comme "grave" sur cette branche, cette valeur a été définie par de précédentes études qui ont permis de déterminer le seuil à partir duquel le sinistre est considéré comme une valeur extrême

— la charge hors atypiques, qui comprend tous les sinistres dont le montant n'excède pas 2 millions d'€uros : ce sont des évènements rares qui ne représentent pas la sinistralité réelle d'un contrat, les sinistres dits "atypiques" ne seront donc pas pris en compte dans cette modélisation

— la charge écrêtée à 30 000€ et celle à 2M€, qui correspond à la prise en compte des sinistres respectivement graves et atypiques mais seulement jusqu'à la hauteur du seuil : par exemple un sinistre de 40 000€ sera valorisé à hauteur de 30 000€ dans la charge écrêtée de ce seuil, au lieu de ne pas du tout le considérer comme dans le cas de la charge hors graves. Cependant cela peut créer un biais dans la modélisation du fait de la présence d'un montant plafonné, mais cela permet pour autant d'être plus proche de la charge globale observée.

### Agrégation des bases

La structure de fusion de ces bases peut se synthétiser par le schéma suivant, jusqu'à atteindre la base agrégée :

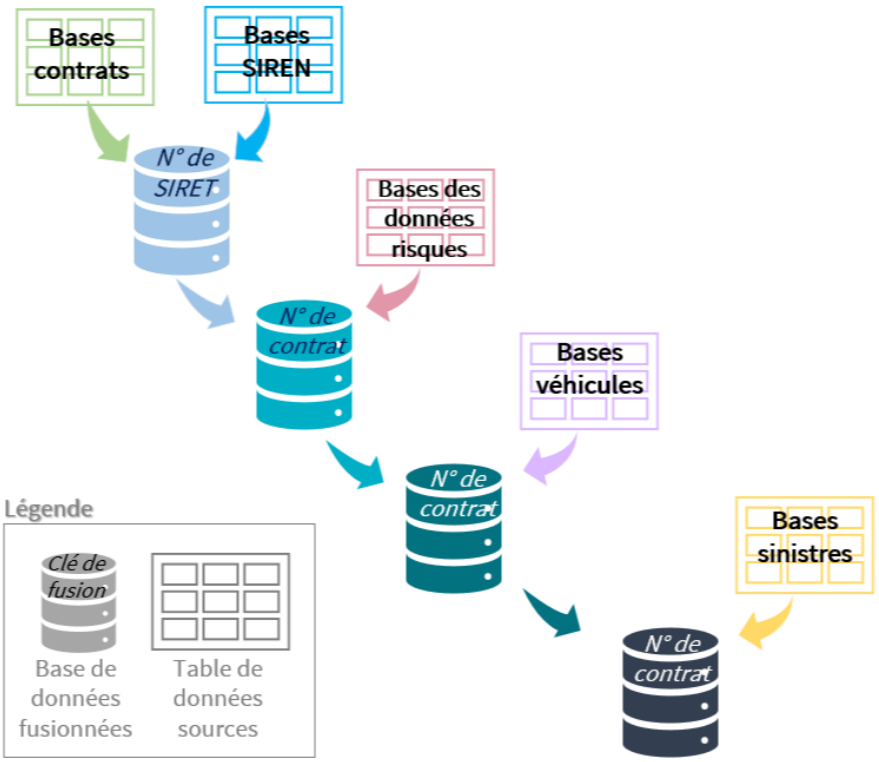



FIGURE 2.1 – Schéma d'agrégation des bases sources

### 2.2.2 Variables retenues

#### Complétion avec de nouvelles variables

Malgré les nombreuses sources disponibles sur les serveurs, certaines données risques ne sont pas remontées dans ces bases mais restent disponibles sur les outils de souscription. Ainsi après avoir fait un récapitulatif des données présentes sur l'outil de souscription le plus complet : AXAPAC, une récupération de ces données supplémentaires a été effectuée. Pour cela, un travail de recensement de ces données et de leur emplacement exact sur l'outil a été nécessaire, la récupération n'est possible qu'en indiquant de manière précise les élements suivants : le numéro de la procédure à entrer dans l'outil pour accéder à la page qui permet de récupérer la donnée, et les coordonnées de la position de l'information sur la page de la procédure correspondante. En effet, une macro fonction Excel permet de capturer à partir de ces informations les données sur cet outil. Ainsi les valeurs de 36 nouvelles variables pour l'ensemble des contrats souscrits sous AXAPAC ont pu être capturées.

**Format de la base de données**

Avant la modélisation, une mise en forme de la base a été nécessaire. En effet, une fois toutes les données regroupées dans une même base de données, les données sont présentées de la façon suivante : chaque variable est déclinée en plusieurs colonnes correspondant à la version de l'année de vision considérée, et ce pour toutes les variables dont l'évolution est jugée probable. Un retraitement du format de cette base a été réalisé afin de faciliter la lecture et la modélisation. Ainsi, la base a été retraitée pour obtenir la forme suivante : une nouvelle variable donnant l'année de vision considérée a été créée. De cette façon, chaque variable est représentée par une colonne et sa valeur correspond à celle de l'année de vision indiquée par la variable éponyme. La base de donnée comprend ainsi au maximum trois lignes par contrat, chacune correspondant à la présence ou non du contrat sur les années de vision étudiées, à savoir de 2016 à 2018 comme précisé plus tôt. De plus, le taux de présence annuel associé à chaque année est fourni afin d'associer l'exposition du contrat correspondante à la sinistralité observée sur cette période. Concernant les données de sinistralité, une historisation a été conservée sur trois ans pour les nombres et charges de sinistres de la garantie. Par conséquent, cette historisation est représentée par trois colonnes : l'une correspondant aux données dont la survenance est celle de l'année de vision, et les deux autres correspondant aux deux années de survenance antérieures à l'année de vision considérée. Ce formatage de la base peut être représenté avec le schéma présenté ci-après :

| | Données risques du contrat | | | charge sinistre observée | | | | |
|---|---|---|---|---|---|---|---|---|
| numéro de contrat | à vision 2018 | à vision 2017 | à vision 2016 | survenue en 2018 | survenue en 2017 | survenue en 2016 | survenue en 2015 | survenue en 2014 |
| contrat n°1 | | | | | | | | |
| contrat n°2 | | | | | | | | |

FIGURE 2.2 – Avant - Schéma de la base des données agrégées

| numéro de contrat | année de vision N | Données risques pour le contrat et l'année correspondants | | | | | charge sinistre observée de l'année N | de l'année N-1 | de l'année N-2 |
|---|---|---|---|---|---|---|---|---|---|
| contrat n°1 | 2016 | | | | | | | | |
| | 2017 | | | | | | | | |
| | 2018 | | | | | | | | |
| contrat n°2 | 2017 | | | | | | | | |
| | 2018 | | | | | | | | |

FIGURE 2.3 – Après - Schéma de la base formatée pour la tarification

**Liste des variables retenues**

- Activité principale : activite1_
- Activité secondaire, le cas échéant : activite2_
- Effectif de l'entreprise : effectif_
- Chiffre d'Affaires annuel de l'entreprise : caht_
- Nombre de véhicules propriétaires, avec la distinction par type de véhicule : nbveh_
  - Véhicules de courtoisie
  - Véhicules du dirigeant
  - véhicules de démonstration
  - Véhicules d'exploitation
  - Véhicules en plaque W (plaque provisoire)
- Mode de gestion du contrat : gestion_
  - Forfaitaires
  - Révisables au CA ou à l'effectif
- Forme juridique de l'entreprise : juridiq_
  - Nom propre
  - Société
- Qualité de l'occupant : qualite_
  - Locataire (LO)
  - Locataire Exonéré de RC locative (LE)
  - Locataire qui souscrit pour le Compte de son propriétaire (LC) : il est alors chargé d'assurer les bâtiments également
  - Locataire d'Alsace-Moselle (LA) : la réglementation locale peut impacter la définition de la responsabilité incombée au locataire
  - Propriétaire Occupant (PO)
  - Copropriétaire (CO)
- Nombre de sites de l'entreprise : nbsitessuppl_
- Situation du risque : isolemt_
  - Agglomération (1)
  - ZI ou ZA (2)
  - Centre commercial (3)
  - Zone rurale (4)
- Antécédents de sinistralité de l'entreprise si elle n'est pas en création : ANTE_
- Code postal du risque : codepostal_rsq

### 2.2.3 Complétion des données manquantes

Certaines variables peuvent avoir un fort pouvoir discriminant dans la modélisation mais les bases de données ont des valeurs non renseignées ce qui peut biaiser la tarification. Il est alors important de traiter les données manquantes dans une base de données pour avoir une modélisation fiable. Plusieurs techniques sont possibles pour le traitement de ces valeurs.

#### Remplacement par la modalité la plus représentée de la variable

Une méthode qui peut être utilisée pour la gestion des valeurs manquantes est le remplacement de ces valeurs inconnues par la modalité la plus représentée de la variable. Cela n'est justifié uniquement lorsque la part des valeurs manquantes ne représente pas une proportion trop importante de la modalité de remplacement, c'est à dire la modalité ayant le plus d'observations. Ces valeurs manquantes ne doivent également pas représenter une part trop élevée dans l'ensemble de la base. En effet, dans le cas contraire cela introduirait trop d'incohérences dans la modélisation.
Ainsi, les remplacements suivants ont été effectués dans les proportions indiquées dans le tableau ci-dessous :

| Variable concernée | Part de N/A sur la variable concernée | Modalité de remplacement *(part sur la variable concernée)* | Part de NA dans la modalité de remplacement | Nouvelle part de la modalité sur la variable |
|---|---|---|---|---|
| Activité principale | *3%* | Garages M3T5 *(42%)* | 6% | 45% |
| Qualité de l'occupant | *12%* | LO *(41%)* | 22% | 53% |
| Isolement | *12%* | 1 - Agglomération *(45%)* | 21% | 57% |
| Forme juridique | *15%* | S – Société *(63%)* | 19% | 78% |
| Nombre de sites suppl. | *14%* | Oui – *(9%)* | 60% | 23% |

Hypothèse : NA issus du haut de segment

FIGURE 2.4 – Part des valeurs inconnues dans la variable de remplacement

#### Régression linéaire selon une autre variable

Une autre méthode est également applicable dans le traitement des valeurs manquantes : regarder s'il existe une relation entre des variables de la base de données afin de retrouver les valeurs de l'une à partir des valeurs de l'autre.
L'une des variables où la part de valeurs inconnues était très importante était l'effectif, or cette variable est une variable majeure dans le tarif actuel et semble très discriminante dans les premiers résultats de modélisation qui étaient pratiqués en conservant les valeurs manquantes telles qu'elles. Une étude des corrélations a permis de montrer que cette variable avait une corrélation importante avec la variable de Chiffre d'Affaires.
Dans un premier temps, une régression linéaire a été modélisée entre ces deux variables lorsque toutes deux étaient renseignées.

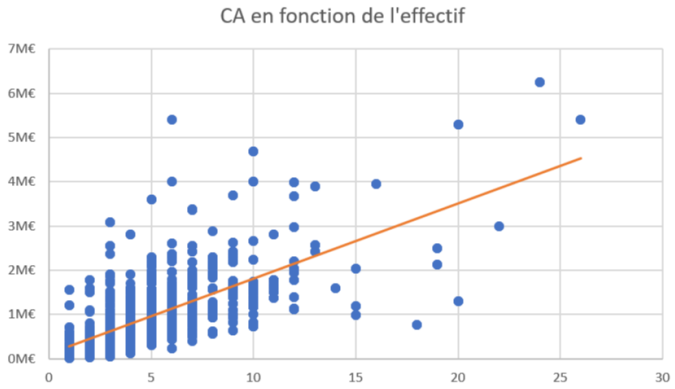



FIGURE 2.5 – Relation entre le CA et l'effectif déclarés à la souscription

Les résultats confirment une relation linéaire entre l'effectif et le CA. Afin d'affiner et de s'ajuster au mieux aux données, une étude par activité principale a été menée. En effet, ces variables sont également corrélées avec l'activité exercée par l'entreprise. Quelques regroupements de modalités de l'activité principale ont été réalisés pour une meilleure stabilité des résultats : certaines activités ayant trop peu d'observations pour avoir une conclusion fiable. Les résultats de quelques unes de ces régressions sont présentés ci-après : il s'agit ici des activités les plus représentées en termes de nombre de contrats, dans le cadre de cette étude.

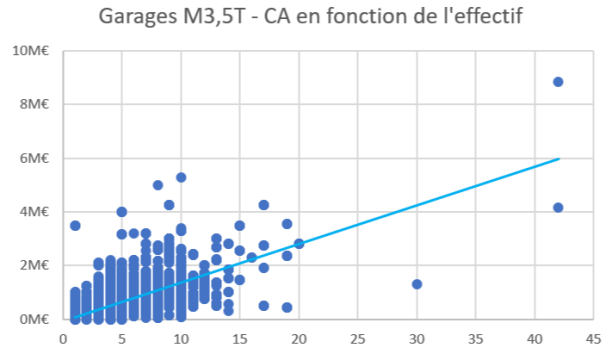



FIGURE 2.6 – Relation entre le CA et l'effectif pour l'activité Garages pour les véhicules de moins de 3,5T

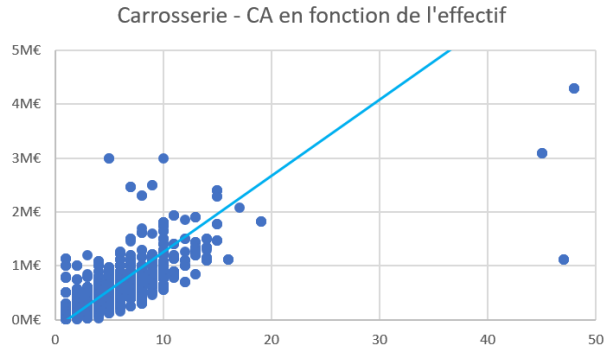



FIGURE 2.7 – Relation entre le CA et l'effectif pour l'activité Carrosseries

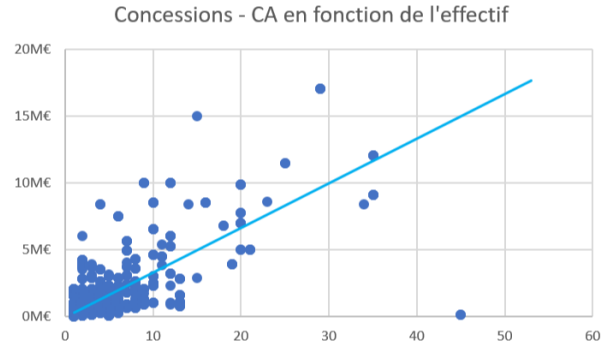



FIGURE 2.8 – Relation entre le CA et l'effectif pour l'activité Concessions

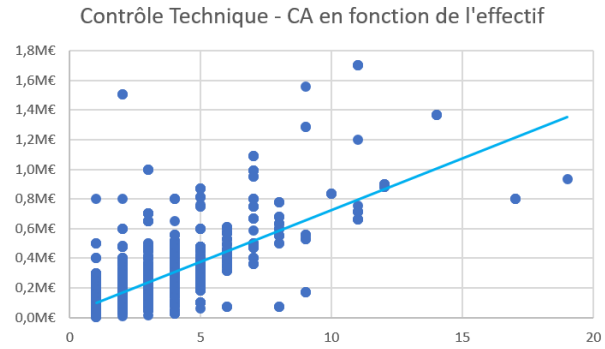



FIGURE 2.9 – Relation entre le CA et l'effectif pour l'activité Contrôle technique

Les résultats sont ici meilleurs que ceux sans la distinction par activité présentés plus hauts. De plus, cela valide la relation de linéarité existante. Ainsi, dans la cadre du remplacement des valeurs manquantes, les variables de CA et d'effectif ont été remplacées par la relation linéaire correspondante spécifique à chaque activité exercée. Cela a été effectué dès lors qu'au moins l'une des deux variables était renseignée, ce qui a permis de diminuer de 66% la part des valeurs manquantes.

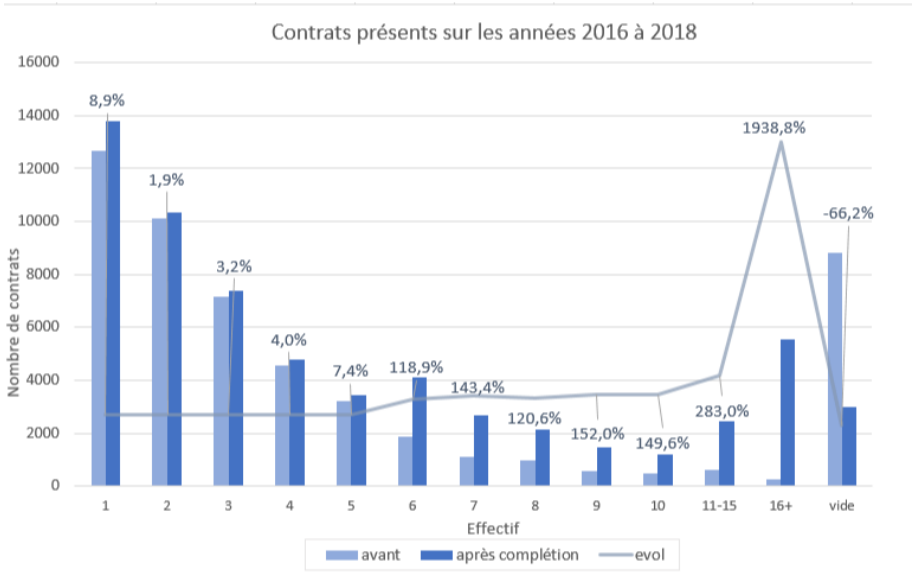



FIGURE 2.10 – Répartition des remplacements effectués en termes d'effectifs grâce à la régression

## 2.3 Retraitement des modalités

### 2.3.1 Le cas spécifique de la variable d'activité

Parmi les variables les plus discriminantes de ce produit, l'activité a été étudiée plus particulièrement. Afin d'avoir une vision complète, la variable d'activité principale et celle d'activité secondaire le cas échéant ont été prises en compte.
Au vu du nombre important d'activités disponibles dans les bases créées, il est apparu utile de procéder à des regroupements pour diminuer le nombre de modalités avec une très faible part de contrats concernés. Les catégories retenues ont été les suivantes :

- Garagistes pour les véhicules de moins de 3,5T (Garagistes M3T5)
- Garagistes pour les véhicules de plus de 3,5T (Garagistes M3T5)
- Garagistes pour les deux-roues (Garagistes 2R)
- Concessions
- Contrôles Techniques
- Réparation rapide : pour regrouper tous les contrats qui étaient apparentés comme des activités de garagiste mais dont le risque est plus spécifique. En effet, l'étude de la sinistralité sur cette sous-segmentation a révélé une rentabilité dégradée par rapport aux autres activités de réparation. Cela pourrait s'expliquer par le fait que les réparations proposées ne sont généralement pas les mêmes et les procédures de vérification seraient moins approfondies du fait des contraintes de temps qu'impliquent ce type d'activité.

— et une activité qui regroupe toutes les autres activités, comme celles qui sont définies comme des holdings notamment (Autres)

Cependant la distinction entre les activités de réparation et de concession étant très discriminante en termes de risques, des études supplémentaires pour valider les activités affiliées à chaque contrat ont été menées.

**Réconciliation entre l'activité renseignée et l'activité des bases INSEE liée au SIRET**

Afin de vérifier que les informations déclarées à la souscription et les informations recensées par l'INSEE correspondaient, une comparaison de ces données pour chaque contrat a été effectuée. Les contrats dont l'activité était différente ont été analysés de plus près et notamment lorsque l'un indiquait une activité de réparation (ex. garagiste) et l'autre une activité de vente (ex. concession). Cela représente 300 contrats pour 2M€ de primes, ce qui correspond à une part non négligeable du portefeuille. Afin de prendre en compte cette incohérence et pour éviter de potentiellement sous-estimer les activités de réparation et sur-estimer les activités de vente en incluant ces contrats dans l'une ou l'autre de ces activités, une activité hybride a été créée pour prendre en compte cette spécificité, qu'on appellera "Garages-Concessions".

**Fiabilisation avec l'activité secondaire**

L'information de l'activité secondaire et du CA associé le cas échéant étant demandée à la souscription cette donnée a pu être récupérée dans la base de tarification. En plus d'inclure cette activité secondaire dans les variables explicatives du modèles, cette variable a permis d'améliorer la fiabilité de la variable d'activité principale.
La première étape a été de vérifier que les contrats dont l'activité principale était le contrôle technique n'avait pas d'activité secondaire puisque cette profession ne peut pas exercer d'autre activité comme la vente de véhicules ou la réparation, et cela est bien vérifié sur nos données.
Ensuite l'activité secondaire a été croisée avec sa part en CA : cela a permis de mettre en avant les contrats ayant une activité secondaire à 50%, c'est à dire qui compte pour autant que l'activité principale. Dans ce cas, lorsque l'une est une activité de réparation et l'autre une activité de vente, ces contrats ont été placés dans la variable hybride "Garages-Concessions" créée plus tôt à partir de l'information INSEE. Ce type de risque représente 70 contrats mais permet tout de même de ne pas biaiser les autres activités.
De plus, ont également été ajoutés dans cette modalité certains contrats issus de l'outil XF. En effet, cet outil comprend principalement les contrats haut de segment, puisqu'il s'agit majoritairement de Concessions. De plus, les contrats avec une activité principale de réparation mais ayant une activité secondaire de vente, peu importe la part de cette activité dans le CA, ont été placés dans l'activité hybride.

#### Autre cas concerné par l'activité hybride

Par ailleurs, une autre variable a été prise en compte dans le cadre de cette analyse : la variable de CA associé à la vente. Cette variable n'est renseignée que pour les contrats dont cette spécification est demandée par les constructeurs ou autres marques automobiles dont dépendent les garages, ce qui représente 1 000 contrats. Parmi ceux-ci, les contrats dont l'activité principale est "Garagistes" et dont le rapport du CA de vente sur le CA total renseigné est supérieur à 50% ont été basculés vers l'activité hybride. Cette règle a été prise en compte uniquement pour les assurés dont le CA total était supérieur à 1M€ afin de ne pas pénaliser les contrats de bas de segment pour qui cette répartition n'est pas révélatrice d'une activité se rapprochant des Concessions. En effet, parmi les Concessions ayant un CA supérieur à 1M€, 90% ont une part de CA vente supérieure à 50% du CA global, cela est donc bien représentatif de ce type d'activité. Ainsi concernant les activités Garages entrant dans cette règle de CA, 75% des activités de Garages sont devenues des activités hybrides.
Cette activité "Garages-Concessions" représente alors 2 140 contrats, soit 4% du portefeuille.

### 2.3.2 L'importance des antécédents de sinistralité

#### Récupération de la sinistralité la plus récente

Au moment de la souscription du contrat, la sinistralité passée est déclarée à l'assureur si l'entreprise n'est pas en création à ce moment-là. Cette information est fournie grâce au Relevé d'information [1]. Ce document regroupe notamment l'ensemble des sinistres du client en détaillant la nature du sinistres, s'il s'agit de sinistres subis ou causés et la responsabilité associée à ceux-ci. Ces données sont stockées dans les bases risques et permettent donc de prendre en compte ces informations dans le cadre de cette modélisation.
Cependant, cette donnée correspond à la situation de l'assuré au moment où il a souscrit le produit, et elle a pu évolué depuis que le contrat est en portefeuille. Ainsi, l'objectif est de mettre à jour ces données lorsque cela est possible afin de se placer dans une situation d'affaire nouvelle, c'est-à-dire une situation où toutes les données renseignées seront des données actualisées à la date de la souscription.
Dans ce but, les données de sinistralité déclarées à la souscription avec le relevé d'information ont été remplacées par la sinistralité observée de l'assuré dans nos bases sinistres lorsque l'ancienneté du contrat est supérieure à un an. En effet, lorsque le contrat a une duration inférieure à un an, il est plus prudent de se baser sur la sinistralité déclarée à la souscription pour avoir des données plus fiables. Ainsi, lorsque le contrat est présent depuis plus d'un an, la sinistralité est prise en compte au prorata de sa présence sur trois ans, qui correspond à la période utilisée dans cette étude pour la prise en compte des antécédents.

1. Document qui retrace les antécédents de sinistres et de situation de l'assuré sur une durée maximale de cinq ans, il est fourni par le précédent assureur au moment de la souscription auprès d'un nouvel assureur

### Calcul du ratio de sinistralité

**Ratio actuel** Actuellement, la sinistralité passée est prise en compte à travers le nombre de sinistres par effectif :

$$\text{Antécédents à l'effectif} = \frac{\text{nombre de sinistres Auto et RC pro sur la période}}{\text{effectif} \times \frac{\text{période de référence sinistre en mois}}{12}}$$

Avec une période de référence comprise entre un et trois ans.
En effet, le nombre de sinistres automobile est corrélé positivement à l'effectif de l'entreprise : plus il y a d'employés, plus le nombre d'accidents est susceptible d'être élevé. Cependant, dans le cadre de la refonte de ce tarif, la possibilité de modifier la variable qui sert de dénominateur au calcul de la sinistralité a été étudiée. Plusieurs variables ont été candidates à l'élaboration de ce nouveau ratio en plus de l'effectif : le chiffre d'affaires et le nombre de véhicules. Ces variables semblaient effectivement tout aussi pertinentes que l'effectif pour re-baser le nombre de sinistres par période de temps. Afin de déterminer la variable la plus pertinente dans le cadre de cette tarif, trois modélisations différentes ont été réalisées, chacune avec un dénominateur du ratio d'antécédents différents toutes choses égales par ailleurs : l'un avec l'effectif, un autre avec le CA et le dernier avec le nombre de véhicules. Le ratio d'antécédents qui s'est révélé le plus discriminant est celui à l'effectif, ce résultat a donc été challengé mais a finalement été conservé.

**Sinistres pris en compte** Par ailleurs, dans le tarif actuel, l'ensemble des sinistres déclarés sont pris en compte, quelle que soit leur nature : par exemple les sinistres de type Dommages Automobiles sont pris en compte au même titre que les sinistres RC Automobile dans le calcul des antécédents de sinistralité des sinistres RC Automobile. Il a également été décidé de remettre en cause cette méthode de la même façon que pour le choix du dénominateur : à savoir une modélisation pour chaque méthode et l'antécédent qui apparaît le plus discriminant dans la modélisation sera conservé. Cependant, la modélisation a montré que les antécédents de la garantie étudiée étaient plus discriminants que ceux qui prennent en compte toutes les garanties, mais pas manière suffisamment significative pour écarter l'une ou l'autre des possibilités. De plus, la refonte du tarif étant globale, et la prise en compte des antécédents étant présente pour toutes les garanties du tarif, une approche ne se basant que sur une garantie n'est pas suffisante pour prendre de décisions lorsque les résultats ne permettent pas d'avoir une conclusion suffisamment claire. En effet, si les antécédents de la garantie RC Automobile semblent légèrement plus discriminants que ceux qui englobent toutes les garanties, ce ne sera pas obligatoirement le cas pour la modélisation des autres garanties. Ainsi, il faut garder une certaine homogénéité dans la prise en compte de ce type de variable pour éviter la redondance des informations et l'incohérence dans la modélisation. Cette question reste alors en suspens jusqu'à la modélisation des autres garanties, et pour la suite de l'étude, seront considérés les antécédents de la garantie RC Automobile étant donné qu'ils restent plus discriminants.

**Notion de responsabilité des sinistres** Une autre information pourra être retenue dans la modélisation des antécédents : la notion de responsabilité dans les sinistres avec tiers, c'est-à-dire les sinistres de la garantie RC Automobile. Les sinistres non responsables seraient alors comptabilisés de manière moins pénalisante que les sinistres responsables lors du calcul des antécédents. Cette information étant disponible sur le relevé d'information, l'intégration de cette information n'impliquerait pas une demande supplémentaire lors de la souscription. De plus, elle permettrait une prise en compte plus juste de la sinistralité passée. Par ailleurs, dans le cadre de cette étude, les informations de part de responsabilité dans la sinistralité observée sont également disponibles dans les bases utilisées. Au vu de l'apport en termes d'équité pour les assurés, et de sa simplicité d'intégration, cette notion sera retenue dans la suite pour les modélisations. De manière concrète, les antécédents seront calculés ainsi :

— un sinistre RC Automobile responsable aura une valeur de 1

— un sinistre RC Automobile non responsable ne sera retenu que pour une valeur de 0,5

Il faut alors définir la notion de sinistres responsables et non responsables, pour cela une introduction à la convention IRSA qui en donne une définition permettra de mieux comprendre cette notion.

**La convention IRSA** Lors d'un accident de la circulation, la garantie RC Auto est impliquée en tant que garantie obligatoire (en France) et peut engendrer des recours entre assureurs selon la responsabilité de leur assuré dans l'accident. Il s'agit toujours à l'assureur d'indemniser son propre assuré, et charge à lui de faire un recours auprès de l'assureur de la personne responsable le cas échéant pour récupérer le montant de l'indemnisation correspondante selon la part de responsabilité du tiers. Afin de faciliter cette gestion de recours entre assureurs, une convention a été créée, il s'agit de la convention IRSA. La Convention IRSA est la Convention d'Indemnisation directe de l'assuré et de Recours entre Sociétés d'Assurance automobile. Signée par la plupart des sociétés d'assurance en France, elle est destinée à faciliter l'indemnisation des dommages matériels sous certaines conditions.

Concrètement, si le montant des dommages estimé par un expert en assurance automobile est inférieur au plafond de 6500 € (fixé par la convention), le recours est forfaitaire et s'établit à 1568 € (valeur en 2020). Le recours exercé est proportionnel au niveau de responsabilité de l'auteur des dommages. Si le montant des dommages est supérieur à 6500 €, le recours est réel, c'est à dire correspondant au montant réel des dommages.

Cette notion de recours est donc non négligeable dans la modélisation d'une telle garantie car cela peut amener à des sinistres enregistrés comme négatifs une fois que le recours est effectif.

Par ailleurs, et c'est l'objet de ce paragraphe, cette convention définit la part de responsabilité des assurés dans un accident, cette responsabilité ainsi définie est appelée

responsabilité conventionnelle. Elle peut différer de la responsabilité de droit commun dans certains cas spécifiques, par exemple dans des situations non prévues par la convention comme les accidents causés par une surcharge du véhicule, une usure des pneus ou par des infractions comme les excès de vitesse ou d'alcoolémie. La part de responsabilité a un impact sur la distinction entre le nombre de sinistres responsables et non responsables. En effet, cette distinction est réalisée en considérant comme responsables les sinistres dont la part de responsabilité est supérieur à 50% et non responsables sinon. Par ailleurs, lorsque la responsabilité conventionnelle n'a pas pu être déterminée, la responsabilité de droit commun est alors appliquée. Ces deux types de responsabilité sont stockées dans les bases sinistres et peuvent donc être utilisées comme information supplémentaire. Tous les cas sont alors prévus pour déterminer la responsabilité et l'appliquer aux antécédents.

**Application aux antécédents de sinistralité** L'approche de la prise en compte de la responsabilité avec la variable d'antécédents est la suivante. Pour rappel, la variable d'antécédents de sinistralité retenue dans cette étude a le format suivant :

$$\text{Antécédents à l'effectif} = \frac{\text{nombre de sinistres RC Auto sur la période}}{\text{effectif} \times 3}$$

Dans le cas où la période retenue est fixée à trois ans.
L'approche des antécédents consiste alors à re-traiter le nombre de sinistres pour intégrer la notion de responsabilité. Ainsi, il a été défini la règle suivante :

— un sinistre responsable (au sens défini plus haut, à savoir une part de responsabilité supérieure à 50%) aura une valeur de 1

— un sinistre non responsable (donc une part de responsabilité strictement inférieure à 50%) ne sera retenu que pour une valeur de 0,5

Cela permet de diminuer le ratio de sinistralité des assurés ayant des sinistres non responsables et ainsi cela diminuerait la pénalité accordée par le modèle avec la variable d'antécédents.

### 2.3.3 Regroupement de modalités

Les remontées opérationnelles sur le produit actuel mettent en évidence le fait que le tarif ne serait pas assez lissé sur des variables comme l'effectif ou encore le CA : pour l'ajout d'une personne dans l'effectif, le tarif augmenterait de façon trop prononcée. Afin de pallier à cette critique, l'élaboration de tranches a été retenue pour éviter des effets de seuils trop fréquents, notamment pour le bas de segment ayant de faibles effectifs, mais qui représentent une part non négligeable du portefeuille. Ces tranches ont été définies après des premières modélisations qui ont permis de regrouper les modalités grâce aux coefficients attribués : les valeurs ayant le même coefficient ou un coefficient très proche ont alors été regroupées pour créer des tranches pertinentes pour cette étude. Ainsi cette méthode a été appliquée aux variables suivantes, dont les regroupements seront présentés dans la partie suivante (2.5.1) :

— Effectif
— Chiffre d'Affaires
— Nombre de véhicules
— Antécédents de sinistralité : chaque type d'antécédents retenu a fait l'objet de tranches spécifiques selon les résultats des modélisations
— Nombre de sites supplémentaires
— Qualité de l'occupant
— L'activité secondaire (activité_2) qui suivait le découpage de la variable d'activité principale a également été retraitée pour prendre en compte l'importance des contrats n'ayant pas d'activité secondaire déclarée à la souscription

## 2.4 Retraitements de la charge sinistre : variable à expliquer

Il existe plusieurs méthodes pour considérer la charge sinistre, à savoir la variable à expliquer, du fait des éléments spécifiques inhérents à la modélisation de la Responsabilité Civile. Parmi ces éléments, les sinistres dits "graves", c'est-à-dire dépassant un certain seuil, et les sinistres liés aux recours, que nous détaillerons par la suite. Afin que l'analyse et les conclusions sur la modélisation attritionnelle ne soient pas impactées par la particularité de ces derniers, nous avons effectivement dans un premier temps exclu ces sinistres de la modélisation.

### 2.4.1 Sinistres graves

Étudions dans un premier temps les sinistres extrêmes, il existe plusieurs définitions pour les catégoriser. En ce qui concerne la branche Auto, des seuils sont prédéfinis afin d'avoir une harmonisation sur les communications au sein de l'entité. Ainsi, deux niveaux de sinistres extrêmes peuvent se distinguer :

— les sinistres "graves" : il s'agit des sinistres dont la charge dépasse 30K€. Ces sinistres sont suivis spécifiquement afin de noter s'il n'y a pas une dérive de sinistralité sur ce type de sinistres, et le cas échéant, essayer d'expliquer cette augmentation en procédant à une analyse plus poussée.
— les sinistres "atypiques" : il s'agit de sinistre dont la charge dépasse 2 millions d'€uros. Ces sinistres sont toujours étudiés en profondeur car très peu nombreux par définition mais ayant un lourd impact sur le portefeuille. Dans le cadre de l'étude de la garantie RC Automobile, les sinistres seront plus souvent concernés par des sinistres corporels ayant de lourdes conséquences sur le tiers assuré que pour d'autres garanties du produit.

#### Charge hors graves

Une première approche pour éviter de biaiser la modélisation avec des sinistres graves est de les exclure de la modélisation pour ne prédire que les sinistres attritionnels. Un

taux moyen observé sur les dernières années serait alors appliqué pour prendre en compte cette charge supplémentaire, sans avoir à la modéliser.

### Charge écrêtée

Une autre option possible dans la prise en compte des sinistres extrêmes est l'écrêtement des sinistres supérieurs, jusqu'à un certain seuil, ici de 30K€. Cela correspond à la prise en compte de la part du sinistre jusqu'au seuil préalablement défini. Ainsi, un sinistre d'une valeur totale de 40 mille €uros aura une valeur écrêtée de 30 mille €uros, qui est le seuil défini pour les sinistres graves, et les 10 mille €uros supplémentaires sont appelés la sur-crête.
Cette méthode permet de ne pas exclure les extrêmes en les comptabilisant dans la modélisation, mais en minimisant la possibilité qu'ils soient considérés comme points aberrants, ce qui minimiserait alors leur importance : cela reviendrait alors à faire une modélisation attritionnelle.

### Mutualisation de la charge des graves

La modélisation écrêtée permet de prendre en compte la fréquence réelle des sinistres en n'excluant pas les sinistres extrêmes, mais elle ne permet cependant pas de modéliser toute la charge réellement observée. Or la prime pure qui est modélisée dans cette étude doit couvrir toute la sinistralité à venir des contrats en portefeuille. Cependant, une typologie de contrat ayant été impactée par un sinistre grave voire atypique sur la période étudiée aura alors une prime pure bien plus élevée qu'une autre typologie de contrat qui n'aura pas été impactée sur la période retenue pour la modélisation, mais les évènements graves peuvent impacter tout type de contrat en portefeuille.
Ainsi, et afin de respecter le principe de mutualisation propre au fonctionnement de l'assurance, une autre méthode utilisée est la mutualisation du montant de la sur-crête des sinistres extrêmes sur l'ensemble du portefeuille. Concrètement, cela signifie que la valeur correspondant à l'ensemble des sur-crêtes de chaque évènement grave ou atypique est redistribuée sur tous les contrats au prorata de leur sinistralité réellement observée afin de conserver un principe d'équité. En effet, les contrats n'ayant aucun sinistre sur la période étudiée ne seront pas impactés par cette redistribution : leur part dans la sinistralité globale observée étant de 0%. En revanche, les contrats ayant une sinistralité non nulle auront une charge sinistre majorée de $x\%$, $x$ étant la part que représente la charge réelle observée mais écrêtée du contrat sur la charge réelle écrêtée du portefeuille, pour une année donnée. Le schéma suivant pourra permettre de mieux comprendre ce principe :

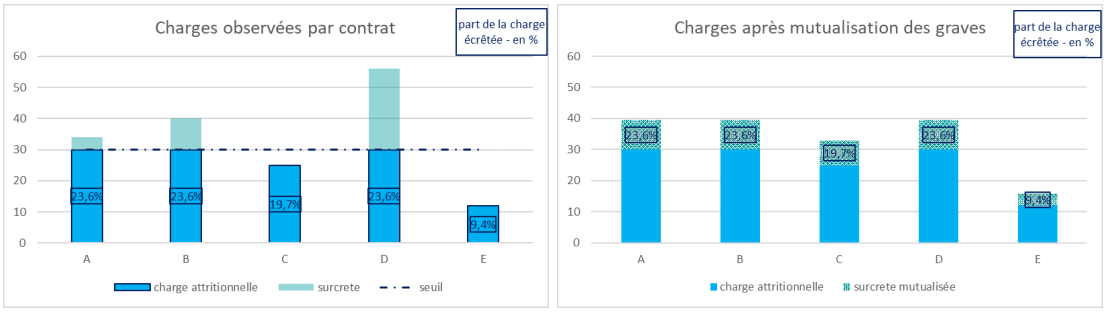



FIGURE 2.11 – Schéma de mutualisation des sinistres graves

#### Charge hors atypiques

La mutualisation est un principe fondamental de l'assurance, cette méthode qui est ici ré-appliquée sur les données comme pour les sinistres "graves" dans le paragraphe précédent. Néanmoins, dans cette étude, les sinistres atypiques sont très peu nombreux (au nombre de deux sur l'année modélisée 2018) mais dont la charge représente plus de 70% de la charge des sinistres graves (c'est-dire supérieurs à 30 mille €uros). De ce fait, la mutualisation de ces sinistres crée de très fortes majorations sur les contrats sinistrés. En effet, une forte proportion du portefeuille ayant une charge nulle, la mutualisation s'établit sur une moindre proportion du portefeuille. Le principe d'équité et de solidarité, sous-jacent à la la mutualisation peut alors ici prendre des proportions qui peuvent biaiser la modélisation.
Par conséquent, afin d'avoir une modélisation au plus près de la sinistralité réellement survenue sur chaque contrat, les sinistres atypiques ont alors été exclus de l'analyse. Ils seront ainsi mutualisés sur l'ensemble du portefeuille à la fin de l'analyse. Cela permet de ne pas altérer la modélisation tout en prenant en compte l'ensemble de la charge observée dans le tarif finalement payé.

### 2.4.2 Gestion des recours pour la garantie RC Auto

#### Modélisation de la responsabilité

La part de responsabilité peut avoir des conséquences non négligeables sur la charge des sinistres liés à la garantie RC Automobile. Il faut alors prendre en compte cette notion dans la modélisation mais il ne s'agit d'une variable explicative standard. En effet, ce n'est pas une variable représentative du risque assuré en lui même mais une information supplémentaire sur la charge modélisée. Dans ce contexte, plusieurs approches ont été testées pour ajouter cette notion, en plus de la prise en compte de cette information dans les antécédents :

— modélisation distincte des sinistres responsables et des sinistres non responsables à travers des modèles en approche prime pure, il faut alors distinguer la charge des responsables de celle des non responsables.

— modélisation distincte des sinistres responsables et des sinistres non responsables à travers des modèles en approche Fréquence × Coût Moyen agrégés, ici la variable à modéliser se distingue ainsi :
  — Deux variables de comptage : le nombre de sinistres responsables et le nombre de sinistres non responsables ;
  — Deux autres variables de sévérité, qui attribuent le coût-moyen des sinistres respectivement responsables et non responsables, en prenant la charge annuelle associée lorsqu'elle est strictement positive, et divisée par le nombre de sinistres sur l'année considérée, toujours en distinguant par type de responsabilité

**Modélisation des recours**

Les recours sont présents sous plusieurs formes dans la modélisation de la charge :

— les sinistres dont la charge est forfaitaire s'il s'agit d'un évènement qui entre dans le cadre de la convention IRSA présentée plus tôt ;

— les sinistres ayant un montant négatif, c'est-à-dire pour lesquels le montant remboursé par AXA à son assuré tiers est inférieur au montant du recours touché in fine.

**Charge forfaitaire** Les charges sinistres forfaitaires créent des irrégularités dans la répartition de la charge totale du fait de pics de volumétrie présents sur ces montants-là. Par conséquent, l'adéquation aux lois de modélisations peut être dégradée par la sur-représentation de ces montants dans les charges sinistres. C'est pour cela qu'il peut s'avérer intéressant de modéliser ce type de charge de façon distincte du reste de la charge globale.

**Charge négative** Les sinistres ayant un montant négatif peuvent permettre de diminuer la charge globale par assuré au moment de l'agrégation de celle-ci. En effet, la modélisation se base sur la somme des charges pour chaque contrat, et sur chaque année de vision retenue pour l'étude. Le fait de diminuer la charge associée au contrat permet de prendre en compte une charge plus faible, ce qui intègre l'information de la non responsabilité du sinistres, et cela pénalise alors moins l'assuré.
Cependant, lorsque la charge agrégée à la maille du contrat est négative car elle n'est pas compensée par d'autres sinistres ayant une charge strictement positive suffisante, alors cette charge doit être exclue de la modélisation. En effet, les lois utilisées pour modéliser la charge globale ou le coût moyen, à savoir respectivement la loi de Tweedie ou la loi Gamma par exemple, ne peuvent pas modéliser des montants négatifs.
Ainsi, dans les premières étapes de modélisation les charges dont les valeurs étaient négatives ont été fixées à 0, mais cela exclut alors une part non négligeable de la charge à modéliser. Par conséquent, cette charge sera modélisée de façon distincte. L'ensemble des sinistres ayant une charge négative sera considéré, et non pas seulement la charge agrégée par contrat et par année qui est négative car non compensée par d'autres sinistres.

**Modélisation par nature de charge sinistre** Cette modélisation qui prend en considération les sinistres forfaitaires et négatifs séparément peut se faire à travers un modèle de Fréquence × Coût Moyen, en ajoutant les sinistres qui n'entrent pas dans l'une ou l'autre de ces catégories de sinistre. Cela aboutirait alors de manière analogue à la modélisation des sinistres responsables d'un côté et non responsables de l'autre comme expliqué plus haut.

**Modélisation en distinguant les sinistres issus de la convention IRSA** Sur le même principe, il est possible de modéliser la charge en reprenant la distinction entre sinistres responsables et non responsables, à laquelle on applique la distinction entre les sinistres issus de la convention IRSA ou non (cf 2.3.2). Cependant, au vu des volumes du portefeuille, la distinction sur les sinistres non responsables ne permettra pas d'avoir une fiabilité des résultats dans le cas où ces derniers seront distingués par application de la convention ou non, mais dans le cas de volumes suffisants, cela serait une piste de modélisation à ne pas négliger pour gagner en finesse.

**Modélisation par mutualisation des forfaitaires et négatifs** Ces modélisations de recours peuvent grandement complexifier le modèle et s'éloignent alors du principe de simplicité qui est le fil conducteur de la refonte de ce produit. Une autre manière de prendre en compte les recours peut être étudiée : la mutualisation des sinistres liés aux recours.
Pour cela, l'ensemble de la charge des sinistres forfaitaires est comptabilisé, et est exclue de la charge globale, mais pour les sinistres négatifs, seule la charge négative restante après agrégation à la maille du contrat sera retenue. En effet, cette agrégation permet de capter une partie de l'information apportée par la présence de sinistres négatifs en réduisant d'autant la charge associée au contrat.
Ainsi, la charge globale est retraitée des charges forfaitaires qui peuvent dégrader l'ajustement aux lois de modélisation, et retraitée de la charge négative qui ne peut être modélisée ainsi. Par suite, et de la même façon que pour la mutualisation des sinistres dits "graves", la somme de ces deux types de charge peut alors être mutualisée sur le portefeuille. Il est possible de procéder de la même façon en reportant cette somme sur chaque observation, proportionnellement à l'importance de la charge retraitée de l'observation concernée par rapport à la charge retraitée totale.
Cependant, si la mutualisation s'effectue au prorata de la charge retraitée des forfaitaires et des négatifs alors les contrats ayant uniquement des sinistres forfaitaires ou négatifs ne seront pas concernés par la mutualisation : leur charge étant passée à 0 une fois retraitée. L'assiette pour le calcul de la part à mutualiser ne sera alors pas la charge retraitée mais les primes acquises du contrat, et ce uniquement pour les contrats dont le nombre de sinistres, quel qu'en soit le type, est strictement positif.
Pour reprendre le schéma de mutualisation des graves, la mutualisation des recours retenue peut se représenter comme ci-dessous.
Si le portefeuille initial est réparti ainsi

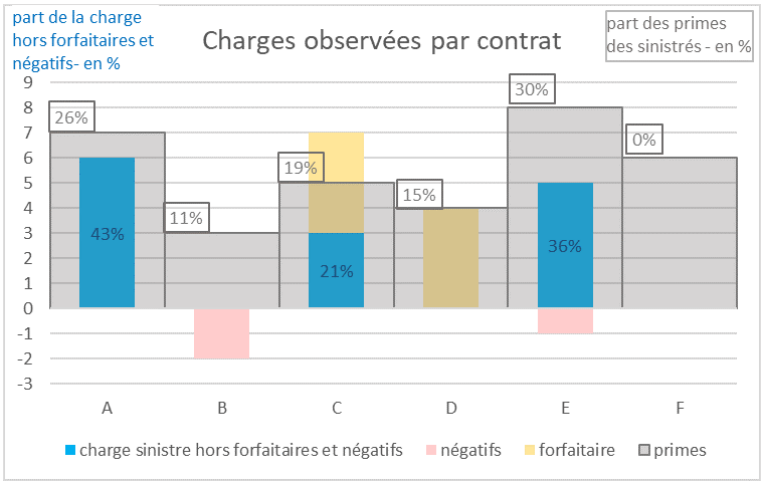



FIGURE 2.12 – Schéma de mutualisation des sinistres graves

Alors les deux types de mutualisation possibles sont les suivantes :

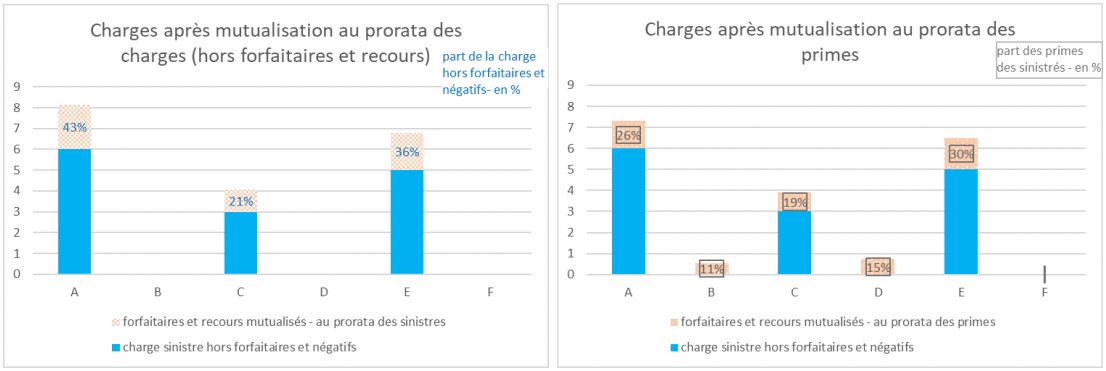



FIGURE 2.13 – Schéma de mutualisation des sinistres graves

La méthode retenue sera celle présentée à droite.

## 2.5 Statistiques descriptives

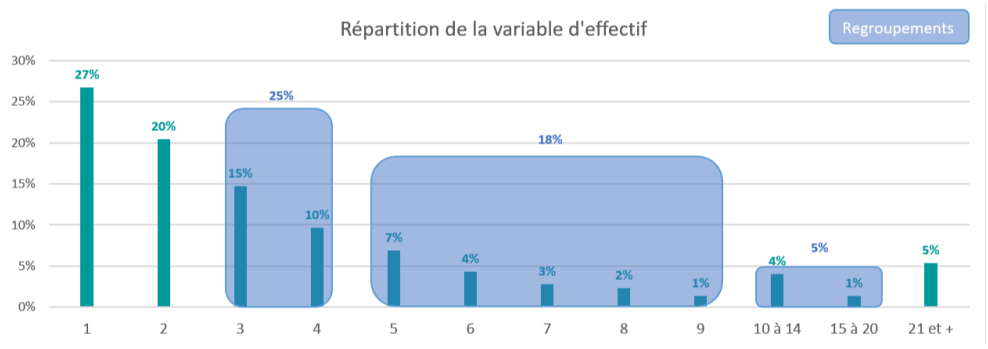
Avant toute étape de modélisation, il est nécessaire d'étudier en détail les variables retenues en tant que variables explicatives afin de mieux connaître les données utilisées et ainsi ne pas faire de conclusion erronée lors de l'analyse des résultats de la modélisation. Une analyse des statistiques de quelques une des principales variables du modèle sera présentée, mais l'ensemble des variables utilisées comme variables explicatives du modèles auront également été analysées en procédant de la même façon.

### 2.5.1 Tris à plat

Une première étape dans la description statistique des variables est l'étude du tri à plat des variables. Cela signifie que la répartition des modalités de la variable est analysée en termes de nombre d'observations sur la base étudiée. Cela permet de mieux comprendre la typologie majoritaire des contrats sur les données retenues après retraitements.

#### Effectif

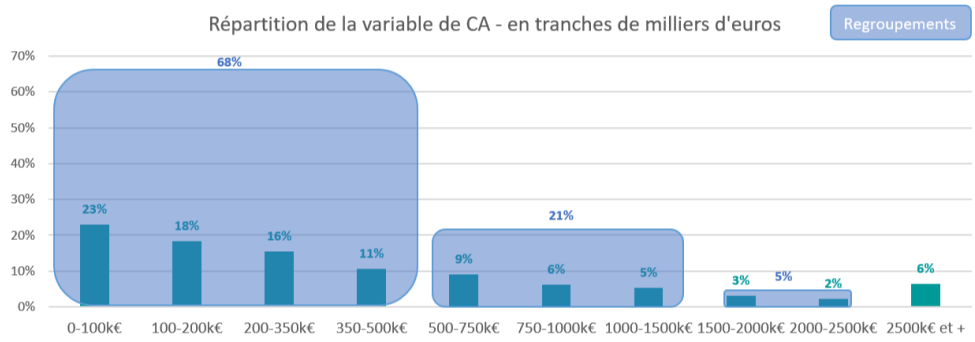
L'effectif est une variable ayant une importance non négligeable dans le tarif actuel puisqu'elle fait partie des variables permettant de fixer la cotisation de base du tarif, tandis que les autres variables ont un format de coefficient d'ajustement.




Il est important de remarquer qu'il s'agit principalement de contrats avec des effectifs peu élevés :72% des observations concernent des entreprises de moins de cinq salariés et pour 27% d'entre elles l'effectif se réduit à une personne. La plupart des contrats étudiés pourraient faire partie de la catégorie des TPE (Très Petites Entreprises) si le critère de l'effectif était retenu. En effet, les entreprises ayant un nombre de salariés inférieur à dix personnes entrent dans cette catégorie. Un autre critère utilisé pour définir ce type d'entreprise est le Chiffre d'Affaires annuel. Étudions alors cette variable.

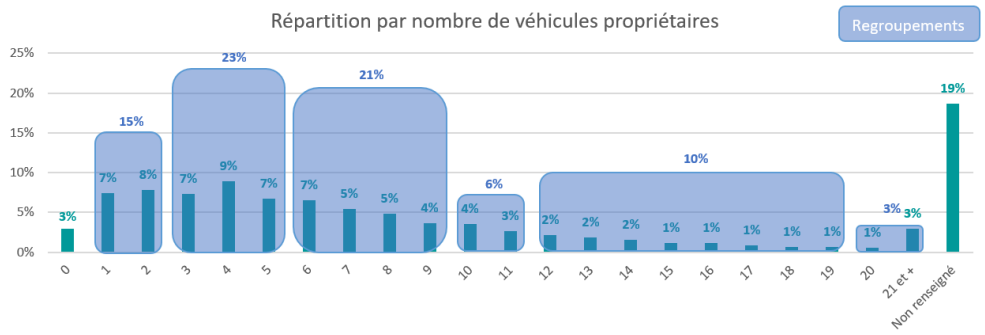
#### Chiffre d'Affaires

Les chiffres indiqués sur le graphique ci-dessous sont en milliers d'euros.




A nouveau, cette tendance majoritaire du portefeuille à être concernée par la catégorie des TPE s'observe également ici : le CA annuel doit être inférieur à 2M€ pour être considérée comme telles!!! . Ce type d'entreprise est plutôt tourné vers l'économie locale, en tant que service de proximité. Les TPE représentent donc les deux tiers du paysage des entreprises françaises [1], ce qui permet de mieux comprendre et appréhender cette répartition : ils représentent ici 90% du portefeuille.
Par ailleurs, la plupart de ces contrats pourraient être concernés par la distribution du produit "Tout-en-un" [2]. Pour rappel, ce produit comprend à la fois les garanties Auto et Non Auto, dans un package dédié qui permet ainsi de simplifier l'assurance sur ce type d'activité. En effet, le public visé cherche généralement la simplicité des démarches plutôt que la multiplication du nombre de contrats pour une seule entreprise (par ex. contrats Auto, Risques Industriels, RC professionnelle).

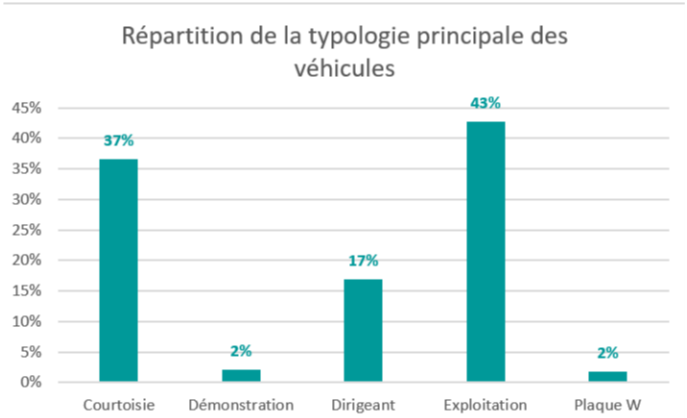
### Nombre de véhicules propriétaires




De manière analogue, le portefeuille est principalement composé d'entreprises ayant peu de véhicules propriétaires, c'est-à-dire de véhicules lui appartenant : les véhicules confiés ne sont par exemple pas considérés ici car les propriétaires de ces véhicules sont les clients, le détail de la typologie de ces véhicules est présentée ci-après avec la variable du type de véhicule principal.

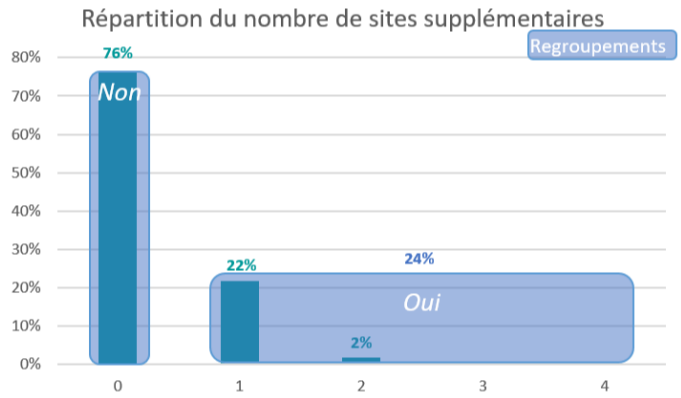
1. Donnée de 2018, issue de la Banque de France
2. La limite de CA annuel pour avoir accès à ce produit est de 2,3M€

### Type de véhicule principal




Cette variable permet de visualiser la typologie majoritaire des véhicules propriétaires au contrat. Ces différentes catégories sont issues des données à la souscription : la répartition des véhicules selon ces différentes catégories est déclarée au moment de la souscription uniquement. Il n'y a donc pas de mise à jour de cette variable le long de la vie du contrat.

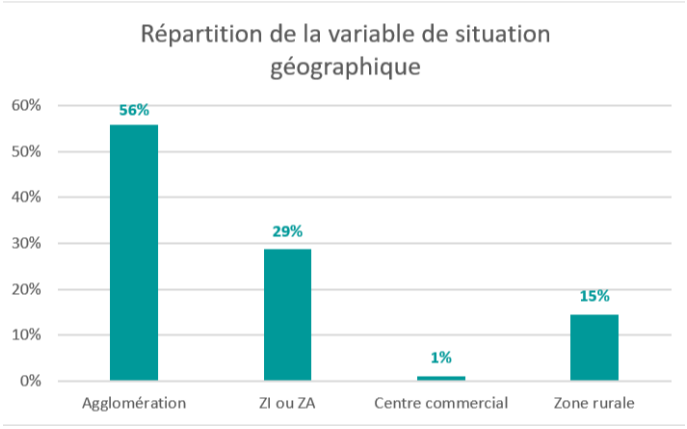
### Sites supplémentaires




La notion du nombre de sites appartenant une même entreprise peut avoir un impact non négligeable sur des garanties comme la RC Automobile du fait de la multiplication potentielle de la circulation des véhicules entre les différents sites, ce qui peut alors multiplier les risques d'accidents de la circulation. Les données présentes dans les bases indiquent le nombre de sites supplémentaires détenus par l'entreprise. Cependant au vu des faibles volumes et des niveaux de risques observés dans les différentes modélisations effectuées lors de cette étude, une transformation de cette donnée en variable binaire a été retenue.

### Situation du risque

La situation du risque peut également avoir un impact important sur le niveau de risque. En effet, si l'entreprise se situe en agglomération, les problématiques ne seront les mêmes que si elle se trouve en zone rurale : la circulation étant plus dense en agglomération, le risque d'avoir un accident de la route peut être plus important notamment en cas de collision avec un piéton.




La prépondérance de risques se situant en agglomération ou Zone Industrielle s'observe ici, or il s'agit de zones ayant potentiellement plus de fréquentation routière et donc plus de probabilité d'accidents. Cette donnée est ainsi à ne pas négliger et pourra être prise en compte de manière brute ou retraitée notamment avec l'information de la zone géographique du risque qui peuvent être corrélées.

### Antécédents de sinistralité

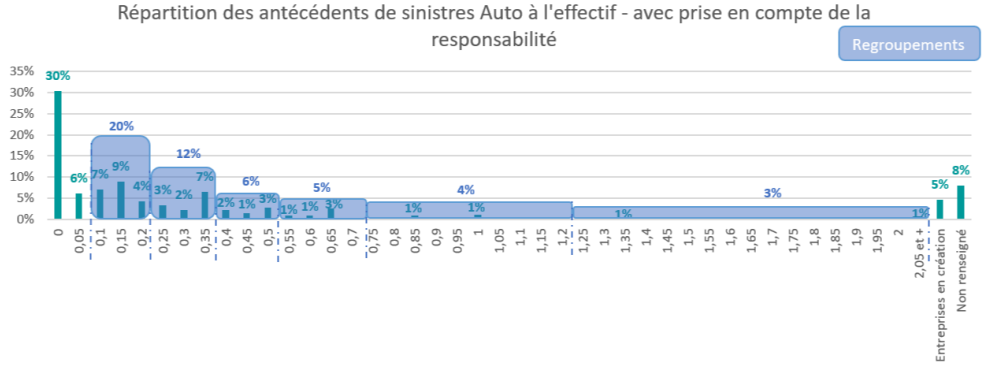



FIGURE 2.14 – Antécédents de sinistralité

Les antécédents choisis pour représenter au mieux la sinistralité passée sont ceux qui prennent en compte les sinistres des garanties Automobile, pondérés par l'effectif. Des arrondis et regroupements ont été appliqués pour faciliter l'usage de cette variable et la compréhension du tarif

### Gestion du contrat

La variable de gestion du contrat peut permettre de mieux comprendre le produit en indiquant la proportion de contrats qui sont révisables chaque année. Le choix de ce type de gestion est laissé à l'assuré pour les contrats ayant un CA annuel supérieur à 1M€, en-deçà de ce montant, les contrats sont automatiquement forfaitaires.

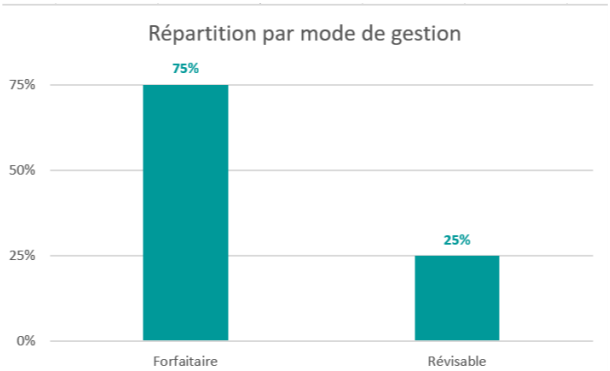



FIGURE 2.15 – Mode de gestion

Une majorité de contrats est sous une gestion forfaitaire. Étudions alors de plus près le fonctionnement de ce type de gestion.

Les contrats révisables sont des contrats dont la prime est réactualisée chaque année en fonction de l'évolution du CA annuel et/ou de l'effectif, qu'ils soient à la hausse ou à la baisse afin d'avoir un tarif qui s'ajuste régulièrement à la réalité observée. Cela permet au client d'avoir une prime qui s'adapte à sa situation. Du côté de l'assureur, cette typologie permet de suivre l'évolution de ces entreprises, en se basant sur le fait que l'objectif d'une entreprise étant de se développer, les primes augmenteront dans le même mouvement que le CA et/ou l'effectif s'accroissent. Cependant, les principaux inconvénients sont la possibilité également de diminution des primes dans le cas où les entreprises sont en perte de vitesse, mais cela implique dans tous les cas une gestion plus lourde afin de mettre à jour les données servant d'assiette à la révision et le re-calcul des primes en fonction des évolutions constatées. C'est pour cela qu'une étude des bénéfices apportés par ce type de contrat a été menée. En effet, l'anti-sélection pourrait jouer un rôle si les entreprises qui optaient pour ce type de gestion étaient des entreprises ayant observé une tendance à la baisse de leur activité de manière générale ou sur certaines années, ce qui les conduiraient à prévoir une baisse de leur prime le cas échéant. A l'inverse, si ce type de contrat montre une tendance majoritairement à la hausse, alors une étude sur la possibilité d'étendre la souscription des contrats révisables à des entreprises ayant un CA supérieur à 1,5M€ au lieu de 1M€ ne serait pas souhaitable car cela constituerait en l'état la perte d'une assiette de primes qui aurait été à la hausse avec une révision du contrat.

### Activité principale

Une autre variable ayant une importance notable dans le tarif actuel est celle de l'activité principale :

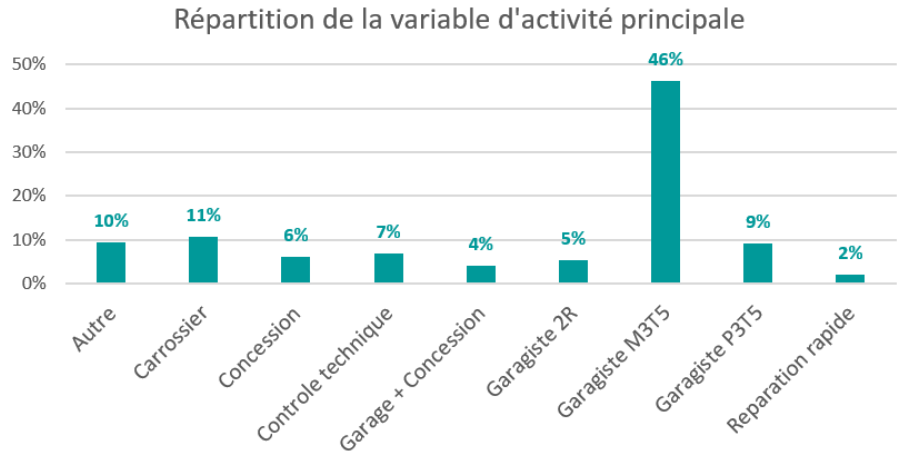



FIGURE 2.16 – Activité principale

La tendance observée avec le CA et l'effectif se confirme avec la prépondérance de contrats ayant une activité de Garagistes pour les véhicules de moins de 3,5T. En effet, ce type d'activité n'a pas les mêmes problématiques qu'une Concession, ayant généralement une structure plus importante et plus orientée vers la vente que l'après-vente.

### Activité secondaire

L'activité secondaire est non négligeable en termes de risque puisqu'il peut fortement varier d'une activité à l'autre. Il est alors important de prendre en compte l'activité secondaire notamment lorsque celle-ci est associée à une typologie de risque différente. Le premier découpage suivait celui de l'activité principale pour garder une cohérence dans la modélisation.

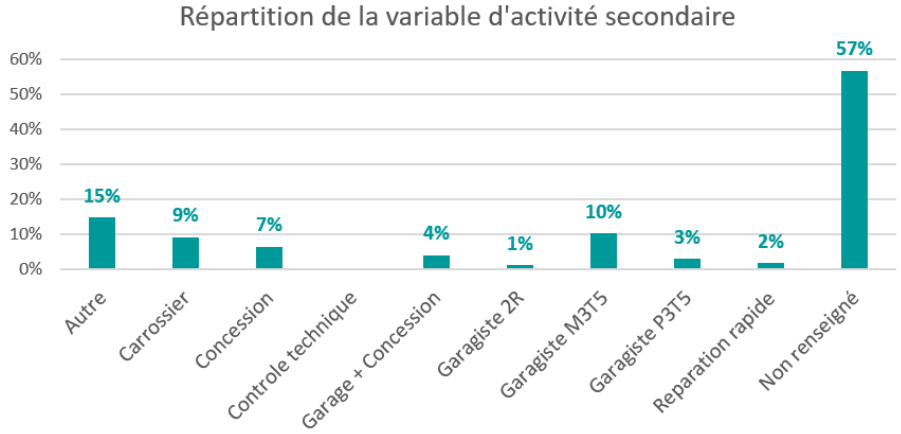


FIGURE 2.17 – Activité secondaire

Cependant, au vu de la répartition, cette transformation ne semble pas la plus appropriée : il faudrait distinguer parmi les "Non renseignés", les contrats dont l'information de l'activité secondaire n'a pas été complétée, malgré le fait qu'ils pratiquent une seconde activité ; et ceux ne pratiquant pas d'autre activité. Par conséquent une modalité "Aucune" a été créée pour correspondre à ce deuxième cas.
Par ailleurs, plusieurs activités ont été regroupées selon la nature du risque :

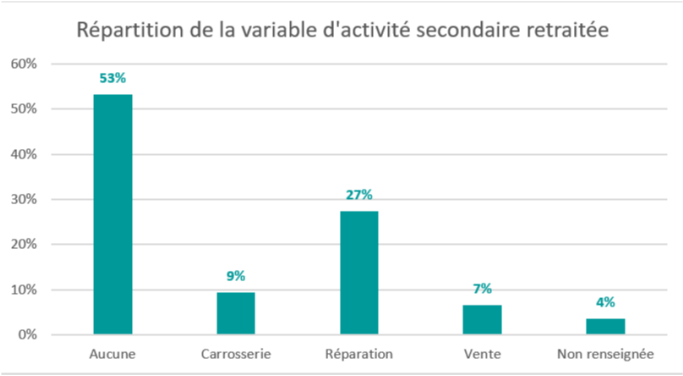



FIGURE 2.18 – Activité secondaire avec regroupements

**Forme juridique**

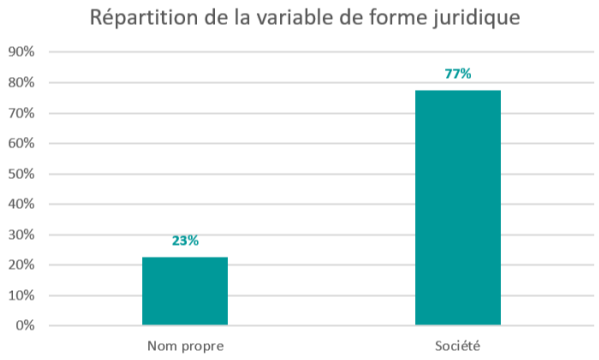



FIGURE 2.19 – Forme juridique

Pour compléter les informations sur l'entreprise, la notion de la forme juridique de celle-ci peut être révélatrice de comportements et aussi avoir un impact sur la sinistralité.

### Qualité de l'occupant

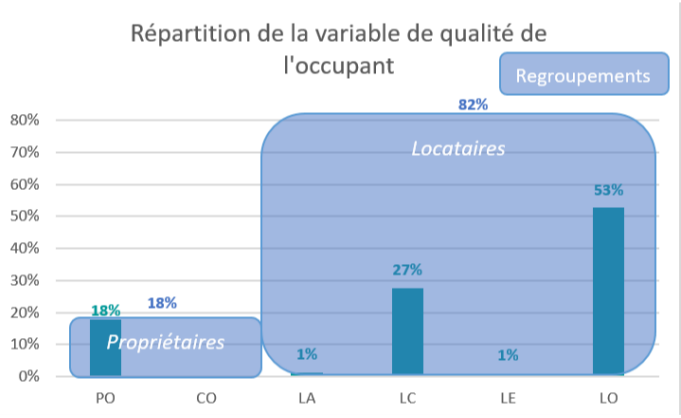



FIGURE 2.20 – Qualité de l'occupant

Sur le même principe, le fait que l'occupant soit un locataire ou un propriétaire peut être corrélé à la sinistralité.

## 2.5.2 Corrélations

Avant de procéder à la modélisation il est également important de connaître les corrélations (*RAKOTOMALALA* (2017)) qui peuvent exister entre les variables explicatives. Ces corrélations, entre variables quantitatives puis qualitatives vont donc être analysées.

### Variables quantitatives

Plusieurs mesures existent pour la calcul des corrélations entre variables quantitatives :

- le coefficient de Pearson qui mesure une dépendance linéaire entre les variables explicatives
- le $\rho$ de Spearman est basé sur le coefficient de Pearson mais sous une forme non paramétrique
- le $\tau$ de Kendall qui s'interprète comme une probabilité de correspondance de deux séries de données

Le coefficient de Pearson est la normalisation de la covariance des variables par leur écart-type :

$$r_{xy} = \frac{COV(X,Y)}{\sigma_x.\sigma_y}$$

avec $COV(X,Y) = \mathbb{E}[(X - \mathbb{E}[X]).(Y - \mathbb{E}[Y])]$ et $\sigma_x = \sqrt{\mathbb{E}[(X - \mathbb{E}[X])^2]}$.
Ce coefficient est compris entre $-1$ et 1 et son interprétation est analogue à celle de la covariance :

— si $r_{xy} > 0$ alors la relation entre les deux variables est linéraire et positive : elles varient dans le même sens et linéairement

— si $r_{xy} < 0$ alors la relation devient linéaire et négative

— si $r_{xy} = 0$ alors cela ne signifie pas toujours que les variables sont indépendantes car l'indépendance implique que $r_{xy} = 0$ mais ce n'est pas une équivalence. Cependant l'équivalence est vérifiée quand le couple $(X, Y)$ suit une loi normale bi-variée, mais ce n'est pas cette cas ici, cette mesure ne sera donc pas utilisée pour l'étude des corrélations.

Le $\rho$ de Spearman s'apparente au coefficient de Pearson mais se base sur le rang des données [1] plutôt que leur valeur observée directement. Cela donne la formule suivante :

$$\rho_S = \frac{COV[rg(X), rg(Y)]}{\sigma_{rg(X)} \cdot \sigma_{rg(Y)}}$$

Avec rg(X) et rg(Y) la variable de rang de X et Y respectivement. Cette mesure permet de retrouver les propriétés de signe associées au coefficient de Pearson sur la corrélation positive ou négative entre les variables, mais en mettant en avant des relations linéaires comme non linéaires alors que le coefficient de Pearson ne concerne que les relations linéaires entre les variables. En outre, du fait du caractère non paramétrique, le $\rho$ de Spearman permet d'avoir une équivalence entre l'indépendance des variables et sa nullité, sans avoir nécessairement la normalité du couple $(X, Y)$.

Le $\tau$ de Kendall a en revanche une interprétation différente puisqu'elle repose sur la notion de probabilité. Ce type de corrélation se base également sur les variables de rang, au même titre que le $\rho$ de Pearson. Ainsi si $\tau > 0$ alors la probabilité de concordance des variables de rang est supérieure à celle de discordance : donc cela s'approche d'une corrélation positive ; inversement lorsque $\tau < 0$. De plus, lorsque $\tau = 0$, alors la probabilité de discordance est égale à celle de concordance, donc il n'y a pas de relation positive ou négative entre les deux variables de rang étudiées.

Pour cette étude, le $\rho$ de Pearson sera retenu, car il permet de concilier à la fois une interprétation adaptée de l'indépendance et du sens de corrélation à tout type de données, et pas seulement pour des données de loi normale bivariée, et un calcul peu coûteux en évitant le calcul de probabilité ligne à ligne.

Ce type de corrélation peut se représenter sous forme de matrices ou de corrélogrammes pour une visualisation graphique, ce qui permet alors de faciliter la prise de décision.

1. Les valeurs sont ordonnées par ordre croissant et son rang dans la série ordonnée est attribué pour chaque valeur

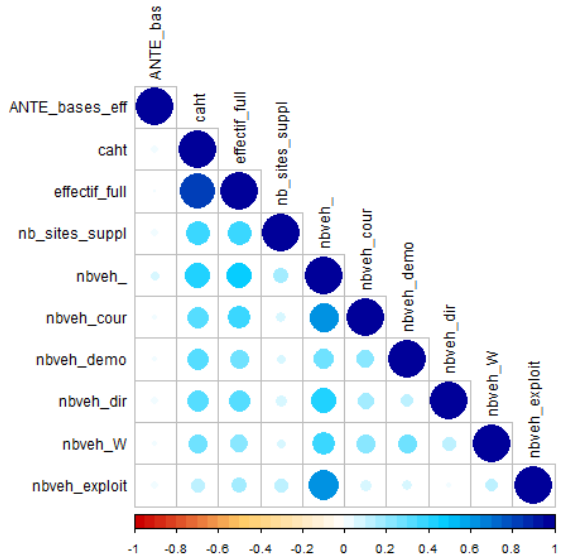



FIGURE 2.21 – Corrélations quantitatives

L'étude de ces corrélations montre notamment la relation plutôt intuitive qui existe entre le CA et l'effectif de l'entreprise : cette matrice permet ainsi de valider cette corrélation et de s'en servir par la suite pour améliorer les données. Par ailleurs, d'autres corrélations semblent se démarquer, à savoir entre le nombre de véhicule total (nbveh_) et le détail par type d'activité de ces véhicules ; ce qui pourrait permettre d'exclure certaines de ces variables de la modélisation, pour éviter la redondance de cette information. Néanmoins, cette information pourrait être remplacée par une nouvelle variable qui permettrait d'identifier le type de véhicule majoritaire au contrat. Une nouvelle variable qualitative : "type_veh_ppal" peut alors être introduite dans ce but.

### Variables qualitatives

Pour l'étude des corrélations entre les variables qualitatives, le $V$ de Cramer sera retenu comme mesure pour ce type de corrélations. Cette mesure se base sur une normalisation du test du $\chi^2$ d'indépendance appliqué à un tableau de contingence des variables qualitatives deux à deux.

Le test du $\chi^2$ d'indépendance permet de comparer les fréquences de deux variables qualitatives et ce en faisant deux hypothèses : $H_0$ : "La fréquence sur les deux variables est identique", donc il n'y a pas de relation entre elles, et $H_1$, l'hypothèse alternative : "La fréquence diffère selon les modalités", donc il y a une relation entre ces variables. La statistique de test associée est donnée par :

$$S = \sum_{i=1}^{p} \sum_{j=1}^{k} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

Avec

- $p$ le nombre de modalités de la variable $X$
- $k$ le nombre de modalité de la variable $Y$
- $o_{ij}$ la fréquence observée des contrats ayant la modalité $i$ de la variable $X$ et la modalité $j$ de la variable $Y$.
- $e_{ij}$ la fréquence théorique si les variables étaient indépendantes, qui est donnée par : $\frac{l_i.m_j}{N}$ avec
  - $l_i$ le nombre de contrats dans la modalité $i$ de la variable $X$,
  - $m_j$ le nombre de contrats dans la modalité $j$ de la variable $Y$
  - $N$ le nombre total d'observations de la base.

Pour un test de d'ordre $(1-\alpha)$ déterminé préalablement, si la valeur de $S$ est inférieure au seuil déterminé par le quantile d'ordre $(1-\alpha)$ de la loi du $\chi^2$ à $(p-1)(k-1)$ degrés de liberté alors l'hypothèse nulle n'est pas rejetée, c'est-à-dire que les variables sont considérées comme indépendantes, et si la valeur de $S$ est supérieure à ce seuil alors l'hypothèse nulle est rejetée, ce qui signifie qu'il y a une relation significative entre ces variables.

Une fois définie $S$ la statistique de test du $\chi^2$ d'indépendance, il est plus aisé de définir le $V$ de Cramer :

$$V = \sqrt{\frac{S}{N.\min(k,p)}}$$

Avec pour rappel, $N$ le nombre total d'observations de la base, et $k$ et $p$ respectivement le nombre de modalités des variables $X$ et $Y$. Le $V$ de Cramer peut prendre des valeurs entre 0 et 1 :

- si $V = 0$ alors il n'y pas de relation statistique entre les variables car cela correspond au cas où $o_{ij} = e_{ij}$
- si $V = 1$ alors il existe une relation entre les variables car cela signifie que la statistique de test $S$ est égale à $N.\min(k,p)$ or cette valeur correspond à la valeur maximale qui peut être prise par la statistique de test
- pour les valeurs comprises dans l'intervalle $]0,1[$, l'importance de la relation est croissante avec la valeur de $V$.

De la même façon, le V de Cramer peut se représenter sous la forme d'un corrélogramme :

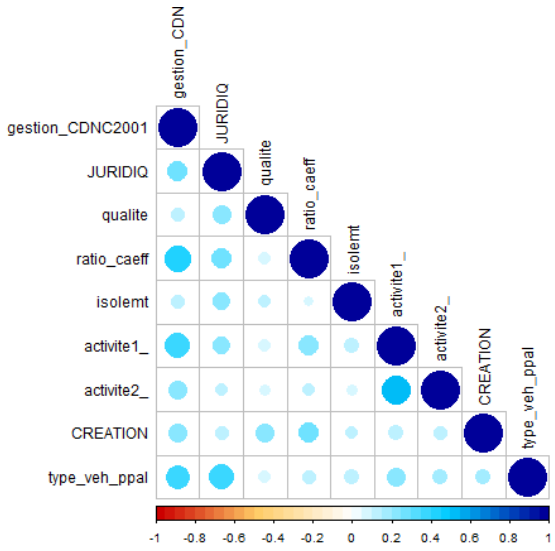



FIGURE 2.22 – Corrélations qualitatives

Ces coefficients révèlent que le ratio du CA sur l'effectif est relativement corrélé au mode de gestion, comme à l'activité principale, et au statu d'entreprise en création. Une analyse plus poussée sur ces variables pourra alors être effectuée pour ne pas conserver l'ensemble de ces variables dans le modèle, mais les corrélations ne permettent pas d'éliminer directement l'une d'elles.

# Chapitre 3

# Modélisation de la prime pure

## 3.1 Les modèles linéaires généralisés

### 3.1.1 Théorie

La notion de Modèles Linéaires Généralisées (GLM en anglais) qui sera utilisée par la suite pour la modélisation de la prime pure des garanties du produit étudié va maintenant être présentée. Comme leur nom l'indique, il s'agit d'une généralisation des modèles de régression linéaire usuels, notamment le modèle linéaire gaussien. Dans ce modèle, une variable aléatoire $Y$ est modélisée grâce à un ensemble de variables explicatives $X_{i=1...p}$. Dans la suite, le vecteur des variables explicatives sera noté $X$ et le vecteur des coefficients de chaque variable explicative $\beta$. Dans un modèle linéaire gaussien l'hypothèse suivante est faite :

$$Y \sim \mathcal{N}(\mu, \sigma^2)$$
$$avec\ \mu = X^t\beta$$

Le modèle s'écrit alors :

$$Y = X\beta + \varepsilon$$

où $\varepsilon$ est le bruit gaussien centré tel que

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$$

L'intérêt des modèles linéaires généralisés est d'étendre le modèle gaussien à un ensemble de lois plus large que la loi normale, à savoir : la famille exponentielle (cf B.1). Les hypothèses du modèle deviennent alors :

- $Y|X = x \sim \mathbb{P}_{\theta,\phi}$ appartient à une famille exponentielle
  avec $\theta$ le paramètre canonique et $\phi$ le paramètre de dispersion de la famille des lois exponentielles.
- $g(\mu(X)) = g(\mathbb{E}[Y|X]) = X\beta$ avec $g$ une fonction bijective appelée fonction de lien, qui traduit la relation existant entre $Y$ la composante aléatoire et $X$ la composante déterministe

Le recours à la méthode des GLM permet de modéliser dans un cadre plus large que celui du modèle linéaire simple. L'emploi d'une fonction de lien permet de se ramener à un cadre mathématique plus aisé, en faisant « comme si » $Y$ suivait une loi Normale alors que l'on suppose qu'il suit une Gamma, une Poisson ou tout autre loi de la famille exponentielle.

La fonction de lien canonique est la fonction de lien telle que : $g(\mu) = \theta$. Ce type de fonction de lien est généralement le plus utilisé dans la modélisation assurantielle, mais d'autres fonctions de liens peuvent être utilisées.

#### Estimation des paramètres

L'estimation des paramètres $\beta$ est obtenue en maximisant la log-vraisemblance du modèle, sachant que la variable à expliquer suit une loi de la famille exponentielle, de densité notée $f$. Les paramètres sont estimés par la méthode de la maximisation de la log-vraisemblance. Soit un ensemble de variables aléatoires $Y_{i=1,\dots,n}$, ici la vraisemblance de $\beta$ est donnée par :

$$l(\beta) = \sum_{i=1}^{n} \log(f(Y_i; \beta, \phi)$$

La même procédure peut s'appliquer pour le paramètre $\phi$.

#### Lois utilisées

Il est possible de modéliser les sinistres de plusieurs manières :

— En ne considérant que le coût des sinistres : il s'agit d'un "Modèle de Prime Pure" : dans ce type de modèle, tous les assurés ont une valeur de sinistre mais ce sinistre peut être nul ou très proche de 0.

— En tenant compte du nombre de sinistres et leur coût moyen : "Modèle Fréquence - Coût moyen".

Ces deux méthodes vont être abordées pour nos modélisations.

### 3.1.2 Approche prime pure

#### Loi de Tweedie

Les distributions Tweedie appartiennent à la classe des modèles de dispersion exponentielle et sont très utiles pour modéliser une distribution continue pour des valeurs supérieures à 0 avec une masse à 0, ce qui est le cadre de la modélisation "Prime Pure". Une particularité de cet ensemble de loi est la relation entre l'espérance et la variance qui se caractérise ainsi :

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

avec $\phi$ le paramètre de dispersion et $p \in \mathbb{R} \backslash ]0,1[$ un paramètre de puissance. En fonction de la valeur de cette puissance p ci-dessus les lois usuelles suivantes peuvent être retrouvées car ce sont des cas particulier de la loi de Tweedie :

— la loi Normale, lorsque p = 0
— la loi de Poisson, lorsque p = 1
— la loi Gamma, lorsque p = 2
— la loi Gaussienne inverse, lorsque p = 3

Pour d'autres valeurs de p, les distributions sont toujours définies mais ne peuvent pas être écrites dans une forme finie, et sont plus compliquées à estimer.
Le cas où $1 < p < 2$ va maintenant être étudié, il implique une loi Poisson composée avec sauts Gamma, qui est positive, très asymétrique à droite et qui possède une masse en 0. L'intérêt de ce type de lois est qu'elle permet de gérer un nombre important de valeurs nulles comme c'est le cas pour nos données, de modéliser la distribution de la charge non nulle par une loi adéquate, et de ne pas imposer une hypothèse d'indépendance entre la fréquence et le coût moyen.
Dans le cadre de cette étude, considérer une loi de Tweedie pour modéliser la charge des sinistres du produit Garages revient donc à faire les hypothèses suivantes :

— La charge annuelle est composée d'un nombre aléatoire de sinistres, ce nombre est supposé suivre une loi de Poisson
— Les montants des sinistres sont indépendants et identiquement distribués selon une loi Gamma
— Le nombre de sinistres est indépendant de leur coût : il s'agit de l'hypothèse d'indépendance entre fréquence et coût.

La fonction de lien utilisée ici sera la fonction *log*, qui est une fonction de lien adéquate avec la modélisation Tweedie et qui permet d'obtenir une formule de type multiplicatif pour le tarif. La fonction de lien *log* sera toujours utilisée dans la suite pour cette raison.

### 3.1.3 Approche fréquence / coût moyen

Le principe de l'approche "Fréquence / Coût moyen" est de modéliser de manière distincte et indépendante la fréquence de survenance d'un sinistre, et le coût d'un sinistre lorsqu'il y a survenance. Cette hypothèse d'indépendance est très forte mais cette méthode permet de vérifier l'efficacité et la stabilité de ce modèle de prime pure.

#### Modélisation de la fréquence

La fréquence correspond à la modélisation d'une variable de comptage. Plusieurs lois peuvent correspondre à ce type de problématique, parmi lesquelles : la loi de Poisson ou la loi Binomiale Négative (cf. B.2). La fonction la plus utilisée pour la modélisation de la fréquence de sinistres est la loi de Poisson et c'est ce qui sera utilisé ici. La fonction de lien utilisée pour la modélisation est la fonction de lien canonique, c'est-à-dire la fonction *log* pour la loi de Poisson.

#### Coût moyen

Le coût moyen correspond à la modélisation d'une variable strictement positive. Plusieurs lois possibles pour cette problématique, comme la loi Gamma et la loi log-Normale. La loi *Gamma* sera ici retenue pour la modélisation du coût moyen.

### 3.1.4 Paramétrage des modèles

#### Validation croisée

Afin de s'assurer qu'il n'y ait pas de sur-apprentissage sur les modélisations, une partie de la base de données (20%) ne sera pas utilisée pour le calibrage des modèles, et servira uniquement pour mesurer les performances de prédictions sur des données qui n'auront pas été retenues lors des modélisations. Cette base est appelée "base de validation", et la base qui servira au calibrage des modèles est appelée "base d'apprentissage".

De plus, afin de minimiser le plus possible le risque de sur-apprentissage, la méthode de la validation croisée va être utilisée. Il s'agit de découper la base d'apprentissage en $k$ échantillons de manière aléatoire pour calibrer le modèle sur $k-1$ échantillons, appelés "échantillons de train" et valider le modèle sur l'échantillon restant, appelé "échantillon de test". Plusieurs validations sont effectuées en prenant à chaque fois un échantillon de test différent parmi le découpage initial, pour ensuite retenir les performances moyennes sur les différentes étapes de modélisation. La base de données est ici découpée en quatre échantillons, ce qui peut se schématiser comme suit :

| | Base initiale | | | | |
|---|---|---|---|---|---|
| Etapes | Base d'apprentissage | | | | Base de validation |
| $1^{ere}$ | Test | Train | Train | Train | |
| $2^{eme}$ | Train | Test | Train | Train | |
| $3^{eme}$ | Train | Train | Test | Train | |
| $4^{eme}$ | Test | Train | Train | Test | |
| Validation finale | | | | | Base de validation |

Cette méthode permet d'assurer la stabilité du modèle en vérifiant que les indicateurs de qualité du modèle ont des valeurs proches sur les différents échantillons de test utilisés.

#### Définition des paramètres de calibrage du modèle

Avant de définir quel modèle appliquer à nos données, il faut définir la variable à expliquer, qui sera la variable à prédire, et les variables explicatives possibles selon les données à disposition. Pour faire une première sélection des variables à expliquer il est possible d'éliminer les variables les plus corrélées à une autre grâce à une matrice de corrélation, la modélisation permettra ensuite de déterminer son pouvoir explicatif sur la variable à prédire. En effet, si pour une variable donnée toutes les observations ont une

même modalité, alors la variable explicative n'aura pas ou peu d'impact prédictif sur la variable à expliquer. Il faut également affecter à chaque observation son exposition au risque, à savoir sa présence sur la période étudiée.
L'année de vision étudiée est prise en compte, sachant que trois années d'observations ont été retenues, ce qui va permettre une comparaison sur les différentes années et ainsi vérifier la stabilité du modèle à travers le temps.
Pour la modélisation avec l'approche prime pure, il faut également définir le paramètre $p$ de la loi de Tweedie, qui peut être calibré automatiquement selon la valeur qui s'adapte le mieux aux données.
Afin d'avoir un modèle qui correspond aux ajustements nécessaires à la distribution d'un produit, il est possible d'imposer des contraintes en amont du calibrage des modèles, comme la croissance des coefficients pour des variables continues telles que le chiffre d'affaires ou l'effectif.

## 3.2 Sélection de modèles

### 3.2.1 Sélection de variables

#### Critères de pénalisation

La sélection de variable est une étape importante dans la refonte d'une tarification puisqu'elle permet d'avoir un modèle parcimonieux. En effet, comme la modélisation GLM est basée sur la maximisation de la vraisemblance, plus le nombre de variables sera important, plus le modèle s'ajustera aux données. Ce comportement, s'il n'est pas contraint, aboutit donc à sélectionner des modèles avec le plus grand nombre de variables possibles. Cependant, pour des raisons opérationnelles et dans le cadre de la volonté de simplification de la souscription d'affaires nouvelles, il est important de procéder à une étape de sélection de variable afin de restreindre le modèle aux seules variables les plus pertinentes.
A cette fin, plusieurs critères peuvent être utilisés afin de prendre en compte une pénalisation des modèles ayant un nombre trop important de variables. Les critères les plus couramment utilisés sont :

— le Critère d'Information d'Akaïke (AIC) donné par : $2.log(L) - 2k$. Le but est de maximiser la log-vraisemblance du modèle notée $log(L)$ en pénalisant par deux fois le nombre du paramètre de ce modèle noté $k$.

— le Critère d'Information Bayésien (BIC) donné par $2.log(L) - k.log(n)$. Il s'agit ici également de maximiser la log-vraisemblance du modèle mais la pénalisation est ici d'autant plus importante que le nombre de variable noté $n$ est importante. Ce critère permet ainsi d'avoir des modèles encore plus parcimonieux.

L'outil utilisé ici n'utilise pas exactement ce type de critère mais se base tout de même sur ce type de structure. En effet, les critères AIC et BIC reposent sur une pénalisation par le nombre de paramètres, c'est-à-dire le nombre de degrés de liberté du modèle, mais

ce n'est pas cette donnée qui est retenue ici.
Dans le cadre des modèles linéaires simples, le nombre de paramètres correspond au nombre de variables. En effet, sur ce type de modèle il n'y a qu'un coefficient par variable, c'est pourquoi le nombre de paramètres coïncide avec le nombre de variables. Ainsi, plutôt que de considérer la pénalisation en termes de degrés de libertés, l'approche se base sur le critère du nombre de variables.
Cette valeur est identique pour les modèles linéaires simples, mais elle ne l'est généralement pas pour les GLM : chaque variable possédant plusieurs modalités donc potentiellement plusieurs coefficients. La méthode utilisée est alors inspirée des critères AIC et BIC, elle aboutira donc également à des modèles parcimonieux, l'objectif étant toujours le même, malgré le fait que la pénalisation ne soit pas basée sur les mêmes informations. Le détail de la pénalisation ne sera cependant pas présenté ici pour des raisons de confidentialité.

### Procédure de sélection de variables

Une fois le critère de pénalisation défini, il faut alors procéder à l'étape de sélection de variables. Le modèle retenu sera le modèle optimal au sens du critère choisi, c'est-à-dire celui qui minimise la valeur de ce critère. Il s'agit alors d'un problème d'optimisation.

**Recherche exhaustive** La meilleure méthode est alors de faire une recherche exhaustive : ce qui signifie tester tous les modèles possibles et retenir celui qui minimise le critère donné. Cependant, plus le nombre de variables est important, plus le nombre de modèles à tester sera élevé : de l'ordre de $\sum_{k=0}^{p} \frac{p!}{k!(p-k)!} = 2^k$ avec $k$ le nombre de variables retenues par modèle et $p$ le nombre de variables disponibles. Cette méthode peut alors s'avérer trop coûteuse en temps et en capacité de calcul pour être retenue dès lors que le nombre de variables potentiellement discriminantes est important comme c'est le cas pour cette étude. Par conséquent, cette méthode ne sera pas utilisée pour la procédure de sélection de variables.

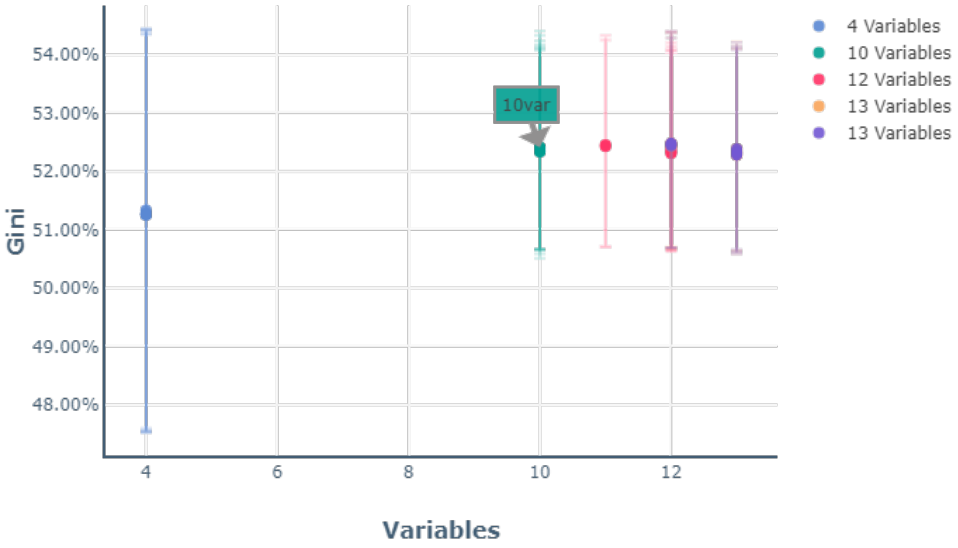
**Méthodes pas-à-pas** Ainsi, dans l'objectif de réduire le nombre de modèles à tester pour trouver celui qui minimise le critère, des méthodes dites "pas-à-pas" peuvent être utilisées. La procédure stepwise et la méthode descendante sont les plus courantes de par l'expérience de leur efficacité. Ces méthodes consistent à ne tester qu'un seul modèle pour un nombre de variables $k$ fixé parmi les $p$ présentes.
La méthode descendante consiste à considérer en premier lieu un modèle avec toutes les variables possibles. Un test de significativité des variables permet alors d'éliminer à chaque étape une variable considérée comme non significative, et ce jusqu'à ce que toutes les variables soient significatives au sens du test. Il y a alors au plus $p$ modèles testés, ce qui réduit de manière non négligeable le coût par rapport à la recherche exhaustive. Cependant, un inconvénient notable est le fait que l'ordre des variables testées peut avoir un impact important : une fois qu'une variable est éliminée, elle n'est plus considérée dans les modèles suivants. Cela peut notamment être problématique lorsqu'il y a des

variables corrélées entre elles.

Il existe également la méthode analogue, appelée méthode ascendante où le premier modèle testé ne contient qu'une variable, et les autres variables sont ajoutées au fur et à mesure en fonction des résultats au test de significativité. Par conséquent, la même problématique que les autres méthodes pas-à-pas se retrouve ici. Néanmoins, une procédure différente mais basée sur le fonctionnement de la méthode ascendante permet de réduire cet effet : il s'agit de la procédure stepwise. La première étape consiste également à faire le test en considérant une seule variable au modèle, mais dans ce cas, les $k$ modèles possibles pour un modèle à une variable sont testés au lieu d'un seul. Le modèle retenu à l'issue de cette première étape est celui qui a la significativité la plus élevée. L'étape suivante consiste alors à réitérer ces tests sur une nouvelle variable, en prenant pour acquis la(les) variable(s) précédemment retenue(s). Cette étape est ainsi réitérée jusqu'à ce que la variable dernièrement ajoutée ne soit plus considérée comme significative sur les modèles testés.

**Méthode retenue** L'outil utilisé reprend les principes de cette dernière méthode, tout en essayant de se rapprocher de la recherche exhaustive en augmentant le nombre de modèles testés à chaque étape. Cette valeur correspondant au nombre de modèles par étape se choisit dans l'outil lors du paramétrage de la modélisation. L'expérience de la modélisation de cette étude sur cet outil a permis de juger que la valeur à retenir pour ce paramètre était de 5. En effet, pour cette modélisation, cela s'est avéré à la fois suffisant pour avoir un panorama de différents modèles pour chaque nombre de variables, et à la fois pas trop coûteux en temps de calcul comme pourrait l'être la recherche exhaustive.

Les différents modèles proposés selon les paramétrages effectués se présentent ainsi :




Les points correspondent à chaque modèle retenu selon les critères de pré-sélection de variables présentés. L'axe des abscisses représente le nombre de variables discriminantes retenues pour les modèles, qui est compris dans l'intervalle indiqué au préalable. Pour chaque valeur en abscisses, un certain nombre de modèles sont proposés selon ce qui a été défini dans les paramétrages. Ici, le paramètre retenu pour ce critère est de cinq et cela s'observe sur ce graphique : sur les différentes possibilités de modèles, pour un nombre de variable fixé, cinq modèles plus ou moins lissés sont présentés.

### 3.2.2 Indicateurs de qualité des modèles

Plusieurs critères peuvent être utilisés pour apprécier la qualité d'une modélisation et il ne faut pas se réduire à l'analyse d'un seul de ces indicateurs. En effet chaque indicateur permet de juger de la qualité d'un modèle à travers un critère donné mais la qualité d'un modèle ne se résume pas à un seul critère. Les différents indicateurs qui seront retenus pour la sélection de nos modèles vont être présentés.
En premier lieu les indicateurs graphiques que sont les courbes Lift et de Lorenz seront introduits, ainsi que l'analyse des résidus. Puis, les indicateurs numériques utilisés à savoir le Gini et l'erreur quadratique moyenne seront présentés plus en détail. De plus, une présentation du Spread permettra de mieux appréhender l'ordre d'importance des variables dans le modèle. Ces indicateurs vont permettre de sélectionner les meilleurs modèles grâce à leur valeur donnée sur les différents découpages de la base d'apprentissage, mais il faut consolider ces choix en les analysant sur la base de validation qui n'aura pas servi à la modélisation.

### La courbe lift

La courbe lift, plus connue sous le nom de lift curve est principalement utilisée pour les modélisations par apprentissage statistique mais permet tout autant d'apprécier la qualité de prédiction des modèles linéaires généralisés. Le principe est de vérifier que l'ordonnancement des prédictions respecte l'ordre des valeurs cibles observées.
Pour cela plusieurs étapes sont nécessaires :

- les observations sont classées par ordre croissant de prédictions,
- puis 20 intervalles équi-répartis sont créés en conservant l'ordre des prédictions une fois triées, pour avoir des quantiles de 5% d'exposition
- enfin la moyenne par quantile est calculée : à la fois pour les prédictions et pour la charge observée
- il est alors possible de représenter la courbe des moyennes modélisées et observées par quantile

. Ainsi une analyse des écarts moyens par quantile et du comportement des courbes l'une par rapport à l'autre est aisée. En effet, la courbe des prédictions sera croissante par construction, mais si la courbe des charges observées est décroissante alors le niveau de risque n'aura pas été modélisé correctement. De plus, pour vérifier la qualité de la prédiction, l'analyse des écarts entre les deux courbes pour chaque quantile permettra d'atteindre un objectif qui est d'avoir des courbes qui se superposent. En effet cela signifie alors que l'observé et le prédit sont égaux en moyenne.

### La courbe de Lorenz

La courbe de Lorenz peut se rapprocher de la courbe ROC utilisée dans les régressions binaires. Le but est d'obtenir la courbe la plus éloignée de la bissectrice dans le cadre $[0, 1]$. Chaque courbe lift passe par les points suivants : $(0, 0)$ et $(1, 1)$ quelle que soit la qualité de la prédiction. Présentons d'abord les deux extrêmes possibles :

- La bissectrice correspond à la représentation de la fonction indicatrice, et peut s'interpréter comme la représentation d'un modèle qui attribue la même valeur à toutes les variables proposées, ce qui n'est pas le but de ce type de modélisation : les variables du modèle doivent être discriminantes pour prédire au mieux la variable cible, à savoir la charge sinistre.
- A l'inverse, la courbe optimale, passant par le point $[0, 1]$, attribue toute la charge à une seule, ce qui représente le cas où la qualité de discrimination du modèle est maximale.

Le but est alors de s'approcher le plus possible de ce deuxième extrême pour avoir un modèle le plus parcimonieux possible, c'est-à-dire où la plupart des coefficients sont nuls afin de minimiser le nombre de variables en maximisant la qualité de prédiction.
De plus, l'indice de Gini se calcule à partir notamment de l'aire sous la courbe de Lorenz, il s'agit donc d'une représentation visuelle de cet indicateur.

### L'analyse des résidus

**Définition des résidus** Définissons d'abord la notion de résidus, aussi appelés le bruit du modèle. Il s'agit de la part qui n'a pas pu être modélisée, une des structures possibles, et la plus courante pour les modéliser est la suivante :

$$R = y_{obs} - \hat{y}$$

Une modélisation par la méthode des GLM, nécessite la vérification des hypothèses suivantes sur les résidus, ils doivent :

— être indépendants

— suivre une loi normale centrée en 0

— être homogènes

L'analyse des résidus va notamment permettre de valider l'homogénéité et la normalité de ceux-ci. Cependant, plusieurs types de résidus peuvent être représentés (*CNAM*) :

— les résidus additifs sont les résidus présentés plus haut, il s'agit des résidus les plus intuitifs et simples d'interprétation : $y_{obs}$ correspond à la valeur du risque observée et $\hat{y}$ à la valeur du risque prédite par la modélisation GLM. Cependant ils ne correspondent pas à une tarification basée sur une formule multiplicative, or le but étant de simplifier la structure tarifaire, cette proposition ne sera pas conservée dans le but de rester sur une structure tarifaire simplifiée

— les résidus multiplicatifs, ceux-ci permettent de prendre en compte cette contrainte de structure. Ils sont donnés par : $R = \frac{y_{obs}}{\hat{y}}$.

— les résidus de Pearson, aussi appelés résidus standardisés, sont des résidus centrés mais pas réduits. Leur structure est la suivante : $R = \frac{y_{obs} - \hat{y}}{\sigma \hat{y}}$.

— les résidus de déviance se basent sur la vraisemblance des données en donnant la contribution de chaque observation à la vraisemblance du modèle, et permettent de prendre en compte la structure du modèle étudié (qui minimise l'erreur quadratique). Ces résidus sont de la forme suivante : $R = signe(y_{obs} - \hat{y}).\sqrt{|d((y_{obs}, \hat{y})|}$, où $d((y_{obs}, \hat{y}) = 2.\sum_i (y_i.\log(\frac{y_i}{\hat{y_i}})^{\smile} y_i + \hat{y_i})$

**Les résidus agrégés** Une autre méthode pour valider un modèle en utilisant la notion de résidus est d'analyser les résidus agrégés (*Emblem User's Guide*). Ce type de résidu est généralement utilisé pour des modèles de Fréquence ou de Prime pure du fait des volumes importants de 0 dans la variable à prédire (valeur non présente dans les modèles de Coût-Moyen). Ces valeurs nulles peuvent en effet biaiser les conclusions des résidus puisque dès lors que la prédiction sera différente de 0, le résidu sera toujours strictement négatif : la forme des résidus ne pourra alors pas correspondre aux critères demandés. Cependant, cela ne signifie pas que le modèle n'est pas correct : un individu n'ayant pas de sinistre ne devrait pas avoir pour autant une prime nulle, et donc aurait des résidus strictement négatifs. Par conséquent ce type de cas aboutirait à une forme inadaptée des résidus ; alors qu'une prime nulle, qui n'est généralement pas souhaitable, permettrait

d'avoir une forme de résidus correcte. Cette contradiction peut être corrigée avec les résidus agrégés.
Ceux-ci correspondent aux résidus de chaque groupe de données ayant la même prédiction, ou une prédiction très proche. Ils sont donnés par :

$$\text{Résidus agrégés} = \frac{\sum_i^m (\text{valeur observée}_i \text{ - valeur prédite}_i)}{\sqrt{V\left(\sum_i^m \text{valeur prédite}_i\right)}}$$

avec $m$ le nombre de valeurs prédites différentes et $V()$ la variance.

**Représentation des résidus** Les modèles retenus seront ceux pour lesquels les résidus agrégés seront proches de 0. Graphiquement cela peut se représenter en fonction des prédictions pour vérifier qu'il n'y ait pas de structure particulière qui montrerait un biais dans la modélisation. En effet, comme vu précédemment, la distribution des résidus doit suivre une loi Normale pour avoir une bonne qualité de prédiction, alors la représentation attendue sera de telle sorte que les résidus seront centrés autour de 0. De plus, afin de faciliter la lecture de ce type de graphique la représentation pourra être sous la forme d'un point par couple de donnée (prédiction, résidu) en insérant une troisième dimension de représentation à travers un jeu de couleur ou un troisième axe pour montrer le volume d'observations pour chaque couple de point (prédiction, résidu).

### Gini

L'indice de Gini peut se calculer à partir de l'aire sous la courbe de Lorenz :

$$\begin{aligned} Gini &= 2 \times \frac{\text{Aire entre la courbe du Lorenz du modèle et la courbe aléatoire}}{\text{Aire entre la courbe optimale et la courbe aléatoire}} \\ &= 2 \times \text{Aire entre la courbe du Lorenz du modèle et la courbe aléatoire} \end{aligned}$$

Avec

- la courbe aléatoire correspondant à la bissectrice
- la courbe optimale correspondant à la courbe passant par le point $(0, 1)$

Le but est toujours de maximiser sa valeur pour avoir une bonne qualité de prédiction.

### L'erreur quadratique moyenne : MSE (Mean Squared Error)

L'erreur quadratique moyenne mesure la moyenne des carrés des écarts entre les prédictions et les observations. Elle est donnée par :

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

Avec n, le nombre d'observations et k le nombre de variables du modèle retenu.

Le meilleur modèle au sens de cet indicateur sera celui qui minimise cette valeur car le but est que l'erreur soit la plus faible possible. Cet indicateur a une version dérivée qui est le RMSE (Root Mean Squared Error), il s'agit de la racine carrée de l'erreur quadratique moyenne : $RMSE = \sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \hat{y_i})^2}$

**Le spread**

Cet indicateur associé à chaque variable discriminante retenue dans le modèle permet d'appréhender le degré d'importance de chacune d'elles à travers la dispersion des coefficients. Il correspond au ratio entre le coefficient le plus élevé associé à une modalité, et le coefficient le plus faible d'une autre modalité :

$$\text{Spread d'une variable} = \frac{\text{coefficient le plus élevé}}{\text{coefficient le plus faible}}$$

Cet indicateur correspond à la dispersion sur tout le portefeuille, mais pour éviter d'avoir une vision trop biaisée par les extrêmes il est possible de le calculer en excluant 5% de l'exposition du portefeuille parmi ceux ayant les coefficients les plus élevés et de même sur ceux ayant les coefficients les plus faibles. Le spread 100/0% sera alors distinct de ce second indicateur : le spread 95/5%.

## 3.3 Zonier

**Objectif** Dans le cadre de la refonte du produit, et notamment de cette garantie, il est apparu important de développer un nouveau zonier adapté à ce risque. Actuellement, il s'agit d'un zonier utilisé précédemment pour la modélisation du risque géographique RC Automobile pour les produits Automobile du Particulier et des Professionnels. Ce zonier n'avait donc pas été calibré sur des assurés du produit Garages étudié ici, le but est ainsi d'avoir un zonier spécifique au risque géographique lié à ce produit et ce risque.
La création du zonier (*SAID*, 2016) est une étape qui intervient après la modélisation GLM avec les variables explicatives présentes dans la base. En effet, la prise en compte de la dimension géographique du risque se calibre sur les résidus de la régression. Il s'agit alors d'une variable qui est construite indépendamment des autres variables explicatives et qui permet de modéliser la part du risque non modélisée par ces variables. Cela améliorera la modélisation si le risque a une dimension géographique, sinon la création du zonier ne diminuera pas les résidus de la modélisation. La maille retenue pour ce zonier sera celle du Code INSEE qui est une maille suffisamment fine pour prendre en compte les spécifités géographiques mais sans aller dans un niveau de détail trop poussé qui ne serait pas stable du fait du faible nombre de contrats présents à cette maille : il existe au plus 20 contrats par Code INSEE, une modélisation plus précise ne serait donc pas adaptée. La modélisation se base sur les résidus de déviance, explicités plus tôt (cf. 3.2.2) car ils prennent en compte les aspects nécessaires à la modélisation du zonier en considérant la structure retenue pour le modèle.

La modélisation du zonier consiste en la prédiction de ces résidus qui doivent répondre à certaines contraintes techniques, mais plusieurs contraintes opérationnelles peuvent également être introduites.

**Théorie** La contrainte principale est le fait qu'il faille considérer que les prédictions doivent être lissées avec celles des voisins géographiques afin de ne pas avoir une carte trop granulaire. Cette contrainte sera prise en compte grâce à la projection des Codes INSEE de chaque risque sur un plan ou sur une carte à travers la latitude et longitude de chaque ville afin de permettre au modèle de s'adapter à cette contrainte géographique. Plusieurs degrés de lissage seront testés et cette appréciation du niveau de lissage géographique est indiquée par la valeur de la distance de Moran [1] : plus sa valeur est élevée, plus le zonier est lissé, et donc moins un Code INSEE a de voisins différents. Il faut alors choisir un zonier qui permet d'avoir un degré de lissage suffisant afin de ne pas avoir de coefficients trop distincts entre deux villes géographiquement proches, mais aussi de créer des zones de risques plutôt que d'associer un coefficient spécifique à chaque Code INSEE. Ce zonier doit cependant modéliser au mieux le risque associé à chaque ville.

**Contraintes opérationnelles** Par ailleurs, d'autres contraintes supplémentaires peuvent être ajoutées, notamment opérationnelles. La contrainte principale étant ici de fixer le nombre maximum de modalités et donc de zones modélisées afin de ne pas avoir plus de 13 coefficients dans la structure tarifaire du zonier, pour reprendre ce qui est utilisé dans le tarif actuel.

1. La théorie de cet indice ne sera pas présentée ici car ce n'est pas l'objet de ce mémoire, mais le lecteur pourra se référer à l'ouvrage suivant pour plus de détails : [S. OLIVEAU] - "Autocorrélation spatiale : leçons du changement d'échelle", *L'Espace géographique 2010/1*, Vol. 39, 51-64

## 3.4 Les modèles GLM testés et résultats sur la base d'apprentissage

De manière synthétique, il est ainsi possible de comparer différents modèles que ce soit en approche prime pure pour apprécier l'impact du nombre de variables retenues et du degré de lissage, ou en comparaison des approches prime pure et fréquence × coût moyen. Plusieurs modèles ont été testés, mais seuls deux d'entre eux, ceux ayant les meilleures performances seront détaillés ici. Pour les autres modèles, le détail de la charge à modéliser se trouve dans la partie de retraitements de la variable à expliquer (2.4).

### 3.4.1 Les modèles GLM testés

De nombreux modèles ont été testés, que ce soit en approche Prime Pure ou Fréquence × Coût-Moyen, avec parmi eux :

- Une modélisation de la charge hors graves
- Un modélisation de la charge écrêtée
- Une modélisation de la charge avec mutualisation des sinistres graves, hors atypiques en fonction de la part de charge écrêtée,
- Une modélisation de la charge hors atypiques non retraitée

Le modèle retenu pour la modélisation des sinistres graves est la charge hors atypiques, pour des raisons de performances de ce modèle, cette charge sera alors celle retenue pour la suite de cette étude.

La notion de recours a ensuite aboutie à de nombreuses autres modélisations :

- Modèle avec distinction des sinistres responsables et non responsables à travers ces deux approches :
  - une approche Prime pure :

    $$\text{Prime Pure} = \text{Prime Pure des responsables} + \text{Prime Pure des non responsables}$$

  - une approche Fréquence × Coût-Moyen :

    $$\begin{aligned} \text{Prime Pure} &= \text{Fréquence des responsables} \\ &\times \text{Coût Moyen des responsables} \\ &+ \text{Fréquence des non responsables} \\ &\times \text{Coût Moyen des non responsables} \end{aligned}$$

- Modèle avec distinction par nature de la charge sinistre, sous la structure suivante :

  $$\begin{aligned} \text{Prime Pure} = {} & \text{Fréquence des forfaitaires} \times \text{Coût Moyen des forfaitaires} \\ & - \text{Fréquence des négatifs} \times |\text{Coût Moyen des négatifs}| \\ & + \text{Fréquence hors recours} \times \text{Coût Moyen hors recours} \end{aligned}$$

mais aussi en approche Prime Pure

— Modèle avec distinction des sinistres responsables et non responsables, avec l'ajout de l'information des circonstances de la convention IRSA ou non, en approche Prime Pure et Fréquence × Coût-Moyen :

$$\begin{aligned}\text{Prime Pure} &= \text{Fréquence des responsables issus de la convention}\\ &\times \text{Coût Moyen des responsables issus de la convention}\\ &+ \text{ Fréquence des responsables non issus de la convention}\\ &\times \text{Coût Moyen des responsables non issus de la convention}\\ &+ \text{ Fréquence des non responsables}\\ &\times \text{Coût Moyen des non responsables}\end{aligned}$$

— Modèle avec mutualisation de la charge forfaitaire et négative au prorata de la prime acquise des contrats sinistrés.

Le modèle retenu pour la modélisation des recours est celui de la charge mutualisée, pour des raisons de performances de ce modèle par rapport aux autres. Ces performances relatives peuvent s'expliquer par les faibles volumes disponibles pour ce produit ; la conclusion pourrait ne pas être identique sur des produits avec davantage de volumes de modélisation.

### 3.4.2 Les modèles retenus et leurs performances

Un tableau permet de synthétiser les valeurs des indicateurs pour les deux modèles retenus : le modèle hors atypiques en approche Fréquence × Coût-Moyen [1], et le modèle avec mutualisation de la charge associée aux recours. Ce récapitulatif permet de juger de la pertinence de ces derniers.

**Comparatif entre un modèle de Prime Pure avec mutualisation et un modèle de Fréquence × Coût-Moyen**

1. Les résultats du modèle de Fréquence et ceux du modèle de Coût-Moyen sont en annexes : cf. C

***Spread***

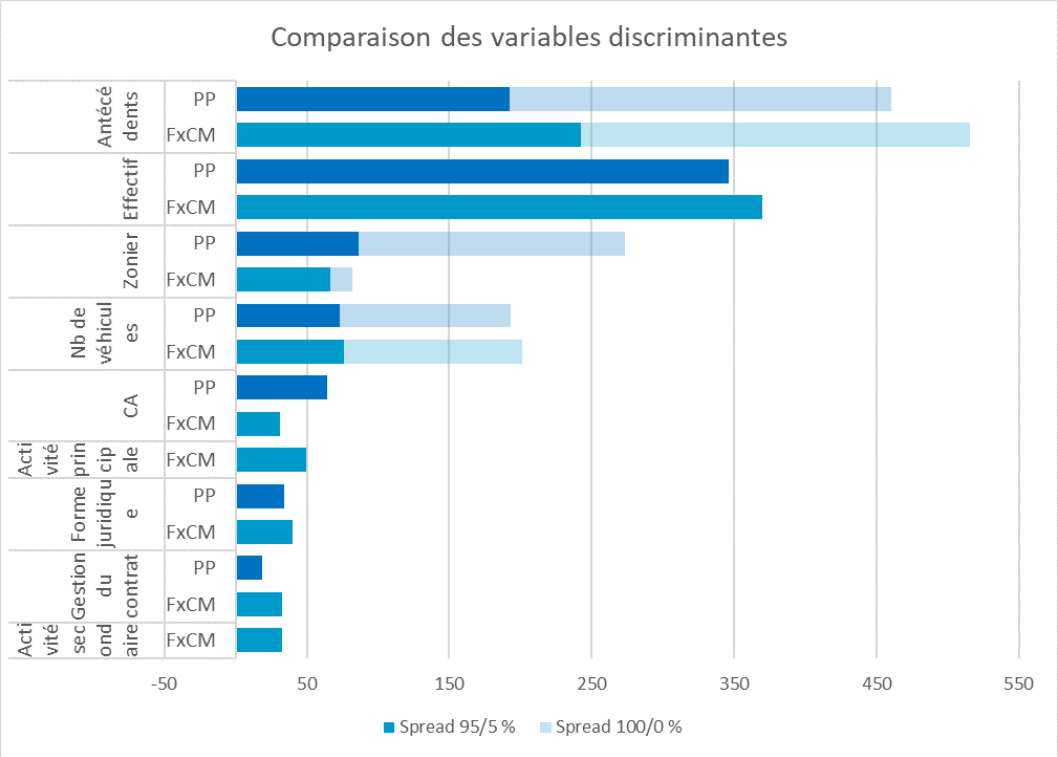



FIGURE 3.1 – Comparatif de l'importance des variables

Le spread permet d'apprécier l'importance des variables et ainsi établir un ordre des variables les plus pertinentes. Il est important de noter que le modèle de Fréquence × Coût-Moyen comporte plus de variables discriminantes, ce qui est dû à la construction de ce type de modèle : les variables du modèle de Fréquence et du modèle de Coût-Moyen sont retenues, ce qui peut démultiplier le nombre de variables dès lors que certaines ne sont pas en commun sur les deux modélisations distinctes. Cependant, l'ensemble des variables du modèle en approche Prime Pure sont contenues dans celui du modèle en approche Fréquence × Coût-Moyen, et l'ordre des variables reste proche.
Les variables d'antécédents et d'effectif sont des variables majeures dans cette modélisation, il ne s'agit pas de variables liées à la typologie de véhicules, comme cela pourrait être mis en avant avec les variables d'activité, qui ne sont présentes que sur un seul des modèles. Il s'agit plutôt d'un effet quantitatif qui joue sur la charge finale comme le montrent les variables d'effectif, de nombre de véhicules, et de chiffre d'affaires.

### *Courbe Lift*

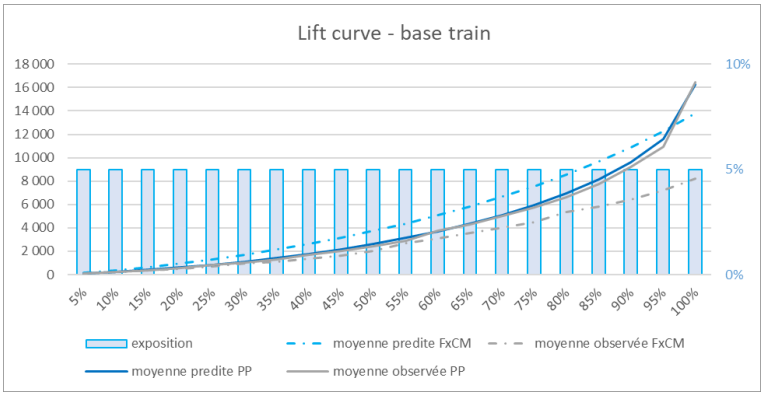



FIGURE 3.2 – Comparatif des courbes Lift

La courbe Lift du modèle de Fréquence × Coût-Moyen (en pointillés) montre d'importantes différences entre l'observé et le prédit, ce qui est révélateur d'une répartition du portefeuille dégradée par rapport à ce qui est réellement observé, contrairement au cas du modèle de Prime Pure où les deux courbes sont proches.

### *Courbe de Lorenz*

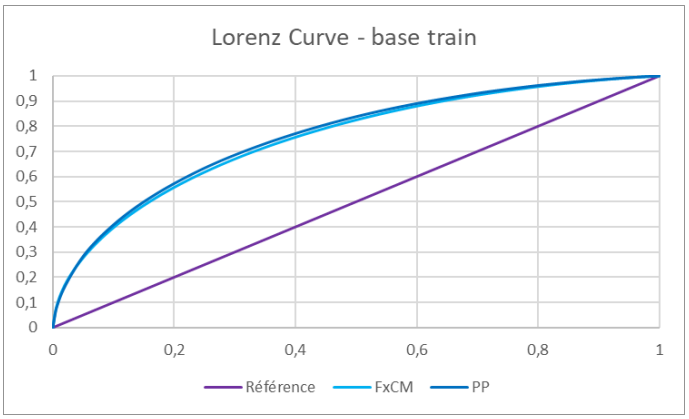



FIGURE 3.3 – Comparatif des courbes de Lorenz

La représentation des courbes de Lorenz des deux modèles permet de montrer une modélisation légèrement plus adaptée du modèle de Prime Pure, mais les deux courbes restent très proches et représentatives de bonnes performances.

***Résidus***

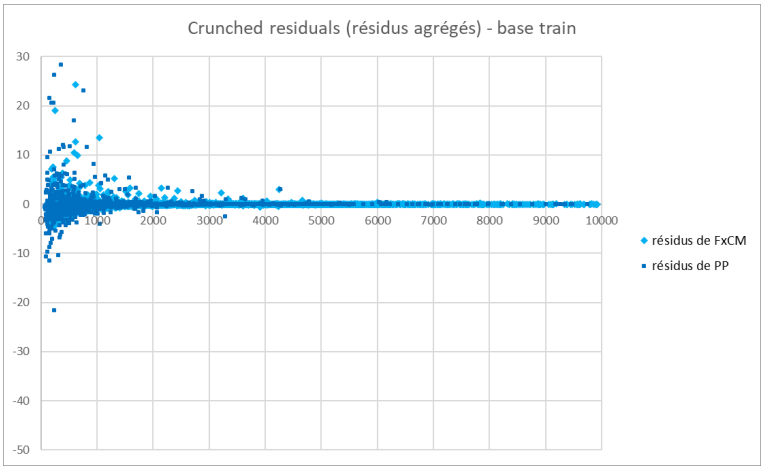



FIGURE 3.4 – Comparatif des résidus agrégés

L'analyse des résidus agrégés permet de vérifier qu'il n'y a pas de biais dans la modélisation et que les hypothèses des GLM sont vérifiées, et ici on peut noter que les résidus sont proches de 0 et n'ont pas de forme spécifique. Cela semble d'autant plus marqué sur le modèle de Prime Pure.

***Autres indicateurs statistiques***

| Indicateurs | Modèle en approche Prime Pure | Modèle en approche Fréquence × Coût-Moyen |
|---|---|---|
| Gini | $52,4\%$ | $50,6\%$ |
| RMSE | 8522 | 7471 |

TABLE 3.1 – Tableau comparatif de modèles en approche Prime Pure et Fréquence × Coût-Moyen

Les indicateurs confirment les conclusions à partir des représentations précédentes. Le Gini du modèle de Prime Pure est légèrement supérieur à celui du modèle de Fréquence × Coût-Moyen, mais le RMSE de ce dernier est plus élevé, ce qui contrebalance les résultats.

***Premières conclusions*** Au vu de l'analyse de l'ensemble des performances, le choix du modèle se porte sur l'approche Prime Pure

Ces données correspondent ici aux valeurs des indicateurs sur la base d'apprentissage (train), et permet de faire une première sélection de modèles, mais la prise de décision doit s'effectuer en prenant en compte les résultats sur la base de validation qui n'a pas été utilisée pour le calibrage des modèles. Ces résultats seront présentés ultérieurement pour répondre à la problématique du sujet.

**Zoniers associés** Dans l'exemple de la modélisation en approche prime pure, différents zoniers peuvent alors être créés selon les contraintes fixées (les zoniers étant similaires avec l'approche Fréquence × Coût-Moyen, seuls ceux de l'approche Prime Pure sont présentés ici). Une comparaison visuelle permet de montrer les différences entre les cas suivants de contraintes :

— un zonier sans contrainte du nombre de modalités et en retenant le lissage qui maximise le Gini, c'est-à-dire sans contrainte spécifique sur le lissage

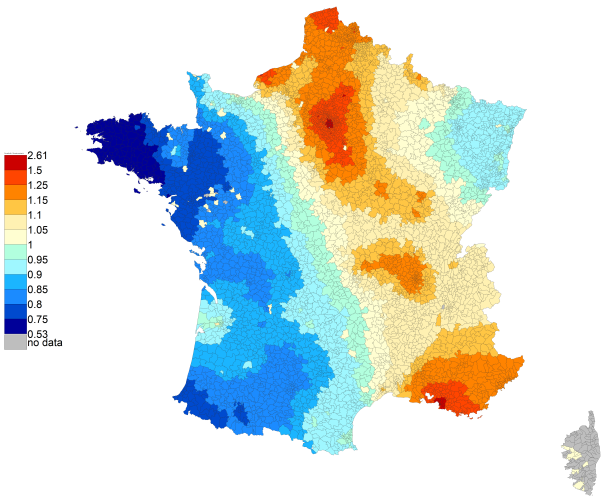



FIGURE 3.5 – Zonier sans contrainte

— des zoniers sans contrainte du nombre de modalités mais en considérant dans un premier temps un lissage très faible, ce qui correspond à une très forte granularité dans les zones, et dans un second temps un lissage très poussé afin de diminuer le plus possible les écarts entre des zones proches, qui permettent de répondre, de façon extrême ici, à des potentielles contraintes opérationnelles d'application et de compréhension du tarif

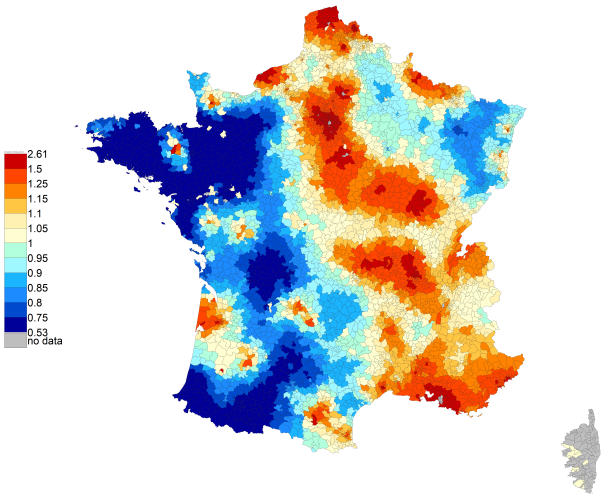



FIGURE 3.6 – Zonier avec une contrainte de lissage très faible

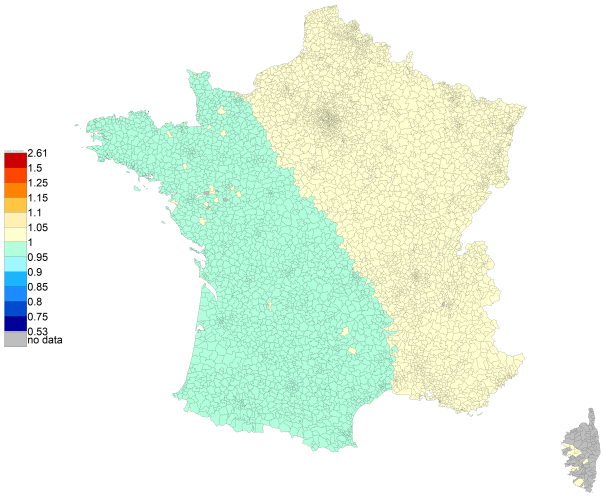



FIGURE 3.7 – Zonier avec une contrainte de lissage très forte

— un zonier où le nombre de zones est fixé a priori, avec un lissage maximisant le Gini, pour des contraintes opérationnelles plus fortes où le nombre de modalités ne peut dépasser un certain seuil

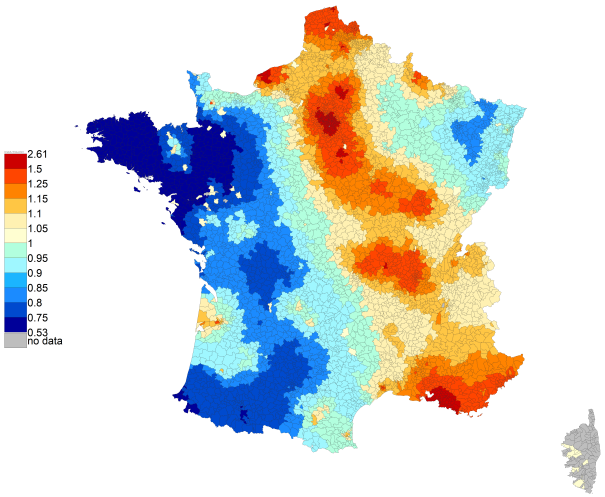



FIGURE 3.8 – Zonier avec une contrainte de nombre de modalités

Le degré de lissage est l'une des contraintes les plus fortes sur le zonier comme le montrent les visuels de ces cartes. En revanche le nombre de modalités, fixé à 13 sur le dernier cas ne montre pas de variation aussi notable par rapport à l'exemple sans contrainte, et la fixation du nombre de modalité facilite l'implémentation opérationnelle, ce dernier zonier sera donc conservé.

## 3.5 Comparaison avec une méthode de Machine Learning : les forêts aléatoires

### 3.5.1 Théorie des méthodes

Les méthodes qui seront utilisées pour challenger les modèles linéaires généralisés sont des méthodes de régression dites "supervisées" (*BELLINA*, 2014). En effet, cette base comporte des données déjà labellisées, les algorithmes vont alors pouvoir se baser sur les variables explicatives pour prédire la variable cible, à savoir la charge sinistre du contrat pour une année donnée. Ainsi, le but est ensuite d'appliquer les résultats de cette modélisation sur de nouvelles variables explicatives afin d'obtenir la prime pure prédite, de la même façon qu'avec les GLM. De nombreux algorithmes de ce type existent mais nous n'allons retenir que quelques uns d'entre eux pour cette comparaison. En effet, le but est de montrer les différences de résultats entre la modélisation classique utilisée en assurance, à savoir les GLM, et des modélisations permettant de diminuer les hypothèses sur les données. En effet, les GLM se basent sur des hypothèses d'indépendance des variables et d'adéquation des données aux lois usuelles qui sont des hypothèses fortes. Les méthodes de Machine Learning permettent ainsi de s'affranchir de ces hypothèses : il s'agit de méthodes non paramétriques, à l'inverse des GLM qui se basent sur des familles de lois pour décrire la distribution des données.
Les algorithmes utilisés aujourd'hui ont été créés depuis de nombreuses années mais les puissances de calculs n'étaient alors pas suffisantes pour les mettre en pratique. La résurgence de ces derniers est ainsi due à l'apparition de machines permettant de faire tourner ces algorithmes qui nécessitent un nombre important de données pour un bon calibrage des modèles. En effet, la puissance de ces algorithmes réside dans l'expérience acquise grâce aux informations retenues par chaque observation, ainsi plus la quantité de données sera importante, meilleure sera la calibration du modèle.
Dans la suite nous allons alors nous focaliser sur des modèles dont l'interprétation est facilitée grâce à l'information de la contribution de chaque variable dans le modèle. Il s'agit des algorithmes basés sur les arbres de régression (*Pericles Actuarial*).

#### Arbres de régression

Les arbres de régression font partie des méthodes d'apprentissage supervisées les plus appréciées en assurance du fait de l'interprétabilité des résultats qui est importante pour des problématiques métier. En effet, les méthodes d'apprentissage statistiques n'offrent pas ou peu la possibilité de comprendre l'effet de chaque variable dans une modélisation, comme cela est possible avec les GLM. Néanmoins, les méthodes basées sur des techniques arborescentes permettent par construction d'obtenir des degrés de contribution des variables dans l'apprentissage.
Le principe des arbres de régression (*BROWNLEE*, 2020) est de définir de manière récursive, par des tests sur les variables explicatives, des groupes homogènes d'observations. Cette structure se forme à partir de l'expérience des données. Pour synthétiser, cette méthode peut se représenter grâce au schéma suivant :

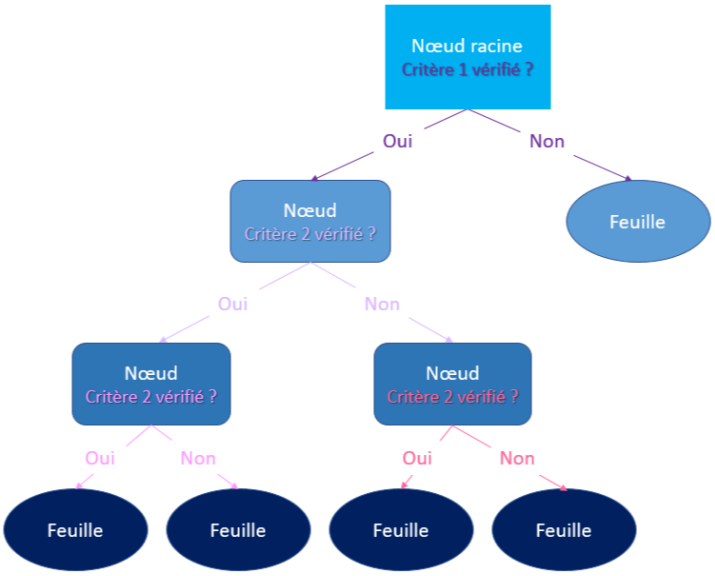



FIGURE 3.9 – Schéma d'un arbre de régression

Les feuilles correspondent aux groupes homogènes recherchés qui nécessitent plus ou moins de tests pour être obtenus, et les tests sont ici appelés des noeuds : à chaque noeud, l'observation va être dans l'un des sous-groupes selon le résultat du test. Le but est alors de définir quels sont les tests les plus pertinents afin d'établir des groupes homogènes de risque, et donc permettre de déterminer la valeur à prédire pour chaque groupe.

**Algorithme CART** Plusieurs algorithmes existent dans la catégorie des arbres de régression, nous allons ici retenir l'algorithme CART : Classification And Regression Trees ie. Arbres de classification et de régression (*CHESNEAU*, 2020).
Cet algorithme nécessite de déterminer en amont les paramètres suivants :

- le test qui va permettre de définir les sous-groupes à savoir le critère de construction qui se décompose en deux parties :
  - la variable sur laquelle va porter le test et donc l'ordre dans lequel les variables sont testées
  - le seuil de la variable qui va permettre de classer dans l'une ou l'autre des feuilles l'observation selon sa valeur par rapport au seuil
- la règle d'arrêt afin de ne pas avoir un découpage trop fin qui aboutirait à un sur-apprentissage, plusieurs types de règles peuvent être retenus comme :
  - la profondeur maximale de l'arbre : c'est à dire le nombre de niveaux de noeuds
  - le nombre minimal d'observations par feuille
  - la valeur minimale d'un indice donné, par exemple le Gini

— la règle d'assignation qui correspond au type d'erreur pour définir la réponse au test : l'observation sera classée de telle sorte que la valeur de la règle d'assignation soit minimale au sein d'une "feuille",
  — dans le cas d'une régression, il s'agira de minimiser la variance
  — alors que dans le cas de la classification, l'indice de Gini ou de l'entropie pourront être utilisés

La difficulté associée à cette méthode est précisément la détermination de ces critères et notamment de la règle d'arrêt. En effet, dans le but de maximiser la qualité de la prédiction, il y a de nombreux risques de sur-apprentissage et cela va dépendre de chaque jeu de données. Il y alors généralement soit des prédictions de faible qualité, soit un sur-apprentissage des données. Cependant, cette méthode, de part son interprétabilité, sert de base à d'autres méthodes plus poussées qui permettent d'obtenir de meilleurs résultats. Cela est notamment le cas pour une autre méthode retenue ici, à savoir les Forêts aléatoires (Random Forest).

**Random Forest**

Les forêts aléatoires font également partie des méthodes basées sur une structure arborescente. Cependant, la différence réside dans l'approche utilisée pour l'apprentissage. En effet, il s'agit de reprendre le fonctionnement des méthodes dites de "Bagging" pour les appliquer aux arbres de régression.
La dénomination "Bagging" provient de la contraction des termes suivants :

— Bootstrap : qui est une méthode d'échantillonnage par tirage avec remise. Le principe du bootstrap est de répliquer des échantillons à partir de la base de données initiale. Cela est notamment très utilisé lorsque la quantité de données est insuffisante pour avoir un apprentissage statistique qualitatif. En effet, ces méthodes nécessitent une quantité très importante de données pour être fiables.
— Aggregating : le principe est alors d'agréger les résultats des méthodes appliquées sur des échantillons. Cela a pour but de diminuer la variance et d'avoir de meilleurs performances sur chaque modélisation, pour ensuite les agréger pour avoir la modélisation finale.

La dimension aléatoire des forêts est ainsi en partie apportée par cette agrégation de plusieurs arbres calibrés sur des échantillons restreints. Ces échantillons ayant été créés aléatoirement par la méthode de "bootstrap". Une autre dimension aléatoire est également apportée par le fait que seule une partie des variables est retenue à chaque étape de la modélisation, et celles-ci sont tirées aléatoirement.
L'apprentissage se calibre alors à la fois sur une part aléatoire des données ayant été ré-échantillonnées, et sur une part aléatoire des variables, ce qui multiplie les sources d'apprentissage et améliore ainsi les performances par rapport à des modèles d'arbres de régression simples comme la méthode CART.

Plusieurs paramètres restent cependant à calibrer :

— le nombre d'échantillons à créer avec la méthode bootstrap, ce qui correspond au nombre d'arbres à agréger, qui forment la forêt

— le nombre de variables qui peuvent être sélectionnées aléatoirement à chaque noeud. Un critère pour aider à définir ce paramètre est : le tiers du nombre de variables (en partie entière) ne doit pas être inférieur à 5 dans le cas des régressions

— au vu du nombre d'arbres qui peut être important, il faut également définir un nombre maximal de feuilles par arbre pour ne pas trop démultiplier la modélisation.

Ces paramètres peuvent également être calibrés en amont par des méthodes d'optimisation, car la détermination n'est pas aisée si ce n'est par l'expérience des données.

### 3.5.2 Retraitements supplémentaires de la base de données

De nombreux retraitements ont déjà été effectués sur la base de données pour la tarification avec la méthode GLM mais les méthodes d'apprentissage statistique nécessitent des retraitements plus poussés.

#### Données manquantes

Dans les étapes précédentes, les variables catégorielles dont certaines valeurs étaient manquantes ont été remplacées dans la plupart des cas par la modalité la plus représentée, ou par une autre modalité selon des hypothèses établies. Pour les variables qualitatives il n'y a pas d'autre possibilité que de remplacer les valeurs manquantes par une autre modalité si le but est de ne pas supprimer toutes les observations avec des données manquantes. Mais il reste encore des observations manquantes concernant les variables quantitatives, comme les variables de nombre de véhicules notamment. Il existe plusieurs méthodes fréquemment utilisées pour répondre à ce type de problématique :

— Suppression des observations avec des valeurs manquantes : au vu du nombre non négligeable de valeurs non renseignées, cette méthode ne sera pas retenue car cela entraînerait la suppression d'un nombre trop important d'observations

— Remplacement par la valeur moyenne ou médiane : le choix entre ces deux possibilités dépend de la base de donnée et des informations d'expérience du produit, pour être au plus près de la réalité. Dans cette étude, la méthode retenue sera celle de la médiane car elle permet de ne pas avoir une valeur biaisée par des extrêmes.

— Régression selon les autres variables quantitatives renseignées : cette méthode a déjà été appliquée sur les variables de CA et d'effectif suite à l'analyse des corrélations qui a montré un lien important entre celles-ci. Cependant, il est également possible d'appliquer cette méthode.

#### Variables catégorielles

Les variables qualitatives ne peuvent pas être traitées ainsi dans les modèles d'apprentissage statistique : seules les valeurs numériques sont utilisables. Par conséquent, il

est possible de transformer ces variables qualitatives pour pouvoir conserver l'information fournie par ces variables. Cela s'effectue en créant à la place de chaque variable, plusieurs variables indicatrices pour chaque modalité de la variable concernée. Par exemple, sur la variable de situation du risque, les modalités sont à l'origine :

— Agglomération
— Zone Industrielle ou d'Activité (Z.I. ou Z.A)
— Centre commercial
— Zone rurale

Elles deviennent alors après retraitements les variables suivantes :

— Situation_Agglomération : 1 si oui, 0 sinon
— Situation_Zone Industrielle ou d'Activité (Z.I. ou Z.A) : 1 si oui, 0 sinon
— Situation_Centre commercial : 1 si oui, 0 sinon

Et la dernière modalité : Zone rurale correspond alors au cas où les trois nouvelles variables présentées ci-dessus sont égales à 0. Il est également possible de créer une quatrième variable pour représenter cette dernière modalité mais cela crée un nombre important de variables alors que cela n'est pas nécessaire, étant donné que l'information est contenue dans les trois premières nouvelles variables crées en remplacement de la variable initiale de Situation.
De même, et de façon intuitive, les variables ayant une réponse binaire Oui/Non, ou des autres variables n'ayant que deux modalités, n'ont pas à être dupliquées, seule une transformation de la modalité Oui (ou de tout autre modalité le cas échéant) en 1 et la seconde modalité en 0 permettra le traitement de ces informations de manière correcte.
Il est important de noter que les qualitatives ayant auparavant été encodées sous des variables numériques doivent tout de même être traitées comme des variables qualitatives. En effet, le traitement des variables numériques sera ordinal, ainsi s'il n'y a pas d'ordre dans la variable, elle ne doit pas avoir de format numérique, hormis un format binaire comme indiqué ci-dessus.

### 3.5.3 Modélisations et performances

Une fois les données retraitées pour correspondre aux exigences des méthodes d'apprentissage statistique, le modèle de Random Forest peut être implémenté, et l'analyse des performances permettra de juger de la pertinence et de la qualité de modélisation de ce type de méthode.

***Importance des variables*** Les méthodes d'apprentissage statistiques n'attribuent pas de coefficients aux modalités des variables qui pourraient permettre d'analyse le spread de celles-ci, mais il est possible d'obtenir l'ordre d'importance des variables dans le processus d'apprentissage. Ces informations sont ici présentées selon deux critères :

— le pourcentage d'erreur moyenne quadratique (MSE) supplémentaire en cas de suppression de la variable en question : ici, les fait de supprimer la variable de chiffre d'affaire (caht) aboutirait à augmenter l'erreur moyenne quadratique de 2%.

— l'impureté des noeuds mesure la qualité de la discrimination des variables : plus sa valeur est élevée, meilleur est le pouvoir discriminant de la variable.

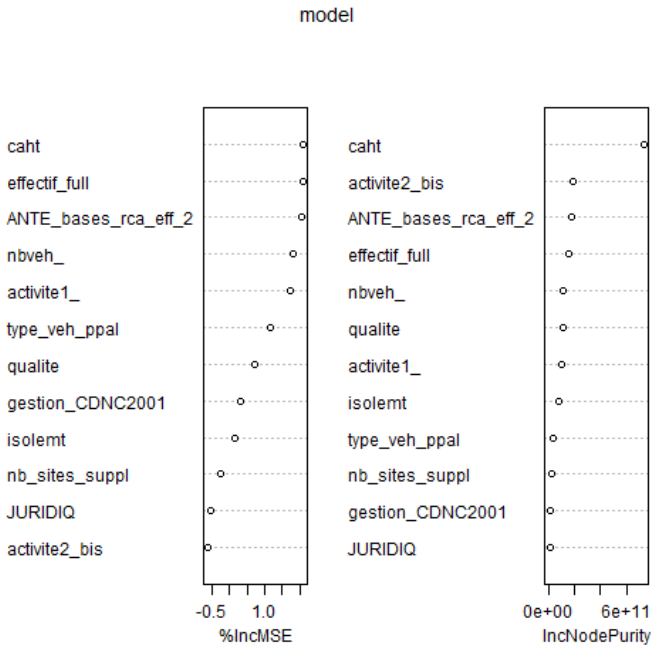



FIGURE 3.10 – Ordre d'importance des variables du modèle de Random Forest - sans contrainte de nombre d'arbre

Cet ordre est défini pour une modélisation sans contrainte du nombre d'arbres pour l'apprentissage, mais il n'est pas forcément nécessaire d'avoir un nombre d'arbres trop important pour l'apprentissage. Le graphique suivant permet de déterminer à partir de quel nombre d'arbre l'erreur moyenne quadratique ne diminue plus significativement.

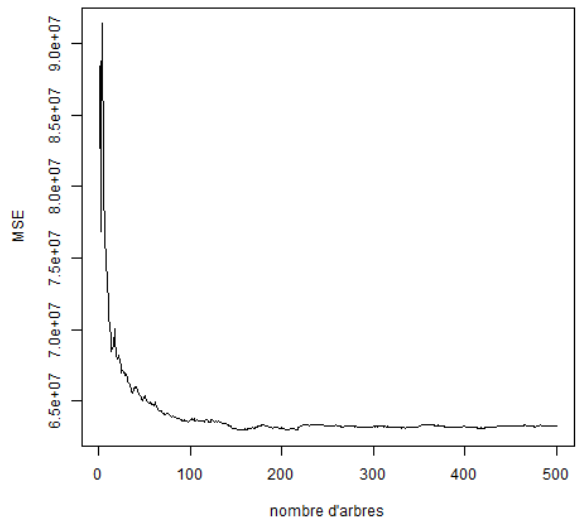



FIGURE 3.11 – Évolution des performances en fonction du nombre d'arbres utilisés

Ce graphique permet alors de considérer que seuls 200 arbres peuvent être nécessaire pour avoir un apprentissage qualitatif sur cette problématique. L'ordre d'importance des variables pour un modèle fixé à 200 arbres est alors le suivant :

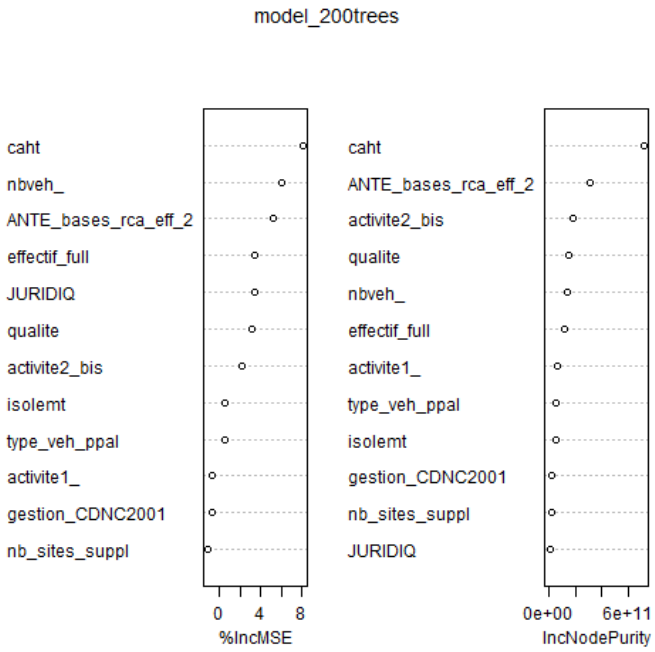


FIGURE 3.12 – Ordre d'importance des variables du modèle de Random Forest une fois le nombre d'arbres fixé

L'ordre d'importance des variables est modifié mais ce n'est qu'un indicateur pour mieux comprendre le fonctionnement de l'apprentissage sur ces données mais le plus important est la mesure des performances du modèle, présentée ci-après en reprenant les critères utilisés pour les méthodes GLM.

### *Lift Curve*

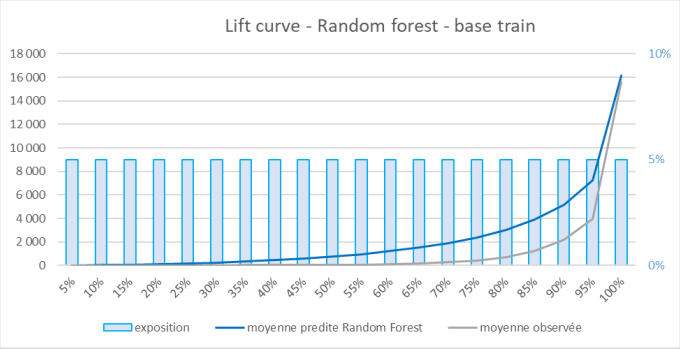



FIGURE 3.13 – Lift Curve pour le modèle de Random Forest

La Lift curve semble montrer une très bonne adéquation sur les prédictions de faibles montants mais se dégrade lorsqu'ils augmentent.

### *Lorenz Curve*

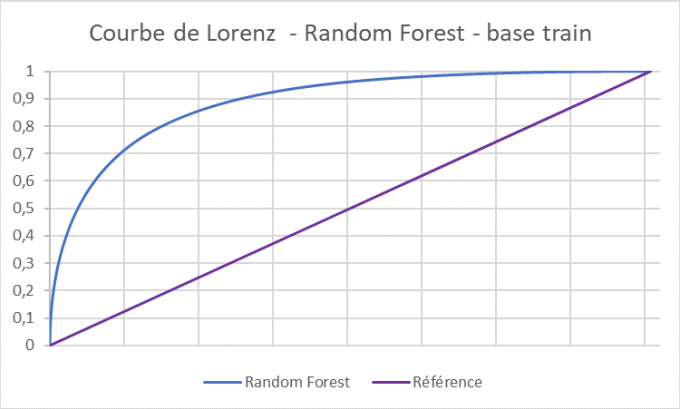



FIGURE 3.14 – Lorenz Curve pour le modèle de Random Forest

En revanche, la courbe de Lorenz montre une très bonne adéquation aux données, avec une courbe qui se rapproche du point $(0, 1)$.

***Résidus agrégés***

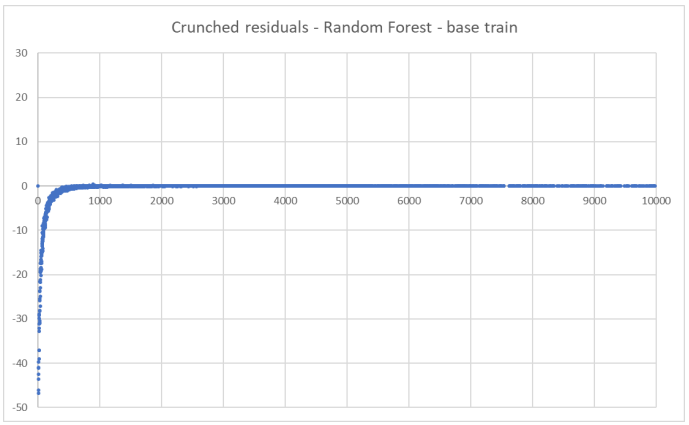



FIGURE 3.15 – Résidus agrégés du modèle de Random Forest

Les résidus nuancent cependant ces conclusions : ils sont moins centrés autour de 0.

***Autres indicateurs statistiques***

| Indicateurs | Random Forest |
|---|---|
| Gini | $77,9\%$ |
| RMSE | 4859 |

TABLE 3.2 – Indicateurs de modèles

Ces indicateurs confirment les bonnes performances sur la base d'apprentissage de ce modèle. Les résultats sur la base de test permettront d'apprécier plus justement la qualité des modèles d'apprentissage statistique et GLM.

# Chapitre 4

# Résultats de la modélisation

## 4.1 Performances et pertinence des modèles retenus sur la base de validation

Les résultats sur la base d'apprentissage présentés plus hauts peuvent être biaisés en cas de sur-apprentissage des données, notamment pour les méthodes d'apprentissage statistique. Les performances sur la base de test permettent ainsi de comparer les performances d'un modèle GLM et d'un modèle d'apprentissage statistique. Le modèle en approche Prime Pure sera le modèle retenu pour la méthode des GLM au vu de sa meilleure qualité d'adéquation présentée sur la base d'apprentissage.

**Comparatif entre le modèle de Prime Pure et le modèle de Random Forest**
Les indicateurs présentés sur la base d'apprentissage seront retenus pour juger des performances sur la base de validation.

***Courbe de Lorenz***

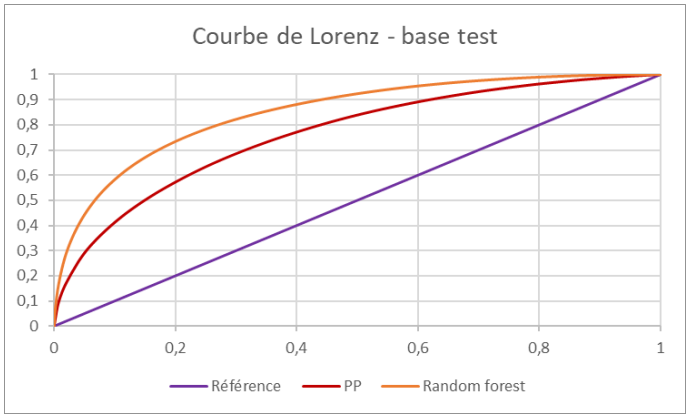



FIGURE 4.1 – Comparatif des courbes de Lorenz

Les conclusions au vu des courbes de Lorenz montrent clairement une meilleure adéquation aux données avec le modèle de Random Forest, l'écart entre les courbes étant non négligeable et au profit du Random Forest, la courbe de ce dernier étant plus proche du point $(0, 1)$ à atteindre pour maximiser les performances.

***Courbes lift***

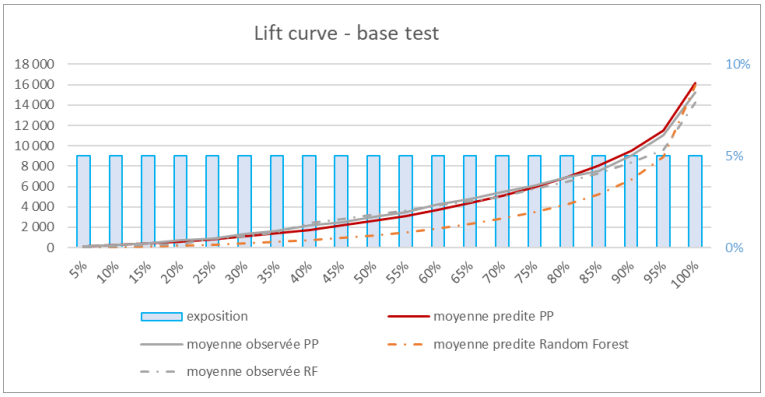



FIGURE 4.2 – Comparatif des courbes Lift

En revanche, les courbes Lift semblent indiquer une meilleure répartition sur le GLM, l'écart entre la courbe des prédictions modélisées et les valeurs observées étant plus faible, révélateur d'une meilleure adéquation à l'observé.

***Résidus***

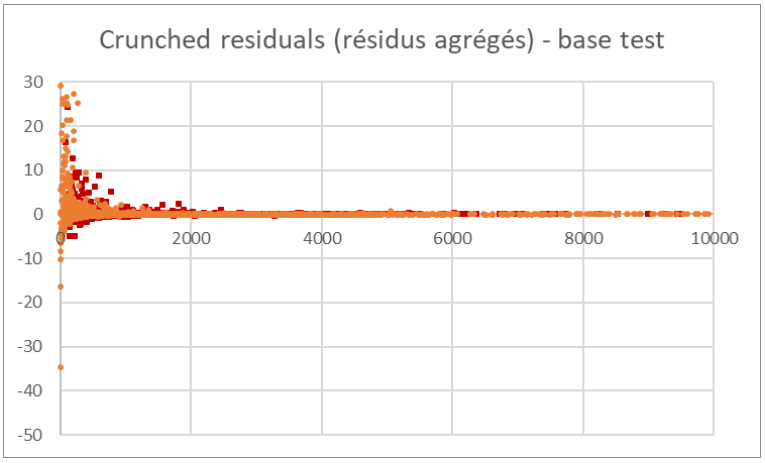



FIGURE 4.3 – Comparatif des résidus agrégés

Les résidus agrégés ne semblent pas départager les deux modèles mais permettent de valider l'hypothèse d'indépendance des résidus dans les deux cas.

***Autres indicateurs statistiques***

| Indicateurs sur la base de test | Modèle en approche Prime Pure | Modèle en d'apprentissage statistique : Random Forest |
|---|---|---|
| Gini | $52,6\%$ | $69,9\%$ |
| RMSE | 5505 | 4766 |

TABLE 4.1 – Tableau comparatif de modèles en approche Prime Pure et Random Forest

Ces indicateurs sont en accord avec les conclusions obtenues grâce aux courbes de Lorenz : le modèle d'apprentissage statistique a de meilleures performances que le GLM, ici présenté en approche Prime Pure. Néanmoins, il est important de noter que le Gini sur la base de test est bien inférieur à celui sur la base d'apprentissage, ce qui peut être révélateur d'un sur-apprentissage, qui n'est pas en faveur de cette modélisation.

Ainsi, pour des raisons opérationnelles, le GLM en approche Prime Pure sera retenu. En effet, les contraintes d'implémentation du tarif ne permettent pas de choisir cette méthode mais cette dernière comparaison aura permis de confronter des méthodes classiques à des méthodes reconnues pour leurs performances. En outre, les résultats du GLM montrent une adéquation très correcte, qui se remarque notamment avec le lift curve. De plus, leur interprétabilité est un principal facteur différenciant qui permet de placer la méthode GLM en tant que méthode la plus utilisée pour la tarification.

## 4.2 Détails sur le modèle retenu

Le but est ici d'entrer plus en détails sur le modèle finalement retenu, à savoir le modèle en approche Prime Pure selon la méthode des GLM. Un zoom par variable permettra de mieux comprendre l'importance de chacune d'elles dans la modélisation. Elles seront présentées par ordre d'importance, les antécédents étant la variable la plus discriminante du modèle.
La valeur relative des coefficients (*Coeffcient value*) est donnée, mais également la valeur de la charge moyenne observée (*Observed Average*) et de la charge moyenne prédite (*Fitted Average*) avec ce modèle pour chaque modalité.

***Antécédents de sinistralité*** Les antécédents de sinistralité présentés ici ont été calculés en prenant en compte différemment les sinistres responsables et non responsables pour que cette dimension ne soit pas négligée dans la modélisation d'une garantie telle que la Responsabilité Civile Automobile. De plus, la modalité "888" permet d'isoler les entreprises en création, qui n'ont donc pas d'antécédents.

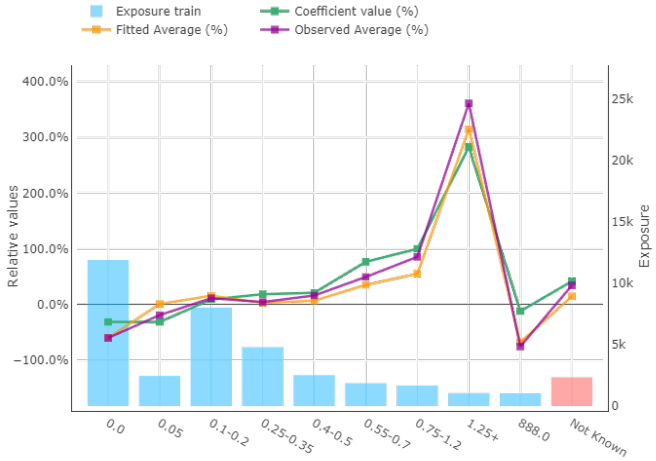



FIGURE 4.4 – Coefficients sur la variable d'antécédents

La tarification actuelle plaçait ce type d'assuré sur le coefficient le plus élevé, par mesure de précaution du fait de l'absence d'information. Cependant, grâce à la prise en compte de la sinistralité réellement observée des clients pour le calcul des antécédents, il est possible d'apprécier leur niveau de risque. Cela permettrait d'avoir un meilleur tarif pour ces entreprises et d'attirer ainsi une clientèle plus large avec des tarifs plus attractifs sur ce segment.
Par ailleurs ces coefficients ne sont pas lissés, mais il est important de noter que, toutes choses égales par ailleurs, la prime pure pour la modalité "$0,75 - 1.2$" est 100% plus élevée que celle pour la modalité "$0.1 - 0.2$", alors qu'il n'y avait pas autant de différence sur les coefficients du tarif actuel. Cela contribue alors à placer cette variable comme la plus discriminante, comme le montre le spread (cf.3.4.2)

***Effectif*** La variable d'effectif était également une variable prépondérante sur la tarification pratiquée jusqu'ici, mais cette modélisation permet de distinguer principalement les profils à faibles effectifs et ceux à très forts effectifs. Ainsi, la volonté d'avoir un tarif plus lissé sur cette variable sur les effectifs centraux est respectée.

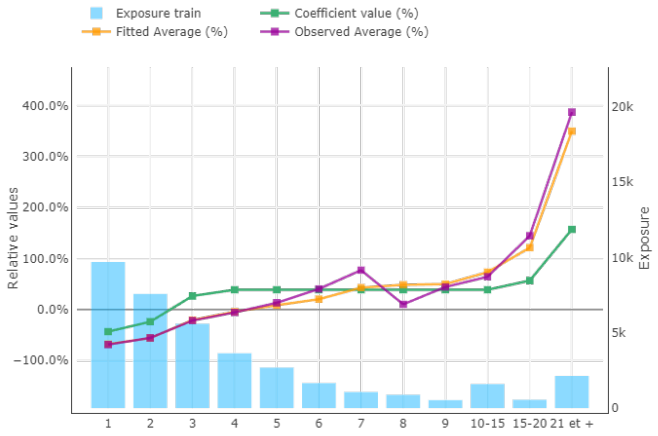



FIGURE 4.5 – Coefficients sur la variable d'effectif

***Nombre de véhicules propriétaires*** Les véhicules propriétaires correspondent aux véhicules qui sont directement immatriculés au nom de l'entreprise et dont le nombre et la typologie du véhicule est renseignée lors de la souscription. Ensuite seul le nombre de véhicule est mis à jour au long de la vie du contrat, c'est pour cela que cette variable a été retenue plutôt que la typologie des véhicules détenus. Par ailleurs, il n'est pas possible à ce jour d'obtenir de données sur les véhicules confiés pour vente ou réparation, mais cette information pourrait se déduire de variables d'effectif ou de chiffre d'affaires, révélatrices des mouvements plus ou moins importants de véhicules.

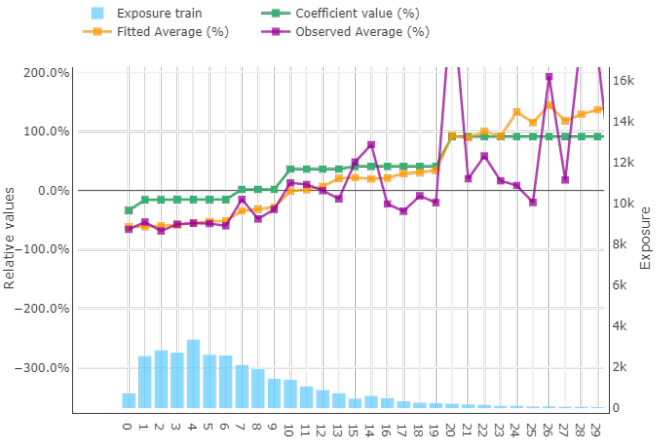


FIGURE 4.6 – Coefficients sur la variable de nombre de véhicules

La modélisation présentée ici permet de distinguer des tranches pour avoir un tarif lissé par rapport à la charge observée qui subit de fortes variations du fait des faibles volumes sur les modalités les plus élevées.

***Chiffre d'affaires annuel*** Le chiffre d'affaire annuel permet, comme l'effectif d'appréhender les potentiels mouvements de véhicules, qui auraient donc un impact sur la garantie Responsabilité Civile Automobile.

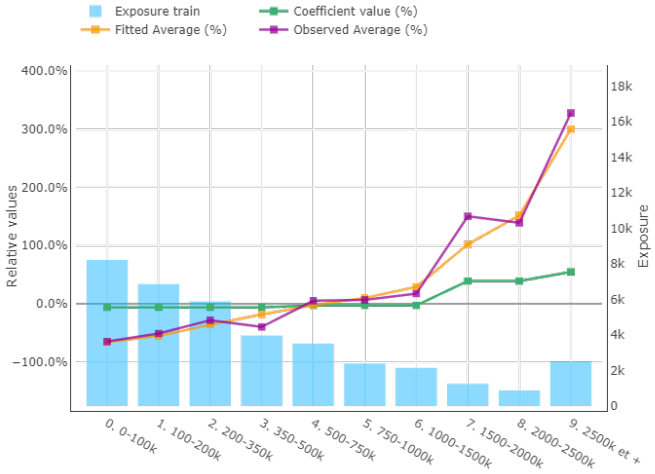



FIGURE 4.7 – Coefficients sur la variable de chiffre d'affaires

Cette donnée permet ici de cibler les chiffres d'affaires importants.

***Forme juridique de l'entreprise*** La forme juridique de l'entreprise se distingue en deux modalités, à savoir *Société* ou *Nom Propre*. Cette variable peut sembler éloignée du risque étudié mais elle permet de cibler des comportements différents du fait de la structure et de l'implication plus ou moins forte du dirigeant dans la santé de l'entreprise.

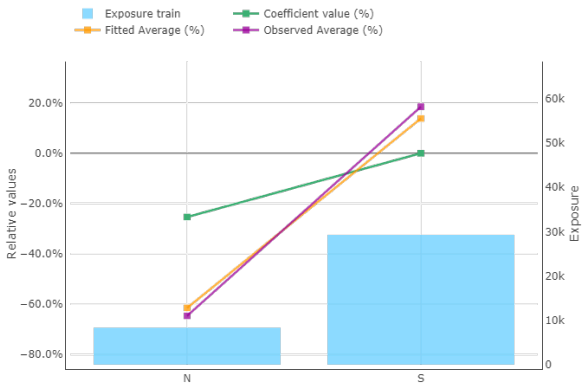


FIGURE 4.8 – Coefficients sur la variable de forme juridique

Ce raisonnement permet alors de mieux comprendre le coefficient attribué aux entreprises en Nom propre dont la responsabilité est plus engagée que les Sociétés.

***Mode de gestion du contrat*** Le mode de gestion du contrat est principalement lié à la quantité d'activité de l'entreprise. En effet, une gestion *Forfaitaire*, c'est-à-dire sans révision du tarif selon des variations de chiffre d'affaires ou d'effectif, est réservée aux entreprises n'ayant pas un chiffre d'affaire trop élevé. Cette variable appuie donc les conclusions obtenues avec les variables de chiffre d'affaires et d'effectif.

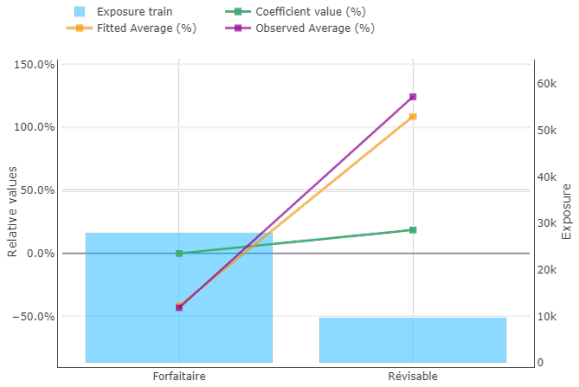



FIGURE 4.9 – Coefficients sur la variable de mode de gestion du contrat

***Zonier modélisé pour la garantie*** Différents zoniers ont été présentés sur ce modèle, et celui retenu est celui avec une contrainte du nombre de zones, qui apparaissait comme le plus pertinent :

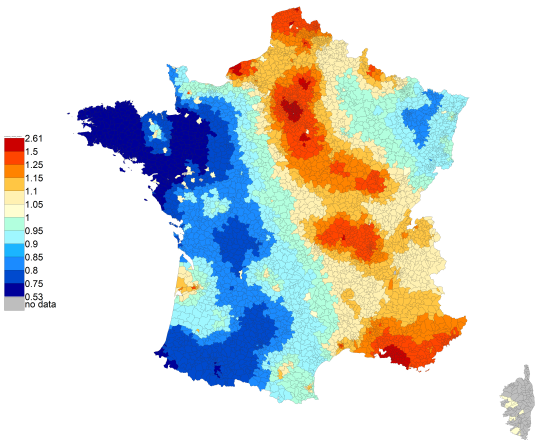



FIGURE 4.10 – Zonier avec une contrainte de nombre de modalités

Le but étant d'apprécier au mieux les variations géographiques du risque, sans pour autant définir un coefficient par ville, ce qui n'irait pas dans le sens de la simplification du tarif. Ainsi, treize zones ont été retenues pour ce zonier, à savoir le même nombre que celui utilisé actuellement, à la principale différence que ces zones sont spécifiques au risque étudié.

## 4.3 Comparaison des primes : entre l'actuel et le modélisé

L'un des objectifs de cette refonte était la simplification du tarif pour passer à une structure multiplicative. Cette structure pourra notamment faciliter la comparaison des tarifs en cas de mise à jour à venir. A ce jour, la comparaison des coefficients de tarification est donc impossible en l'état, la comparaison entre le tarif actuel et le modélisé s'effectue alors par une analyse de la dispersion des primes.
Cette analyse permet de définir un premier impact chiffré des évolutions de la garantie sur le portefeuille, et d'identifier les profils les plus impactés par des hausses ou des baisses de tarifs.

Cependant, il est important de noter que les primes modélisées dans le cadre de ce mémoire n'ont pas subi de lissage significatif. Il ne s'agit donc pas des évolutions qui seront réellement observées lors de la commercialisation du produit après refonte. Les étapes de lissage pour rendre le tarif opérationnel seront effectuées une fois toutes les garanties du produit modélisées. En effet, cette étape nécessite d'avoir une vision globale des évolutions pour obtenir de meilleurs résultats, c'est pour cela qu'elle sera mise en oeuvre lorsque les tarifications des garanties auront été modélisées et agrégées.

**Impact général sur la garantie** La mesure de l'impact de l'évolution de la tarification du produit peut se mesurer en analysant la dispersion des primes en prenant comme comparaison la valeur des primes pures modélisées sur la base retenue et la valeur des primes pures réelles du contrat. Cette dernière est obtenue en prenant la prime commerciale disponible dans les bases contrats en retraitant les chargements correspondants.
Dans un premier temps, une analyse à l'échelle de la garantie pourra être proposée afin d'avoir un aperçu possible de l'impact global, les conclusions ne pouvant être définitives qu'une fois toutes les garanties modélisées et la tarification lissée.
Afin de ne pas avoir de biais lié aux éventuelles majorations annuelles, seules les affaires nouvelles des trois dernières années, parmi les contrats présents dans la base de modélisation, ont été retenus pour cette comparaison.

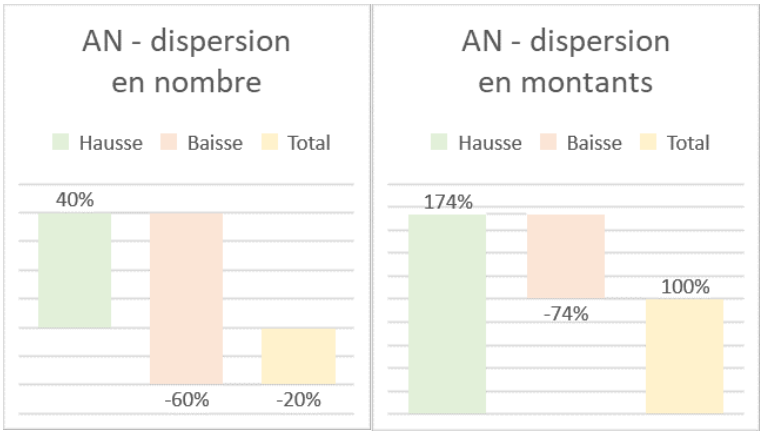


FIGURE 4.11 – Impacts de l'évolution de la tarification RC Automobile

La dispersion permet de montrer qu'en termes de nombre, il y a plus de contrat qui ont des baisses de tarifs : 20% des exemples étudiés ont des tarifs plus faibles avec les primes modélisées. En revanche, la conclusion est inverse si le raisonnement est en termes de montants : l'enveloppe de primes augmenterait de 100%, soit un doublement.

Pour rappel, ces impacts ne seront pas ceux qui seront effectivement observés lors de la commercialisation du nouveau produit, étant donné que les étapes de lissage seront effectuées ultérieurement, à savoir une fois que toutes les garanties auront été re-tarifées. Par conséquent ces résultats pourront être atténués, compensés, voire inversés lorsque ce type d'analyse sera pratiquée à l'échelle du produit.
Néanmoins, il est intéressant d'essayer d'expliquer ces effets pour comprendre les profils potentiellement impactés par ces évolutions marquantes.

**Profils impactés** L'analyse des profils impactés va permettre de mieux cibler l'origine de ces évolutions sur la dispersion de la garantie avant lissage. La modélisation par un arbre de décision sur les différences de primes Responsabilité Civile Automobile entre l'actuel et le modélisé peut fournir quelques pistes sur les profils concernés.

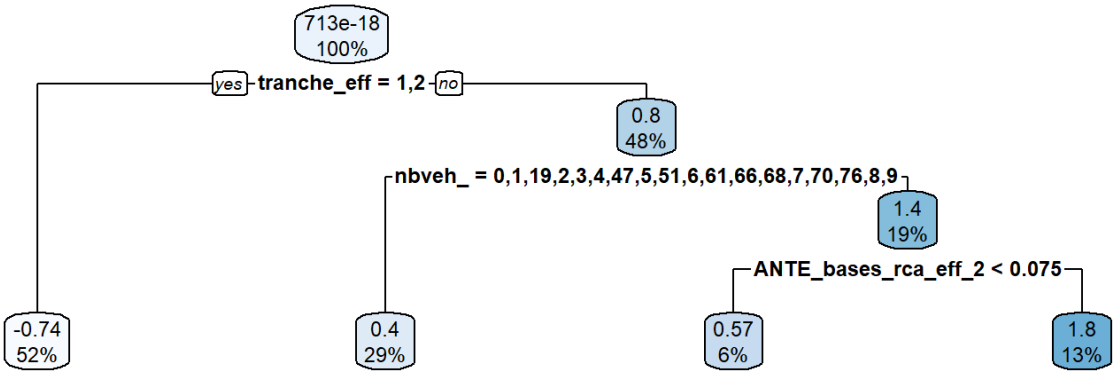



FIGURE 4.12 – Arbre de dispersion des primes RC Automobile entre le tarif actuel et le tarif modélisé

L'arbre de décision permet de montrer que les contrats subissant des hausses de tarif sur cette garantie ont des effectifs élevés et un nombre important de véhicules, cela correspond alors au haut de segment. A l'inverse, les contrats du bas de segment sont plutôt à la baisse.
Ainsi, le portefeuille étant principalement composé de clients de bas de segment, la dispersion des primes en nombre a une tendance à la baisse, expliquée par les volumes du bas de segment (en nombre de contrats) très importants. Cependant, ces profils ont de faibles primes, à l'opposé du haut de segment qui a des primes élevées. Par conséquent,

l'effet de la dispersion en termes de montants peut s'inverser du fait que le haut de segment est celui qui subit une tendance à la hausse.

**Conclusions à nuancer à la maille du produit** Ces résultats montrent des impacts sur la garantie étudiée ici, mais l'effet sur le produit pourra être tout autre une fois que toutes les garanties du produit auront été revues. En effet, plusieurs étapes de la refonte vont entrer en jeu dans la comparaison du tarif avant de pouvoir obtenir des conclusions sur le portefeuille.

— La modélisation de toutes les garanties permettra d'observer de potentiels effets de compensation sur certains profils, comme par exemple ici sur le haut de segment du portefeuille qui est particulièrement impacté à la hausse

— Une fois les garanties modélisées, il sera possible d'avoir une vision d'ensemble et donc de procéder à un lissage du tarif avant commercialisation, afin d'éviter d'avoir des impacts si prononcés à la maille du produit

— Une dernière comparaison pourra également valider les évolutions, une fois que les chargements sur le nouveau produit auront été définis. Cela permettra d'avoir une dispersion entre deux typologies de primes identiques, sans retraitement. En effet, il s'agira d'une comparaison entre deux primes commerciales réelles.

En définitive, il semble que les évolutions tarifaires concerneraient les clients des haut et bas de segment, alors que le segment médian subirait de faibles variations. Les conclusions pourront toutefois être nuancées lorsque les travaux de lissage, nécessaires avant toute commercialisation, seront finalisés. Il sera alors intéressant de comparer les profils impactés dès lors que ces étapes auront pu apporter une cohérence générale au tarif.

# Conclusion

L'objectif de la refonte de la garantie RC Automobile du produit Garages était d'obtenir un tarif simplifié à tous les niveaux pour qu'il soit simple de compréhension pour les clients, mais aussi simple à mettre à jour en fonction des évolutions de rentabilité grâce à une structure tarifaire multiplicative. Cette refonte est intervenue pour redresser les dérives de sinistralité observées sur le produit sur les dernières années tout en entrant dans la volonté de simplification comme fil conducteur de la stratégie d'AXA France.

**Résultats des modélisations** La garantie RC Automobile a été sélectionnée pour entamer ce projet du fait des volumes de primes et de charges sinistres qui étaient majoritaires sur le produit. Par ailleurs, elle a également été retenue pour les spécificités qui lui incombent et qui impactent la modélisation du tarif. En effet, l'une des principales caractéristiques de cette garantie est l'intervention de la convention IRSA et des recours qui impliquent la présence de charges sinistres négatives et forfaitaires. Ce type de charge ne pouvant pas être incluse directement dans les modélisations, plusieurs modèles ont été testés pour essayer de considérer au mieux ces spécificités tout en ayant des modélisations performantes.
Le modèle hors sinistres atypiques, avec mutualisation de la charge forfaitaire et négative au prorata des primes pour les contrats sinistrés est celui qui a été retenu comme le plus adapté pour cette problématique.

**L'importance des données** Ces résultats ont pu être obtenus à la suite de l'analyse et du retraitement des données afin d'avoir des conclusions fiables. En effet, les données constituent la base de bonnes modélisations, et il est important de bien les comprendre et d'apporter les retraitements nécessaires à leur utilisation pour une modélisation de qualité. Le but était alors de rassembler différentes sources de données pour fiabiliser les informations, notamment sur l'activité exercée par l'entreprise, pour obtenir suffisamment de données de qualité. La problématique de la complétion des données manquantes a également été traitée pour essayer d'exclure le moins d'informations possible.

**Limites de l'étude** L'une des principales problématiques des produits en IARD Entreprises est la différence de volumes avec les produits dédiés aux Particuliers. La quantité de données est moins importante alors que le niveau de risque est relativement plus élevé en termes de charge sinistre. Le produit Garages est particulièrement impacté par ce

manque de volumétrie de données, qui est alors à prendre en compte au moment des conclusions faites sur le tarif. Les résultats obtenus pour la garantie RC Automobile de ce produit ne seraient donc pas les mêmes sur cette garantie pour un autre produit automobile. En effet, le nombre de sinistres et la charge associée aux sinistres négatifs et forfaitaires pourraient donner de meilleurs résultats sur une modélisation qui distingue ces charges en plusieurs modèles à agréger ensuite, plutôt que sur une modélisation avec mutualisation comme cela a été retenu pour cette étude. Par ailleurs, les problématiques de volumétrie pourraient également avoir des des effets sur la prise en compte des sinistres graves et atypiques, dont les seuils n'ont pas été revus ici car ce n'était l'objet de l'étude.

**Améliorations futures** Les applications de cette étude étaient prévues pour la création du nouvel outil de souscription, mais ce dernier a été reporté, les résultats vont dans un premier temps permettre de mettre un jour uniquement certains aspects du tarif. Cependant, les travaux de préparation des données et réflexions sur les spécificités du tarif seront conservés pour être utilisés dans un second temps. Cela laisse le temps d'apporter d'autres améliorations grâce à l'inclusion de données qui n'étaient pas disponibles jusqu'ici mais qui le seront à l'avenir. L'une d'entre elles pourrait être l'apport du FVA (Fichier des Véhicules Assurés) qui permettra d'obtenir la liste exhaustive des véhicules et l'usage de ces derniers lorsque cette base de donnée sera totalement fiabilisée. Ces informations auraient pour but d'avoir des données récentes et de qualité sur la typologie et le nombre de véhicules détenus pour amplifier l'importance de cette donnée dans le modèle.

# Table des figures

# Liste des tableaux

# Bibliographie


— Articles
  — BROWNLEE, J. (2020), *Bagging and Random Forest Ensemble Algorithms for Machine Learning.* Machine learning algorithms
  — CHESNEAU C. (2019), *Introduction aux arbres de décision (de type CART).* Université de Caen
  — Pericles Actuarial, *Machine Learning : Du GLM à l'arbre de CART en passant par le Random Forest*
— Cours :
  — RAKOTOMALALA R. (2017) *Analyse des corrélations*
  — CNAM, *Les résidus*, Modèles linéaires généralisés
— EMBLEM User's Guide, *Residuals*, 92-127
— Mémoires :
  — BELLINA R. (2014) *Méthodes d'apprentissage appliquées à la tarification non-vie.* Mémoire d'actuariat, ISFA.
  — SAID S. (2016) *Refonte des modèles de tarification de l'assurance automobile et création de zoniers tarifaires.* Mémoire d'actuariat, Université Paris Dauphine.

# Annexe A

# Présentation AXA France

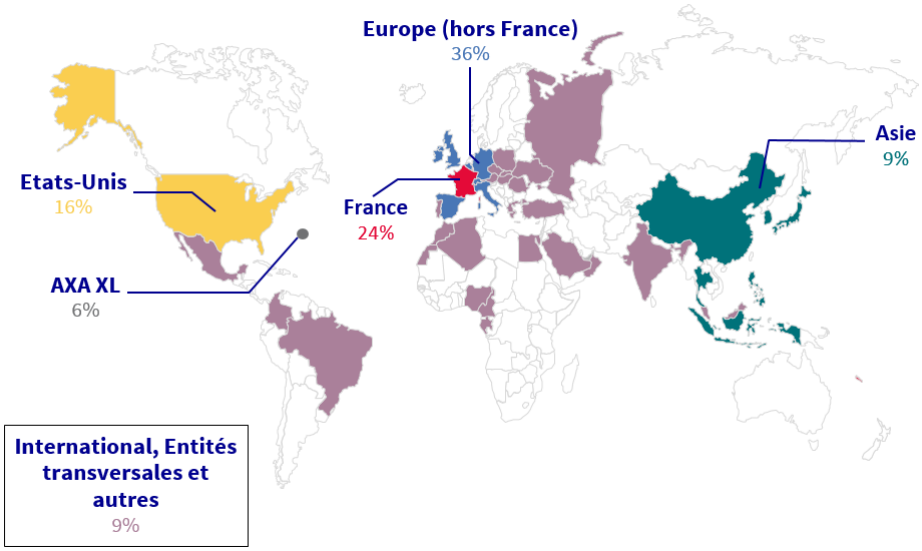
## A.1 AXA France, une société du groupe AXA

### A.1.1 Le groupe AXA

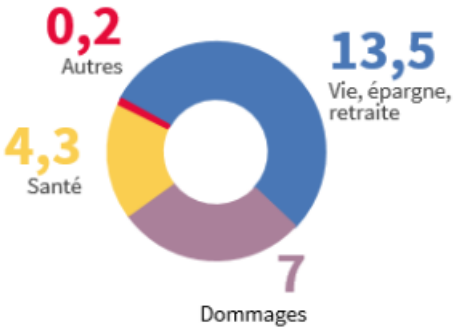
AXA est un acteur majeur français de l'assurance implanté mondialement qui a vu le jour grâce à de nombreuses fusions dont la plus ancienne a eu lieu en 1817. En 1985 a lieu un tournant dans l'histoire du groupe car c'est à cette date que la société prend, sur demande de Claude Bébéar, le nom d'AXA. À partir de 1996, AXA devient l'acteur numéro un mondial de l'assurance grâce à l'acquisition de l'UAP (Union des Assurances de Paris). Depuis cette date, de nombreuses fusions ont permis à AXA de se développer à l'international, on peut citer par exemple la dernière acquisition majeure d'AXA : XL Group, l'un des principaux acteurs de l'assurance dommages des entreprises et de la réassurance, pour la somme de 12,4 milliards d'euros en mars 2018.

### A.1.2 Quelques chiffres

Aujourd'hui AXA est présent dans 61 pays et compte environ 171 000 collaborateurs. Leurs expertises s'expriment à travers une offre de produits et de services adaptés à chaque client dans trois grands domaines d'activité : l'assurance dommages ; l'assurance vie, épargne, retraite et santé ; et la gestion d'actifs. Pour la 10ème année consécutive, AXA est le numéro un de l'assurance dans le monde selon Interbrand. Sa mission est d'apporter à ses clients, qu'ils soient particuliers ou professionnels, une aide et une solution appropriées à leurs différents besoins en produits et services d'assurances, de prévoyance, d'assistance, de banque, d'épargne, et de protection juridique. Pour parvenir à cet objectif, AXA s'appuie sur une véritable orientation client. Le chiffre d'affaires d'AXA en 2018 s'élève à plus de 102 milliards d'euros ce qui représente une hausse de 4% comparé à l'année précédente et est réparti de la manière suivante :




Chiffre d'affaire en France en 2018




## A.2 AXA France

AXA France est le 2ème assureur français derrière le bancassureur Crédit Agricole Assurances avec un chiffre d'affaires d'environ 25 milliards d'euros pour un résultat opérationnel de 1,42 milliards d'euros. Ce classement peut se décomposer avec une 4ème place en assurance de personnes et une 2ème place en termes d'assurance de biens et de responsabilité (très proche du groupe mutualiste Covéa qui le détrône pour un peu plus de 100 millions d'euros). Ci-dessous la répartition du chiffre d'affaires en France en 2018 (en milliards d'euros) :

### A.2.1 La Direction Actuariat et Pilotage Entreprises (DAPE)

Les besoins des entreprises sont divers que ce soit en assurance dommages, responsabilité civile ou risques de spécialité. Il est donc nécessaire de s'adapter aux demandes des entreprises qui sont en perpétuelles évolutions. La Direction Actuariat et Pilotage Entreprises (DAPE) se situe au sein de l'entité AXA Particuliers et IARD Entreprises et est composée de différentes professions : actuaires, data scientists, statisticiens, souscripteurs en réassurance, auditeurs... La DAPE est sous la direction de Thierry Coulloux et est divisée en 3 services :

— Actuariat Comptes, Pilotage Production et Sinistres IARD

— Coordination Conformité Contrôles et Réassurance Facultative

— Actuariat Produits et Data Science

La DAPE contribue au développement du portefeuille IARD Entreprises et à sa rentabilité. Son rôle est essentiel pour assurer le développement d'AXA France sur le secteur de l'assurance d'entreprises qui a ses spécificités par rapport au secteur de l'assurance de particuliers. Au sein de la DAPE, j'ai intégré le service Actuariat Produits et Data Science qui est sous la direction de Véronique Marpillat. Ce pôle est séparé en 2 équipes :

— Risques de fréquence (dirigée par Gérald Lucas) : Automobile, Immeubles, Risques Techniques (RT), Transports Terrestres (TT), Risques Spéciaux Lignes Spécialisées (RSLS), Caution et Data Science Il s'agit de l'équipe dans laquelle j'ai effectué mon alternance.

— Risques d'intensité (dirigée par Loic Chenu) : Responsabilité Civile, Construction, Risques Industriels (RI) et Collectivités Locales

Ce service est effectivement scindé en 2 parties du fait que les méthodes diffèrent selon le type de risque. En effet, le suivi des risques dits « de fréquence » est basé sur un historique court car ce sont des risques ayant une fréquence de sinistres élevée dont le déroulement est rapide et le coût moyen par sinistre modéré, alors que le suivi des risques dits « d'intensité » se base sur une vision de longue durée étant donné que la fréquence des sinistres est plutôt faible due au déroulement très long des sinistres et de leur coût élevé en moyenne. Les missions de ce service sont diverses :

— Développement de la rentabilité

— Refonte et suivi de la tarification

— Processus de Management du Portefeuille : diagnostic, mise à jour des S/C d'équilibre, projection du chiffre d'affaires, revalorisation tarifaire

— Support mensuel pour la direction technique

## A.3 Activités utilisées dans le tarif actuel

| |
|---|
| Concessionnaire |
| Garagiste spécialiste |
| Vendeur de véhicules d'occasion sans réparation |
| Concessionnaire engins de chantier ou agricoles |
| Atelier mobile / Entretien à domicile |
| Campings car, caravanes, vans de tourisme |
| Carrossier peintre jusqu'à 3,5t |
| Carrossier peintre plus de 3,5t |
| dépannage remorquage |
| Garagiste ou agents de marques |
| Garagiste spécialiste Poids lourds |
| Récupérateur / casseur |
| Matériel de chantier ou agricole |
| Sellier bourrelier |
| Speedy |
| station lavage automobile |
| Electricien Auto, station montage auto radio |
| Réparation / Montage de pneumatiques |
| Cycles, motocycles et voiturettes |
| Distribution de carburants avec ou sans réglages et entretien véhicules |
| station service avec boutique |
| Location de boxes / parcs de stationnement |
| Location de véhicules |
| Station de contrôle et de diagnostic Auto |
| centre autosur |
| Centre de contrôle et de diagnostic automobile pour les véhicules de plus de 3,5 tonnes |

# Annexe B

# Notions mathématiques

## B.1 Famille exponentielle

Les distributions de lois de familles exponentielles s'écrivent :

$$f(y|\theta,\phi) = \exp\left(\frac{y\theta - b\theta}{a(\phi)} + c(y,\theta)\right)$$

avec

- $a()$, $b()$ et $c()$ des fonctions
- $\theta$ le paramètre naturel
- $\phi$ le paramètre de nuisance

Les propriétés de ces lois sont : $\mathbb{E}[Y] = b'(\theta)$ et $Var(Y) = b''(\theta).\phi$

## B.2 Lois pour la fréquence

La loi Binomiale Négative correspond à une généralisation de la loi de Poisson La loi de Poisson de paramètre $\lambda \in ]0,+\infty[$ a pour densité :

$$f(y,\lambda) = \frac{\exp(-\lambda)\lambda^y}{y!},\ \forall y \in \mathbb{N},$$

Il est facile de montrer que $\mathbb{E}(Y_i|X_i) = V(Y_i|X_i) = \lambda$, ce qui suppose une homogénéité du portefeuille par rapport au risque, ce qui semble peu réaliste étant donné que les données ont une forte proportion de contrats avec une absence de sinistre, et d'autres avec des valeurs extrêmes. Cela correspond à une sur-dispersion de la variable à modéliser $Y$. Une généralisation de la loi de Poisson peut aussi être considérée : la loi Binomiale Négative (pour plus de détail sur la généralisation : cf **??**). Cette loi peut prendre en compte un paramètre de dispersion pour palier à cette problématique. La densité de cette loi, de paramètres $r \in ]0,+\infty[$ et $p \in ]0,1[$ est donnée par :

$$f(y,r,p) = \frac{\Gamma(r+y)}{\Gamma(r)y!}p^r(1-p)^y,\ \forall y \in \mathbb{N}$$

avec $\mathbb{E}(y|X) = \frac{r(1-p)}{p}$ et $V(y|X) = \frac{r(1-p)}{p^2}$.
Dans ce cas, il n'y a alors plus homogénéité du portefeuille.

## B.3 Lois pour le coût-moyen

Les distributions sont respectivement :

— pour loi Gamma de paramètres $a$ et $b$ :

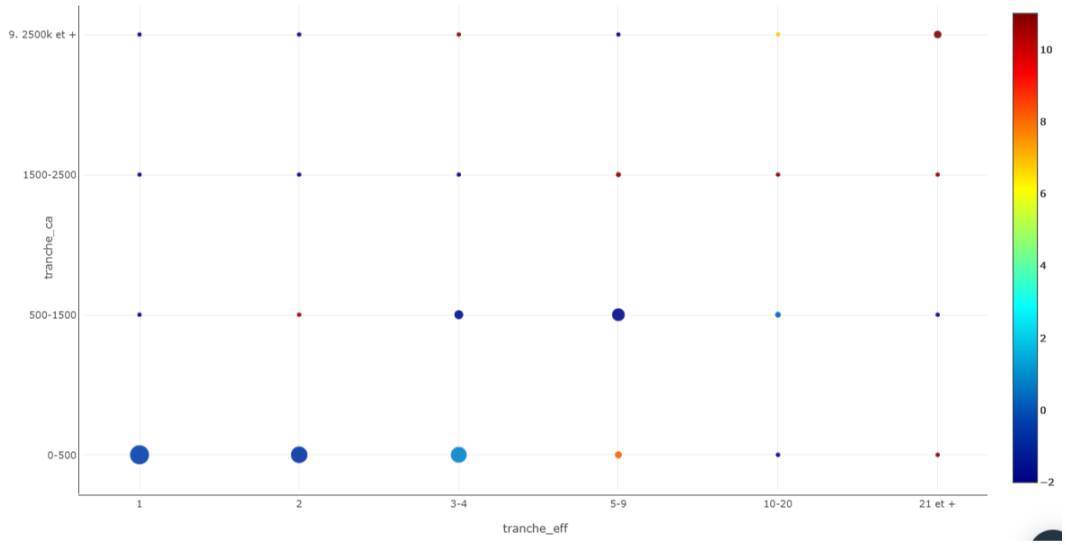
$$f(y, a, b) = \mathbb{1}_{y>0} \frac{b}{\Gamma(a)} (b\,y)^{a-1} \exp(-b\,y)$$

— pour la loi log-Normale de paramètres $\mu$ et $\sigma^2$ :

$$f(y, \mu, \sigma) = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(\ln(y) - \mu)^2\right)$$

## B.4 Interactions entre les variables

Dans le cadre de la modélisation avec la méthode des GLM, l'hypothèse de l'indépendance des variables explicatives entre elles est posée, et validée grâce à l'analyse des corrélations. Cependant, il est possible de choisir de ne pas ignorer la relation entre ces deux variables dans le modèle grâce aux croisements, ou interactions, l'hypothèse d'indépendance des variables n'est alors plus vérifiée. L'ajout d'interactions peut permettre d'enrichir la qualité du modèle mais pour cela il ne faut retenir que la (ou les) interaction(s) la (ou les) plus pertinente(s) et analyser l'effet de cette interaction sur le modèle. Les variables d'interactions viennent en complément des variables concernées.
Suite à l'étude des corrélations dans cette étude, la prochaine étape est l'analyse de l'interaction entre la variable de CA et celle de l'effectif. Celles-ci sont corrélées mais apportent séparément des informations dans le modèle, ce qui pousse à les conserver toutes deux mais en prenant en compte la relation qui existe entre elles. L'ensemble des interactions ont par ailleurs été testées et les interactions les plus pertinentes seront retenues pour cette modèle.
La représentation de ces interactions peut être multiple, par exemple avec une analyse des coefficients de la régression associés à chaque couple de modalités croisées, et ce en prenant en compte l'exposition associée à chacun de ces croisements.




Ces interactions sont représentées en plaçant en abscisses les tranches d'effectif et en ordonnées les tranches de CA pour analyser le comportement au croisement de chaque modalité. La lecture s'effectue alors ainsi :

— la taille du point représente le volume associé à chaque couple de variables

— la couleur du point correspond à la valeur du coefficient modélisé par le GLM

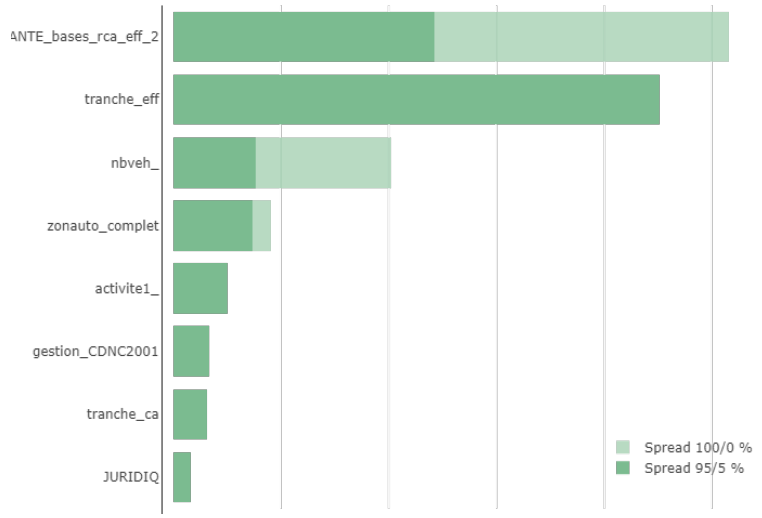
Une linéarité semble se dessiner entre les tranches de CA et d'effectif pour les tranches les plus faibles, cependant les coefficients semblent relativement proches de 0. Le spread (cf. 3.2.2) de cette variable (=120%) est apporté par la relation qui existe sur les tranches les plus faibles de chacune des variables.
Au vu de la faible information supplémentaire apportée par cette interaction, elle ne sera pas retenue dans cette modélisation.

# Annexe C

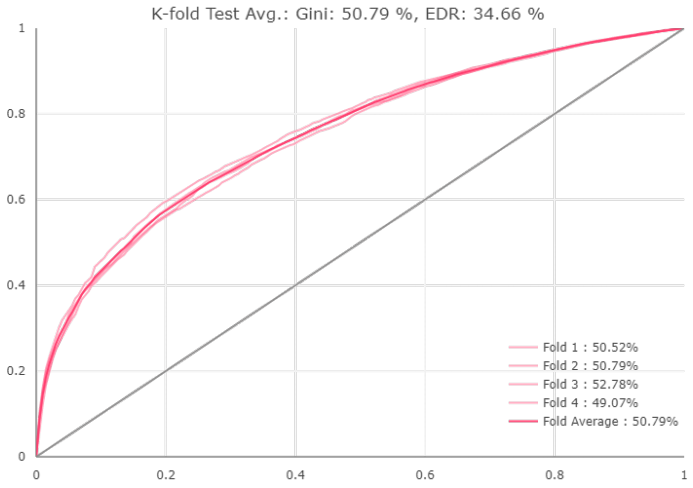
# Performances des modèles de Fréquence et de Coût-Moyen

## C.1 Modèle de Fréquence

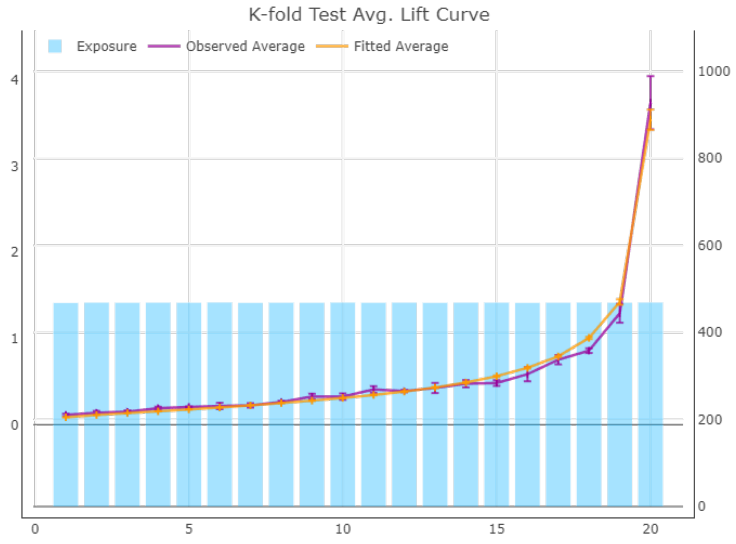
### C.1.1 Variables retenues



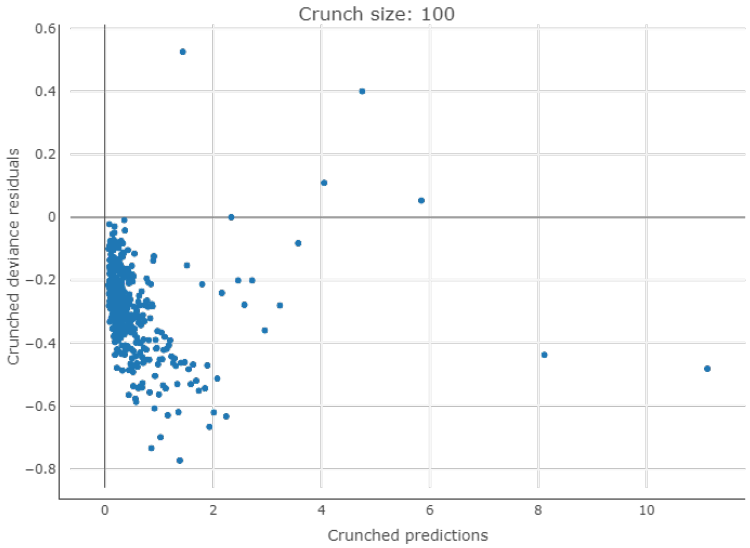
### C.1.2 Courbe de Lorenz




### C.1.3 Courbe lift



### C.1.4 Résidus agrégés

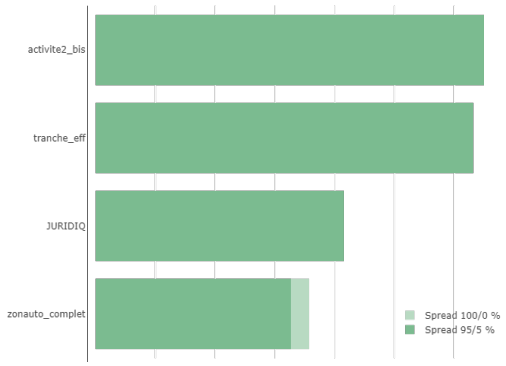



### C.1.5 Indicateurs statistiques

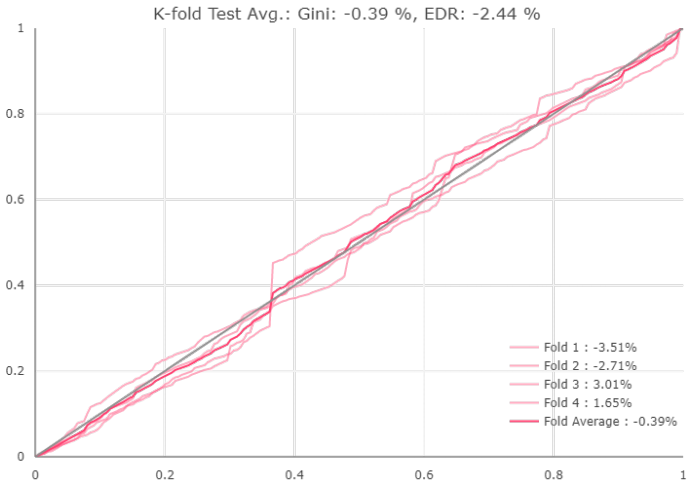
| Indicateurs du modèle de Fréquence | Base d'apprentissage | Base de validation |
|---|---|---|
| Gini | $51,1\%$ | $50,8\%$ |
| RMSE | 1483 | 1492 |

## C.2 Modèle de Coût-Moyen

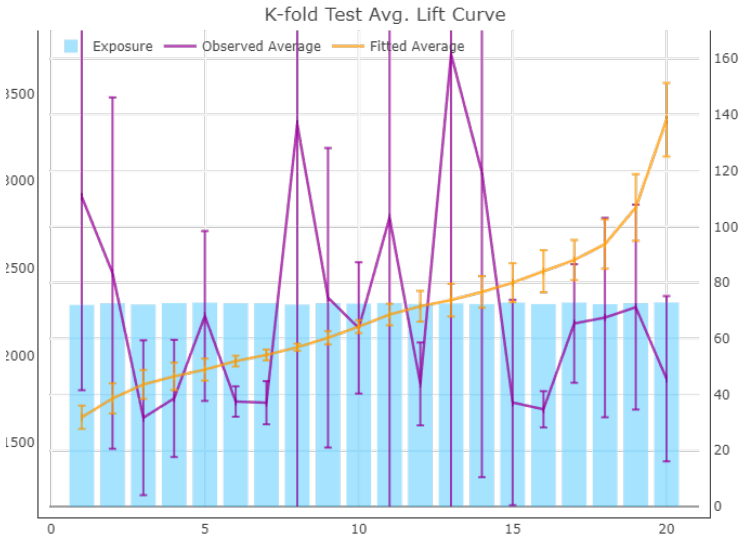
### C.2.1 Variables retenues



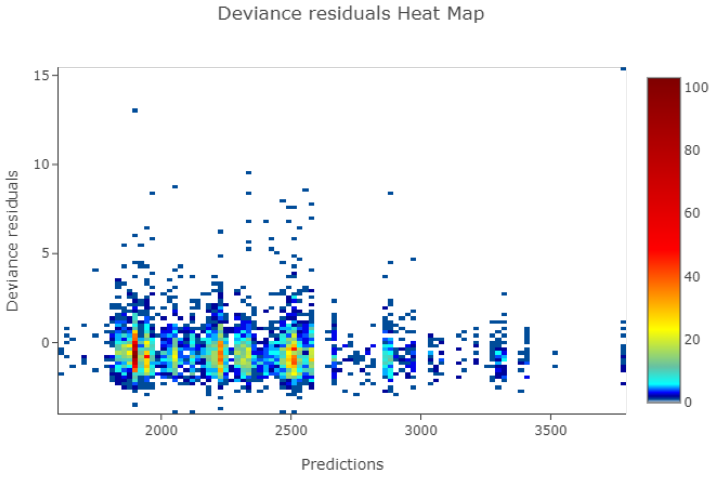
### C.2.2 Courbe de Lorenz




### C.2.3 Courbe lift



### C.2.4 Résidus




### C.2.5 Indicateurs statistiques

| Indicateurs du modèle de Fréquence | Base d'apprentissage | Base de validation |
|---|---|---|
| Gini | $10,6\%$ | $0\%$ |
| RMSE | 8727 | 8232 |

# Annexe D

# Code pour les méthodes d'apprentissage statistique

## D.1 Random Forest

```
library(randomForest)
###Retraitements suppl mentaires pour l'apprentissage statistique
#Selection des variables explicatives possibles
features_rcdrau <- subset(base_rcdrau, select =
    c(outil, reseau, ANTE_bases_rca_eff_2,
    gestion_CDNC2001, JURIDIQ, qualite,
    nbveh_,activite1_, activite2_bis, effectif_full,
    caht, tranche_ca, tranche_eff, isolemt,
    codepostal_rsq, nb_sites_suppl, type_veh_ppal))

##Comptage et traitement des valeurs manquantes
nb_NA <- sapply(features_rcdrau,function(x) sum(is.na(x)))
missmap(features_rcdrau, main = "Missing values vs observed")
sum(is.na(features_rcdrau$nbveh_))
sum(is.na(features_rcdrau$ANTE_bases_rca_eff_2))
#Remplacement par la moyenne
base_rcdrau$nbveh_[is.na(base_rcdrau$nbveh_)] <-
    mean(base_rcdrau$nbveh_,na.rm=T)
base_rcdrau$ANTE_bases_rca_eff_2[is.na(base_rcdrau$ANTE_bases_rca_eff_2)] <-
    mean(base_rcdrau$ANTE_bases_rca_eff_2,na.rm=T)

facteurs <- c('outil', 'reseau', 'gestion_CDNC2001',
    'JURIDIQ', 'qualite', 'activite1_',
    'activite2_bis', 'tranche_ca', 'tranche_eff',
```

```
    'isolemt', 'codepostal_rsq', 'type_veh_ppal')
base_rcdrau[facteurs] <-
    lapply(base_rcdrau[facteurs], function(x) as.factor(x))

#Creation de la base avec les variables explicatives
et la variable    expliquer
base_chg_HG2M_rcdrau <- subset(base_rcdrau,
                        select = c(chg_rcdrau_muRc_p_HG2M_survNvNp1,
                        alea,ANTE_bases_rca_eff_2,
                        gestion_CDNC2001, JURIDIQ,
                        qualite, nbveh_, activite1_,
                        activite2_bis, caht,
                        effectif_full, isolemt,
                        nb_sites_suppl, type_veh_ppal))

training <- base_chg_HG2M_rcdrau[base_chg_HG2M_rcdrau$alea
                                 %in% c(1:3,6:10),]
testing <- base_chg_HG2M_rcdrau[base_chg_HG2M_rcdrau$alea
                                 %in% c(4,5),]
training$alea<-NULL
testing$alea<-NULL

model <- randomForest(data=training,
                      chg_rcdrau_muRc_p_HG2M_survNvNp1 ~ .,
                      importance = TRUE)

plot(model$mse, type = "l", xlab = "nombre d'arbres", ylab = "MSE")

model_200trees <- randomForest(data=training,
                      chg_rcdrau_muRc_p_HG2M_survNvNp1 ~ .,
                      importance = TRUE, ntree=200)
varImpPlot(model_200trees)
```

## D.2 CART

```
library(rpart)
library(rpart.plot)
arbre_disp_ <- rpart(`dif en %` ~ ., data = base_dispersion_)
plotcp(arbre_disp_)
arbre_disp <- rpart(`dif en %` ~ .,
    data = base_dispersion_,
```

```
    control=rpart.control(cp=0.01, maxdepth = 4, minsplit = 500))
rpart.plot(arbre_disp,tweak = 2)
```