Mémoire présenté devant le jury de l'EURIA en vue de l'obtention du Diplôme d'Actuaire EURIA et de l'admission à l'Institut des Actuaires

le 06 Septembre 2023

Par : Audrey LE PERON
Titre : Modélisation des garanties RCM et RCC en Flottes non Dénommées : Tarification par véhicule grâce à l'apport du FVA

Confidentialité : Oui - Durée : 2 ans

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

Membres présents du jury de l'Institut des Actuaires :
Samuel STOCKSIEKER
Romain LAÏLY, Président de Jury
Signature :

Entreprise :
AXA France
Signature :

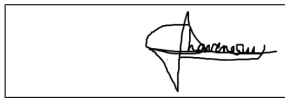
AXA France IARD
Société Anonyme au Capital de 214 799 030 €
Siège social : 313, Terrasses de l'Arche
92727 NANTERRE CEDEX
722 057 460 RCS Nanterre
(Entreprise régie par le Code des Assurances)

Membre présent du jury de l'EURIA :
Catherine RAINER

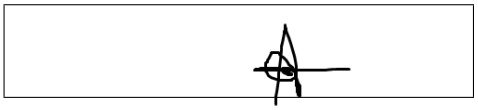
Directeur de mémoire en entreprise :
Lilian CHAVENEAU
Signature :

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 :



Signature du candidat :



EURIA
EURo Institut
d'Actuariat

6, avenue le Gorgeu
CS 93837
29238 Brest Cedex 3

T +33 (0)2 98 01 66 55
euria@univ-brest.fr

Résumé

Un contrat dit Flottes Non Dénommées correspond à la couverture des risques d'une Flotte Automobile (d'un parc automobile) de plus 50 véhicules pour les grandes entreprises ou collectivités locales. La particularité de ces contrats est qu'il n'y a la signature que d'un seul contrat qui décrit les conditions d'octroi des garanties par typologies de véhicules. Il n'y a donc que très peu d'informations sur ces véhicules.

Sur le marché IARD Entreprises d'AXA France, le chiffre d'affaires est équivalent au marché IARD Particuliers/Professionnels (PP) avec un montant qui s'élève en 2021 à 3,5 milliards d'euros contre 3,9 milliards d'euros pour le PP. La branche Automobile et Transport représente 32 % du chiffre d'affaires du marché IARD Entreprises. Dans le cadre de ce mémoire, le sujet portera sur les Flottes Non Dénommées, aussi appelées flottes ouvertes, qui représentent 31 % du chiffre d'affaires de la branche Automobile et Transport.

Avec la mise en application du Fichier des Véhicules Assurés (FVA), une première refonte tarifaire de la garantie Responsabilité Civile (RC) a été effectuée en 2022 afin d'avoir une tarification à la maille véhicule grâce au FVA ainsi qu'aux variables provenant du Système d'Immatriculation des Véhicules (SIV). L'objectif de ce mémoire sera d'ajouter des variables provenant de la base Sécurité et Réparation Automobile (SRA) et sera triple. Dans un premier temps, une modélisation utilisant uniquement les variables SRA sera effectuée. Puis, dans un second temps, une modélisation avec les variables SIV et SRA. Enfin, une modélisation en partant du modèle du tarif 2022 avec ajout des variables SRA sera effectuée afin d'améliorer le modèle déjà existant. Également, une géolocalisation des sinistres sera effectuée afin d'affiner le modèle retenu.

Mots clefs: Flottes ouvertes, IARD Entreprise, Sécurité et Réparation Automobile, Fichier des Véhicules Assurés, Système d'Immatriculation des Véhicules, Responsabilité Civile, Modèles linéaires généralisés

Abstract

On AXA France's P&C compagny market, turnover is equivalent to the Personal/ Professionnal (PP) P&C market, amounting to 3.5 billion euros in 2021 compared with 3.9 billion euros for the PP. The Automobile and Transport branch represents 32% of turnover in the Corporate P&C market. For the purposes of this thesis, the subject will be Unnamed Fleets, also known as open fleets, which account for 31% of the Automobile and Transport branch's turnover.

An Unnamed Fleet policy covers the risks of a vehicle fleet of more than 50 vehicles for large corporations or local authorities. The particularity of these contracts is that there is only one contract to be signed, which describes the conditions for granting cover for each type of vehicle. There is therefore very few information on these vehicles.

With the application of the vehicule register, *"Fichier des Véhicules Assurés"* in french, an overhaul of the Civil Liability (CL) tariff was carried out in 2022 in order to have a tariff at the vehicle level using the FVA and variables from the *"Système d'Immatriculation des Véhicules (SIV)"*. The goal of this thesis will be to add variables from the *Sécurité et Réparation Automobile (SRA)* database, and will be threefold : firstly, to produce a model using only the SRA variables ; secondly, to produce a model using the SIV and SRA variables ; and finally, to produce a model based on the 2022 tariff model with the addition of the SRA variables. Claims will also be geolocated in order to improve the model selected.

Keywords: Vehicle fleets, PC Insurance, FVA, SIV, SRA, Civil Liability, Generalized Linear Model

Note de Synthèse

A. Contexte

Avec un marché de 2,5 milliards d'euros, les flottes automobiles d'entreprise représentent 10,5% de l'ensemble des cotisations automobiles en France en 2021. Afin de répondre à la forte demande, de nombreuses sociétés d'assurance se répartissent ce secteur. Ainsi, le marché des flottes automobiles peut être caractérisé par une forte concurrence, ce qui nécessite que chaque compagnie d'assurances élabore des stratégies tarifaires à divers niveaux pour s'assurer une position privilégiée dans le paysage de l'assurance. Sachant que le marché automobile des particuliers s'avère être très concurrentiel, il est intéressant pour les assureurs de se tourner vers d'autres segments comme le marché des flottes automobiles d'entreprise afin d'augmenter leur rentabilité.

Au sein de la branche des flottes automobiles d'entreprise, deux types de produits s'y distinguent de part leur taille : les flottes fermées, ou Parcs dénommés, et les flottes ouvertes, ou flottes non dénommées. Les flottes fermées disposent d'un nombre de véhicules inférieur à 50 véhicules ainsi que de nombreuses informations tarifaires. À l'inverse des parcs, les flottes non dénommées disposent d'un grand nombre de véhicules, à savoir au moins 50 véhicules.

Avant la mise en place du *Fichier des Véhicules Assurés* (FVA), les informations concernant les véhicules au sein des flottes ouvertes demeuraient inconnues. Seuls le nombre de véhicules par type de véhicule et par garantie étaient renseignés. L'introduction du FVA a eu un double impact. D'une part, il a contribué à lutter contre la non-assurance en France, exigeant aux assureurs la déclaration des véhicules qu'ils couvrent à l'Association pour la Gestion des Informations sur le Risque en Assurance (AGIRA) depuis le 1^{er} janvier 2019. D'autre part, cela a offert à AXA France l'opportunité de recenser l'intégralité de ses véhicules assurés, améliorant ainsi la compréhension des véhicules au sein des flottes ouvertes. Par conséquent, une segmentation plus précise des risques ainsi qu'une tarification plus adéquate de la garantie Responsabilité Civile (RC) ont pu être proposées.

C'est dans ce cadre que s'inscrit le présent mémoire, visant à proposer une nouvelle approche de la modélisation de la prime pure de la garantie Responsabilité Civile (RC), en passant au préalable par la modélisation des garanties Responsabilité Civile Matérielle (RCM) et Responsabilité Civile Corporelle (RCC) au moyen de modèles de fréquence et de coût moyen. Cette nouvelle modélisation est réalisée après une première refonte tarifaire effectuée en 2022 par l'équipe Auto d'AXA France IARD Entreprises, et aura pour objectif de proposer de meilleures performances, grâce à l'ajout de nouvelles variables provenant de l'organisme *Sécurité et Réparation Automobile* (SRA), ainsi qu'en incorporant une variable relative à la distance moyenne des sinistres d'un contrat.

B. Les bases de données utilisées

Dans le cadre de cette proposition de modélisation de la prime pure de la garantie RC, différentes bases de données sont nécessaires. Certaines bases de données sont internes à AXA comme la base contrat, la base véhicule ou encore la base sinistre. Ces bases sont fondamentales pour retrouver les informations telles que le numéro de contrat, le numéro d'immatriculation pour chaque véhicule, les numéros de sinistres avec les informations relatives aux sinistres (garantie concernée, responsabilité, montant du sinistre, date de survenance, etc) mais également des informations liées aux caractéristiques des véhicules.

Comme évoqué précédemment, le fichier des véhicules assurés est une base clé pour les flottes non dénommées, car elle permet de récupérer les informations provenant de la base AGIRA qui contient le numéro de contrat, l'immatriculation des véhicules, les dates d'entrées et de sorties du FVA, la date de première immatriculation ainsi que quelques informations sur le véhicule.

En plus des variables issues de données internes, le numéro d'immatriculation des véhicules a permis la récupération de données externes. Le *Système d'Immatriculation des Véhicules (SIV)* est une base de donnée qui regroupe plusieurs informations liées au certificat d'immatriculation, aussi appelée carte grise. L'Agence Nationale des Titres Sécurisées (ANTS) est en charge de sa mise en oeuvre. Ce système permet d'associer une immatriculation à un véhicule et non à un propriétaire comme cela était fait avant sa mise en place le 15 avril 2019. La base provenant de l'organisme professionnel *Sécurité et Réparation Automobile (SRA)* permet par ailleurs de récupérer des informations liées aux caractéristiques techniques et commerciales des véhicules. Toutes les compagnies d'assurances automobiles sont adhérentes à cet organisme, qui a parmi ces principaux objectifs la diffusion aux adhérents de l'information sur les véhicules.

C. Étude sur la géolocalisation des sinistres / Construction d'une nouvelle variable

Une étude concernant la géolocalisation des sinistres a été réalisée pour approfondir la compréhension de l'activité de l'entreprise. Cette étude visait à déterminer si les véhicules de l'entreprise circulent à échelle locale, départementale, régionale, nationale, voire internationale. L'objectif principal de cette étude était d'évaluer si ces informations pourraient présenter un intérêt dans le contexte de la modélisation des garanties de Responsabilité Civile Matérielle (RCM) et Responsabilité Civile Corporelle (RCC). En d'autres termes, l'étude visait à déterminer si ces informations pourraient contribuer à une tarification et une segmentation plus précise.

Un travail sur la récupération des adresses des souscripteurs ainsi que les adresses des sinistres a été réalisé. Pour les adresses des sinistres, certains sinistres n'avaient que peu d'éléments de renseignés comme le code postal uniquement ou la ville, d'autres avaient l'adresse complète avec le numéro et nom de la rue, le code postal et la ville, et certains sinistres n'avaient aucune informations. Bien que 22,36% des contrats n'aient aucun renseignements sur les lieux de leurs sinistres, cela implique que 77,64% des contrats ont au moins un sinistre avec une information sur le lieu. Après avoir récupéré les coordonnées GPS sur le site `https://adresse.data.gouv.fr/csv` ainsi qu'un site interne à AXA, il a été possible de calculer la distance entre l'adresse du souscripteur et l'adresse des sinistres.

Après avoir calculé la distance pour chaque sinistre, une moyenne des distances des sinistres a été déterminée pour chaque contrat. Cette moyenne permet de définir la distance moyenne à laquelle les véhicules d'un contrat se déplacent (et ont des sinistres) en moyenne. Ainsi, pour chaque contrat, la distance moyenne des sinistres est associée à un indicateur "topage" de 1 si l'entreprise se trouve en Île-de-France, et de 0 dans le cas contraire. Grâce à cette approche, il devient possible de créer une nouvelle variable comportant différentes tranches de distances moyennes : $<$150 km, [150,300 km] et $>$300 km, avec, pour chacune des tranches, un indicateur IDF (Île-de-France) ou bien PROV (Province, hors Île-de-France), ainsi qu'une modalité "International" qui englobe les sinistres à plus de 1000 km. Cette variable nouvellement créée sera nommée "Tranche_dist_moy".

Après étude des tendances pour le coût moyen des sinistres de la garantie RCM et RCC, ainsi que celles de la fréquence des sinistres de la garantie RCC et RCM, il a été décidé de prendre en compte la variable pour la fréquence de sinistres forfaitaires et non forfaitaires RCM, la fréquence de sinistres RCC et le coût moyen des sinistres RCC.

D. Modélisation

Afin de pouvoir modéliser la prime pure de la garantie Responsabilité Civile des flottes non dénommées, il est nécessaire de modéliser d'une part la prime pure de la garantie RCM et d'autre part celle de la garantie RCC. Pour ces deux garanties, le modèle de fréquence $\times$ coût moyen est celui qui a été retenu. La fréquence est modélisée à partir d'une loi de Poisson, et le coût moyen à partir d'une loi Gamma. Ainsi, les différentes modélisations qui ont été réalisées sont les suivantes :

- Modélisation du coût moyen des sinistres de la garantie RCM
- Modélisation du coût moyen des sinistres de la garantie RCC
- Modélisation de la fréquence des sinistres de la garantie RCC
- Modélisation de la fréquence des sinistres forfaitaires de la garantie RCM
- Modélisation de la fréquence des sinistres non forfaitaires de la garantie RCM

Pour toutes ces modélisations, trois approches ont été menées : une approche dans laquelle la modélisation prend en compte les variables provenant de SRA et du SIV, une autre qui prend en compte uniquement les variables SRA et enfin une autre qui reprend le tarif de 2022, avec les mêmes coefficients pour les variables tarifaires, avec ajout des variables SRA. Pour chaque approche, les variables provenant de la base contrat et de la base véhicule sont également pris en compte ainsi que la variable nouvellement créée sur la distance moyenne des sinistres.

Après avoir effectué les différentes analyses et les différents traitements antérieurs à la modélisation, comme la discrétisation ou le regroupement des variables, l'étude des valeurs extrêmes, la corrélation des variables ainsi que la sélection des variables, la modélisation de la prime pure des garanties RCM et RCC a été réalisée ce qui a ainsi permis la modélisation de la prime prime RC.

E. Résultats et conclusion

Finalement, l'approche de modélisation n'utilisant que les variables SRA n'a été retenue pour aucun modèle. Celle utilisant les variables SIV et SRA a été retenu pour le coût moyen des sinistres de la garantie RCM, mais pour les autres modèles, c'est l'approche partant du tarif 2022 avec ajout des variables SRA qui a été retenu. Il est également à noter que pour cette approche, la variable qui en ressort est la nouvelle variable sur la distance moyenne.

Malgré un indice de Gini plus faible que le tarif 2022, la nouvelle modélisation de la prime pure de la garantie Responsabilité Civile reste plus proche de la prime pure observée que la prime pure du tarif 2022. Ceci a pu se confirmer grâce à la double lift curve ainsi que sur l'analyse des variables tarifaires. Finalement, une légère amélioration est observée sur cette nouvelle modélisation. La variable sur la distance moyenne des

sinistres est une variable qui ressort des modèles, ce qui montre que la localisation des sinistres a une importance dans le modèle et que cet axe serait à développer pour le futur. Après comparaison avec le tarif du Parc, appelé E7, il a été remarqué que la nouvelle tarification est en dessous du tarif E7 mais a globalement la même tendance que la prime pure observée. De plus, concernant la variable sur la distance moyenne des sinistres, il a été remarqué que la prime pure observée varie entre chaque tranche, ce qui indique qu'elle pourrait être intéressante pour la modélisation du Parc.

Le modèle reste tout de même assez performant mais nécessite quelques axes d'améliorations. En effet, étant donné que le taux de manquants reste encore assez important, il serait intéressant de le réduire avec la construction d'un modèle qui prédit les manquants à partir des autres variables. De plus, au vu de l'impact de la variable sur la distance des sinistres, il paraît pertinent d'améliorer sa qualité en explorant de nouvelles méthodes comme le Mext Mining sur les rapports de sinistres afin de récupérer de nouvelles informations. Il serait également possible d'ajouter de nouvelles variables externes provenant de bases de l'État ou bien de l'INSEE. Le contrôle de qualité des données pour la jointure avec le FVA, le SIV ou bien avec SRA doit encore être amélioré dans le but de pouvoir avoir des données plus qualitatives et améliorer les modèles.

Synthesis note

A. Context

With a market worth 2.5 billion euros, company car fleets will account for represent 10.5% of all motor insurance contributions in France in 2021. In order to meet demand, a number of insurance companies are sharing out this sector. As a result, the car fleet market can be characterised by strong competition, which requires each insurance company to develop pricing strategies at various levels to ensure to secure a privileged position in the insurance landscape. Knowing that the private car market is proving to be highly competitive, it is interesting-to turn to other segments, such as the corporate fleet market, in order to increase their profitability.

Within the corporate fleet sector, there are two types of product distinguished by their size: closed fleets, or named fleets, and open fleets, or unnamed fleets. open fleets. Closed fleets have a number of vehicles- less than less than 50 vehicles and have a wealth of pricing information. In contrast unnamed fleets have a large number of vehicles, i.e. at least 50 vehicles. Before the introduction of the Fichier des Véhicules Assurés (FVA), the information about the vehicles in open fleets was unknown.

Prior to the introduction of the "Fichier des Véhicules Assurés" ("FVA"), the information about vehicles in open fleets was unknown. Only the number of vehicles per type of vehicle and by cover were given. The introduction of the FVA has had a twofold impact: on the one hand, it has helped to combat non-insurance in France, requiring insurers to declare the vehicles they cover to the "Association pour la Gestion des Informations sur le Risque en Assurance" ("AGIRA") from 1 January 2019. On the other hand, this has given AXA France the opportunity to identify all of its insured vehicles, thereby improving the understanding of vehicles within open fleets. As a result, a more precise segmentation of risks and more appropriate pricing of Third Party Liability (TPL) cover have been proposed.

It is in this context that this report is written, with the aim of proposing a new approach to modelling the pure premium for Civil Liability (CL) cover, by first modelling Material Liability (ML) and Personal Liability (PL) cover using frequency and average cost models. This new modelling follows on from an initial tariff overhaul carried out in 2022 by the AXA France IARD Entreprises Motor team, and will aim to offer better performance by adding new variables from the "Sécurité et Réparation Automobile" ("SRA") organisation, as well as by incorporating a variable relating to the average distance of claims for a policy.

B. The databases used

As part of this proposal to model the pure premium for liability cover, various databases are required. Some databases are internal to AXA, such as the contract database, the vehicle database and the claims database. These databases are fundamental for retrieving information such as the contract number, the registration number for each vehicle, the claims numbers with information relating to the claims (cover concerned, liability, amount of the claim, date of occurrence, etc.) but also information relating to the characteristics of the vehicles.

As mentioned above, the file of insured vehicles is a key database for unnamed fleets, as it can be used to retrieve information from the AGIRA database, which contains the contract number, vehicle registration, FVA entry and exit dates, the date of first registration and some information about the vehicle.

In addition to variables derived from internal data, the vehicle registration number was used to retrieve external data. The *Vehicle Registration System*, also name in french *"Système d'Immatriculation des Véhicules"* ("SIV"), is a database containing information relating to the vehicle registration certificate, also known as the "carte grise" in France. The "Agence Nationale des Titres Sécurisées" ("ANTS") is responsible for its implementation. This system makes it possible to associate a registration with a vehicle and not with an owner, as was the case before it was introduced on 15 April 2019. The database provided by the professional body *Sécurité et Réparation Automobile (SRA)* also makes it possible to retrieve information relating to the technical and commercial characteristics of vehicles. All motor insurance companies are members of this organisation, whose main objectives include the dissemination of vehicle information to its members.

C. Study on the geolocation of claims / Construction of a new variable

A study into the geolocation of claims was carried out to gain a better understanding of the company's business. The aim of the study was to determine whether the company's vehicles travel locally, regionally, nationally or even internationally. The main objective of this study was to assess whether this information could be of interest in the context of modelling material liability and Bodily Injury Liability cover. In other words, the study aimed to determine whether this information could contribute to more accurate pricing and segmentation.

Work was carried out on recovering the addresses of policyholders and the addresses of claims. For the addresses of claims, some claims had only a few details such as the postcode or the town, others had the full address with the number and name of the street, the postcode and the town, and some claims had no information at all. Although 22.36% of contracts had no information on the location of their claims, this means that 77.64% of contracts had at least one claim with location information. After retrieving the GPS coordinates from the `https://adresse.data.gouv.fr/csv` site and an internal AXA site, it was possible to calculate the distance between the policyholder's address and the address of the claims.

After calculating the distance for each claim, an average of the claim distances was determined for each contract. This average makes it possible to define the average distance at which the vehicles in a contract travel (and have claims) on average. Thus, for each contract, the average distance of claims is associated with a 'topage' indicator of 1 if the company is located in the Île-de-France region, and 0 if it is not. Thanks to this approach, it is possible to create a new variable with different average distance brackets: $<$150 km, [150,300 km] and $>$300 km, each with an IDF (Île-de-France) or PROV (Province, excluding Île-de-France) indicator, as well as an "International" modality which includes claims more than 1000 km away. This newly created variable will be called "Tranche_dist_moy".

After studying the trends for the average cost of claims for Material Liability and Bodily Injury Liability cover, as well as those for the frequency of claims for Bodily Injury Liability and Material Liability cover, it was decided to take into account the variable for the frequency of lump-sum and non-lump-sum Material Liability claims, the frequency of Bodily Injury Liability claims and the average cost of RCC claims.

D. Modelling

In order to model the pure premium for Third Party Liability cover for unnamed fleets, it is necessary to model the pure premium for Material Liability cover and that for Bodily Injury Liability cover. For these two types of cover, the frequency × average cost model was chosen. The frequency is modelled using a Poisson distribution, and the average cost using a Gamma distribution. The various models used are as follows:

- Modelling the average cost of Material Liability claims
- Modelling the average cost of Bodily Injury Liability claims
- Modeling the frequency of Bodily Injury Liability claims
- Modeling the frequency of non-lump-sum claims under Material Liability cover
- Modeling the frequency of lum-sum claims under Material Liability cover

For all these models, three approaches have been used: one where the model takes into account the variables from the SRA and the SIV, another which takes into account only the SRA variables and finally another which uses the 2022 tariff, with the same coefficients for the tariff variables, with the addition of the SRA variables. For each approach, the contract and vehicle variables are also taken into account, as well as the newly created variable on the average distance of claims.

After carrying out the various analyses and treatments prior to modelling, such as discretisation or grouping of variables, study of extreme values, correlation of variables and selection of variables, the modelling of the pure premium for RCM and RCC cover was carried out.

E. Results and conclusions

Finally, the modelling approach using only the SRA variables was not adopted for any model. The approach using the SIV and SRA variables was retained for the average cost of claims for RCM cover, but for the other models, the approach using the 2022 tariff with the addition of the SRA variables was retained. It should also be noted that for this approach, the variable that emerges is the new average distance variable.

Despite a lower Gini index than the 2022 tariff, the new modelling of the pure premium for Third Party Liability cover remains closer to the observed pure premium than the pure premium of the 2022 tariff. This was confirmed by the double lift curve and the analysis of tariff variables. Finally, a slight improvement is observed in this new model. The variable on the average distance of claims is a variable that stands out in the models, which shows that the location of claims is important in the model and that this axis should be developed in the future. After comparison with the Park tariff, known as E7, it was noted that the new tariff is below the E7 tariff but has the same overall trend as the pure premium observed. In addition, for the variable on the average distance of

claims, it was noted that the observed pure premium varies between each bracket, which indicates that it could be of interest for modelling the Park.

Even so, the model still performs fairly well, but there are a few areas for improvement. Indeed, given that the rate of missing claims is still fairly high, it would be interesting to reduce it by building a model that predicts missing claims on the basis of the other variables. In addition, given the impact of the variable on the distance between claims, it seems appropriate to improve its quality by exploring new methods such as text mining on claims reports in order to retrieve new information. It would also be possible to add new external variables from government or INSEE databases. Data quality control for the join with the FVA, SIV or SRA still needs to be improved in order to obtain more qualitative data and improve the models.

Remerciements

En premier lieu, je tiens à remercier Véronique Marpillat, responsables des équipes Actuariat Produits et Data Science (APDS) au sein de la Direction Actuariat et Pilotage Entreprise (DAPE), pour son accueil, sa confiance et sa disponibilité qui m'ont permis de réaliser ce mémoire dans les meilleures conditions.

Je remercie également Lilian Chaveneau, manager de l'équipe Actuariat Produits Auto et Transport (APAT), ainsi que mon tuteur d'entreprise Khalil Ouhabi pour leurs conseils et leurs disponibilités pour m'accompagner au mieux durant toute cette expérience professionnelle.

Je remercie également l'ensemble de l'équipe APDS pour avoir fait de cette alternance une expérience professionnelle très enrichissante. Notamment, je tiens à remercier Sami BECK et Lucas JULLIA.

Je n'oublie pas de remercier l'ensemble des professeurs et intervenants qui nous ont transmis leurs savoirs au sein de l'EURo Institut d'Actuariat ainsi qu'à l'Université Gustave Eiffel. En particulier, Thierry Jeantheau, Chi Tran ainsi que mon tuteur de mémoire Ronan Buros.

Pour terminer, je remercie toutes les personnes qui m'ont aidé de près ou de loin lors de ce mémoire, notamment mes parents, mes amis, et Quentin.

Table des matières

Résumé **i**

Abstract **ii**

Note de Synthèse **iii**

Synthesis note **viii**

Remerciements **xiii**

Introduction **1**

1 Mise en contexte et objectifs **3**
1.1 Descriptif de l'assurance des flottes automobiles 3
1.2 Le produit Flottes Non Dénommées 5
1.3 Présentation et objectifs de l'étude 7

2 Le Fichier des Véhicules Assurés **8**
2.1 La mise en application du FVA 8
2.2 Son apport pour les flottes non dénommées 10

3 Les données du portefeuille **12**
3.1 Les bases de données utilisées 12
3.1.1 Bases de données interne à AXA 12
3.1.2 Base de données externe 13
3.2 Traitement/Construction des bases de données 15
3.3 Quelques statistiques descriptives 19

4 Retraitement et analyse des données **22**
4.1 Le traitement des valeurs manquantes 22
4.2 Retraitement des variables 23
4.3 Présentation de la base finale 28
4.4 Études des variables 29

5 Étude sur la géolocalisation des sinistres **40**
5.1 Contexte et objectifs 40
5.2 Étude sur la distance lieu de sinistres / adresse souscripteur 40
5.3 Résultats et conclusion 43

6 Théorie - Modélisation de la prime pure **46**
6.1 Les modèles de tarification 46
6.2 Les modèles linéaires 48
6.3 Les modèles de Machine Learning 52
6.4 Les critères de sélection des modèles et méthode de sélection des variables tarifaires 54

7 Application et résultats des modèles **61**
7.1 Présentation du logiciel AKUR8 61
7.2 Tarification : résultats et comparaison des modèles obtenus 62

Conclusion **83**

Annexes **86**

Bibliographie **94**

Introduction

Le marché IARD (Incendies, Accidents et Risques Divers) Entreprises est un marché qui se développe de plus en plus au fil des années, se rapprochant fortement du marché Particuliers / Professionnels. Avec l'obligation d'assurance en Responsabilité Civile, en 2021, 56,8 millions de véhicules sont en circulation en France dont 8,5% sont des flottes d'entreprises. En terme de cotisations, on obtient un total de 24 milliards d'euros en assurance automobile, dont 2,5 milliards d'euros sur les flottes d'entreprises, soit 10,5% en 2021 [1]. Le marché automobile des particuliers est un marché qui s'avère très concurrentiel ainsi les compagnies d'assurances sont amenées à se tourner vers d'autres segments comme les flottes entreprises afin d'augmenter leur rentabilité.

Sur le segment des flottes automobiles, on peut y retrouver deux types de produits qui se distinguent de part leur taille : les flottes fermées (ou Parcs dénommés) et les flottes ouvertes (ou Flottes non dénommées). Une particularité importante entre ces deux flottes est l'information pour chaque véhicule, qui est historiquement connue pour les flottes fermées étant donné qu'elles sont composées de moins de 50 véhicules, à l'inverse des flottes ouvertes, où le nombre de véhicule est supérieur à 50 et où les informations des véhicules ne sont pas connues. Ainsi, le souscripteur ne renseigne pas les véhicules mais seulement un nombre de véhicule par type par garantie.

Afin de lutter contre la non-assurance, de faciliter l'identification des véhicules en cas de délit de fuite, de vol ou bien encore de trafic, le gouvernement a procédé à la création d'un Fichier national des Véhicules Assurés (FVA). Depuis le 1er janvier 2019, les compagnies d'assurances ont l'obligation de déclarer les véhicules qu'elles vont assurer à l'Association pour la Gestion des Informations sur le Risque en Assurance (AGIRA) dans un délai de 72 heures après la souscription. L'AGIRA a la charge de traiter les flux perçus et de constitution et d'alimenter le FVA dans un délai de traitement maximum de 6 heures. Les véhicules concernés par le FVA sont les voitures, deux-roues, camions, remorques et les véhicules agricoles. Il regroupe les informations sur le véhicule tel que l'immatriculation, le pays d'immatriculation, le nom de l'assureur, le numéro de contrat d'assurance auto ainsi que la durée du contrat.

1. https://www.franceassureurs.fr/wp-content/uploads/2022/09/donnees-cles-2021_juin22_webinteractif.pdf

Ainsi, la mise en place du FVA est un véritable avantage pour le segment des flottes d'entreprises non dénommées. En effet, la connaissance du numéro d'immatriculation permet de prendre en compte des caractéristiques techniques du véhicule qui peuvent être utile lors de la tarification du produit. Cela permettra ainsi d'avoir une tarification comprenant plus d'informations et ainsi avoir une meilleure rentabilité ainsi qu'un tarif plus attractif et compétitif. Le client étant une entreprise, il cherchera à avoir le tarif le plus bas pour sa prime d'assurance couvrant sa flotte, que ce soit à l'affaire nouvelle et d'autant plus au renouvellement où il n'hésitera pas à faire intervenir la concurrence.

Dans ce cadre ci, une refonte tarifaire du produit a été faite en 2022 par les équipes Actuariat Produits Auto et Transports (APAT). Depuis, de nouvelles variables provenant de la base Sécurité et Réparation Automobile (SRA) ont été adressés au service, il serait donc intéressant de regarder si celles-ci pourraient avoir un impact sur la tarification des flottes ouvertes. Également, une étude sur la géolocalisation des sinistres par rapport à l'adresse des clients pourrait être utile afin d'avoir une information de plus sur l'activité des entreprises.

Afin de rendre compte de ces problématiques, une mise en contexte du sujet sera présentée ainsi que ses objectifs avec l'apport de la mise en place du Fichier des Véhicules Assurés. Ensuite, les données utilisées afin de pouvoir créer le modèle de tarification seront explicitées. Par la suite, les étapes de retraitement des données ainsi que leurs analyses vont être précisées. L'étude sur la géolocalisation des sinistres ainsi que les résultats de cette étude seront présentés également. Pour finir, avant de décrire les différents modèles ainsi que leur résultats, une partie théorique sur la modélisation de la prime pure sera présentée.

Chapitre 1

Mise en contexte et objectifs de l'étude

1.1 Descriptif de l'assurance des flottes automobiles

Le marché de l'assurance automobile est le premier marché de l'assurance de biens et de responsabilité avec un montant de cotisations de 24,1 milliards d'euros ce qui représente 38% de ce secteur en 2021.

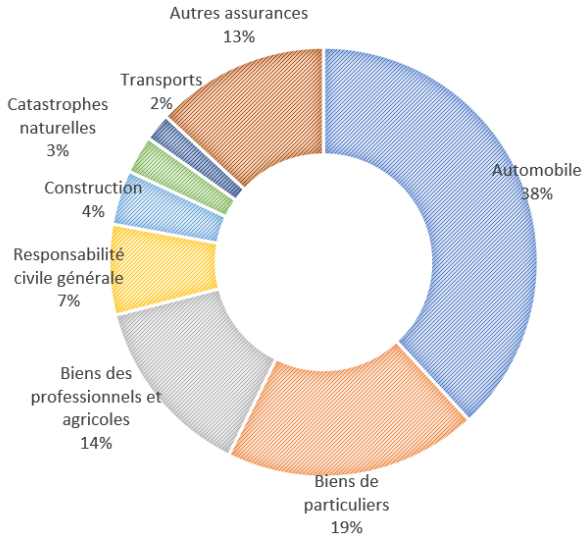


FIGURE 1.1 – Répartition des cotisations dans l'assurance de biens et de responsabilité. Source : Données clés 2021 FFA

Cependant, le marché de l'automobile est composé majoritairement par l'assurance du particulier, qui devient de plus en plus concurrentiel ce qui complique l'évolution de sa rentabilité. Ainsi, il est plus intéressant pour les assureurs de se tourner sur les flottes automobiles, qui représentent 10,6% du secteur et qui a l'évolution la plus importante sur

le marché des flottes automobiles, avec une augmentation de 4,7% des cotisations entre 2020 et 2021 (cf. Annexe A 7.2). De plus, le nombre de véhicules assurés est également en hausse de 5% par rapport à 2020 avec un total de 4,8 millions de véhicules en 2021 ce qui représente 8,5% du parc.

L'assurance de flotte automobile d'entreprise est une assurance qui couvre le risque de plusieurs véhicules dans un même contrat. Le détenteur du contrat est une personne morale qui peut assurer plusieurs types de véhicules : véhicules léger, poids lourds, remorques et semi-remorques, véhicules agricoles ou de chantier, etc. Une des particularités des flottes automobiles est que le souscripteur de risque n'a pas l'obligation d'indiquer les conducteurs des véhicules ni l'historique des sinistres de la flotte. C'est ainsi le seul responsable des dommages causés par les conducteurs véhicules en vertu de l'article L121-2 du code des assurances. Les flottes automobiles sont très populaires dans certains secteurs, particulièrement ceux qui impliquent des déplacements ou livraisons.

Au sein de la Direction Actuariat & Pilotage Entreprise (DAPE), Véronique Marpillat est la responsable de l'équipe Actuariat Produits et Data Science (APDS) dans laquelle elle dirige plusieurs équipes du secteur IARD Entreprises. On peut compter au sein de ses équipes, l'équipe Actuariat Produit Auto et Transport, l'équipe Actuariat Construction/RC/Caution, ainsi que l'équipe Dommages aux Biens (Immeuble, Risques Industriels, Risques Techniques, Risque Spéciaux Lignes Spécialisées). Ce mémoire a été réalisé dans la branche Actuariat Produits Auto et Transport.

Chez AXA France, différents produits d'assurance automobiles sont proposés. Ils peuvent se différencier de part la taille de la flotte ainsi que l'activité de l'entreprise.

1. Le produit *Mono* : Couvre les flottes de 1 à 4 véhicules. En général, ce produit est à destination des petites entreprises et assure des engins ou des véhicules de plus de 3,5 tonnes (noté P3T5). La tarification de ce produit prend en compte les caractéristiques spécifiques de chaque véhicule mais également de chaque conducteur.
2. Le produit *Parcs dénommés*, aussi appelé *flotte fermée*, ou *Parc* : Couvre les flottes de 5 à 50 véhicules. Ce produit assure des véhicules des professionnels ou entreprises jusqu'à 5 000 personnes. Le tarif est calculé par véhicule et en fonction des antécédents mais ne prend pas en compte les caractéristiques du conducteur car pour ce produits, ils ne sont pas affectés à un véhicule précis. Ici, la tarification se fait donc par véhicules.
3. Le produit *Flottes non dénommées*, aussi appelés *flotte ouverte* : Couvre les flottes à partir de 50 véhicules. La tarification de ce produit ne se fait pas par véhicule mais par catégories de véhicule car il n'y a pas de déclaration des véhicules présents sur la flotte. Une description plus complète du produits sera présentée dans la prochaine section étant donné qu'il s'agit du produit étudié pour ce mémoire. Comme pour le produit parc, les conducteurs ne sont pas affectés à un véhicule précis, il ne peut donc pas y avoir de bonus ou de malus. De plus, de part le nombre de véhicules important dans la flotte, les caractéristiques de chaque véhicule n'étaient

pas connues avant la mise en place du Ficher des Véhicules Assurés (noté FVA) en janvier 2019.

4. Le produit *Garages et Concessions* : Couvre les véhicules présents au sein du garage ou de la concession. Il concerne principalement les professionnels de l'automobile.

Au sein du portefeuille d'AXA France IARD Entreprise, l'assurance automobile occupe 32% du chiffre d'affaire de la branche à fin 2021 (3 Mds€). De plus, les flottes non dénommées représentent également la seconde part la plus importante du chiffre d'affaire de l'assurance automobile d'entreprise, derrière les flottes ouvertes :

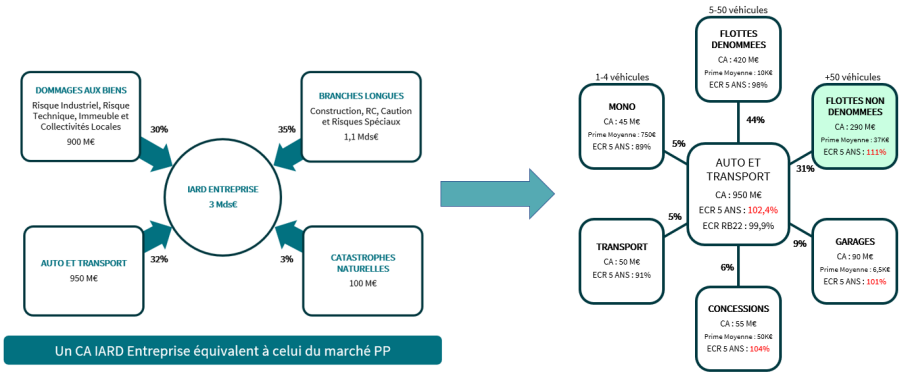


FIGURE 1.2 – Part du Chiffre d'affaire AXA IARD Entreprise en 2021

1.2 Le produit Flottes Non Dénommées

Dans la branche des flottes automobiles d'entreprise, deux grandes familles se distinguent : les flottes fermées (Parc dénommés) et les flottes ouvertes (Flottes Non Dénommées, noté FND). Dans le cas des flottes fermées, les véhicules de la flottes sont connus alors que pour les flottes ouvertes ce n'est pas le cas. En effet, le souscripteur ne renseigne pas les véhicules unitairement comme c'est le cas pour le Parc, mais seulement un nombre de véhicules par type et par garantie :

Type de véhicule	Responsabilité civile	Dommages tous accidents	Incendie	Vol	Bris de glace
Véhicule léger (<3T5)					
Camions (PL > 3T5)					
Bus & Autocars					
Tracteurs					
Semi-Remorques					
Remorques > 3T5					
Engins de chantiers & assimilés					
Engins agricoles					
2 ou 3 roues					
Remorques < 3T5					
Total					

FIGURE 1.3 – Tableau de composition des flottes ouvertes

Afin d'étudier les différents sinistres, il faut distinguer les véhicules de moins de 3,5 tonnes (noté M3T5) des véhicules de plus de 3,5 tonnes (P3T5) étant donné qu'en général, le coût des sinistres est plus important chez les véhicules lourds supérieur à 3,5T. De plus, comme évoqué précédemment, la composition de la flotte n'est pas connue, seul le type de véhicule majoritaire est connu. On distingue différents types de véhicules :

- Les *Transports Public de Voyageurs (TPV)* : majoritairement des véhicules P3T5 mais avec une part non négligeable de M3T5, il s'agit des autocars, autobus ou bien des taxis. C'est un segment avec une sinistralité élevée en dommages et en corporel.
- Les *Transports Public de Marchandise (TPM)* : généralement des véhicules P3T5, véhicules transportant des marchandises qui n'appartiennent pas à l'assuré. C'est un segment qui est peu rentable étant donné que les sinistres sont fréquents et leurs coûts sont élevés.
- Les *Transports pour Propre Compte (TPC)* : véhicules transportant des marchandises appartenant à l'entreprise assurée.
- Les *Engins chantiers*
- Les *Engins agricoles*
- *Autres* : toutes les autres catégories de véhicules, en particulier les véhicules légers (VL) de M3T5.

Il existe également le segment des conservations, qui sont des contrats spécifiques à de grandes flottes pour lesquelles on observe une fréquence de sinistralité élevée. Concernant les garanties au sein des flottes, il y a deux types de garanties :

— Garanties obligatoires : la responsabilité civile (RC). Elle couvre les dommages corporels (RCC) et les dommages matériels (RCM)

— Garanties facultatives :

  - Dommages : couvre les dommages survenus sur les véhicules. Il existe différentes garanties comme le Dommage accidentels (DOMA), le vol (VOL), le bris de glace (BDG), l'incendie (INC).
  - Sécurité du conducteur (SDC) : indemnise le conducteur du véhicule à hauteur des montants plafonnés par le contrat.
  - Protection juridique (PJ) : défend les intérêts de l'assuré. On parle de défense quand la responsabilité de l'assuré est engagée et de recours quand l'assuré a subi un dommage par un tiers.
  - Assistance (ASS) : prend en charge les frais de remorquage et l'assistance au véhicule endommagé.

Il est également à noter qu'il faut distinguer les catégories de flottes de véhicules :

1. Catégorie 1 : Véhicules légers (VL) et Véhicules Utilitaires (VU), ce sont des flottes de véhicules de M3T5.
2. Catégorie 2 : Camions et remorques P3T5, ce sont principalement des TPV, TPM ou TCP. Les remorques de plus de 750 kg doivent avoir leur propre assurance.

3. Catégorie 3 : Bus et Autocar P3T5. Il ne s'agit que de TPV.
4. Catégorie 4 : Engins de chantier
5. Catégorie 5 : Engins agricoles
6. Catégorie 6 : Deux roues.
7. Catégorie 7 : Remorques M3T5, leur assurance est souvent incluse dans celles des catégorie 1 qui les accompagnent au sein d'une même flotte.

La refonte tarifaire des Flottes ouvertes découle de l'avènement du Fichier des Véhicules Assurés (FVA). En effet, avant la mise en place du FVA, les différents véhicules composant les flottes n'était pas connus, mais depuis début 2019, les assureurs ont l'obligation de déclarer à l'AGIRA (Association pour la Gestion des Informations sur le Risque en Assurance) tous les véhicules en RC dans le FVA. Nous pouvons donc récupérer la liste des véhicules à posteriori de l'affaire nouvelle. Cependant, nous ne pouvons pas récupérer l'information pour les autres garanties. Un chapitre sera dédié au Fichier des Véhicules Assurés afin de pouvoir comprendre au mieux son enjeu.

1.3 Présentation et objectifs de l'étude

Dans le cadre de ce mémoire, l'objectif sera de modéliser les garanties Responsabilité Civile Matérielle (RCM) et Responsabilité Civile Corporelle (RCC) sur le produit Flottes non dénommées en faisant une tarification par véhicule grâce à l'apport du Fichier des Véhicules Assurés. Les modèles seront comparés avec le modèle fait par l'équipe Actuariat Produits Auto et Transport en 2022 ainsi qu'au modèle du produits Parc dénommés. Afin de pouvoir challenger le modèle de 2022, des variables provenant de la base Sécurité et Réparation Automobile vont être ajoutées dans l'objectif d'avoir une segmentation plus fine du risque assuré, ce qui permettra une tarification plus précise des flottes ouvertes.

De plus, une étude sur l'activité de l'entreprise sera présentée. L'objectif de cette étude est d'avoir une information de plus sur les flottes afin de savoir si les véhicules se déplacent en moyenne autour de l'adresse de l'entreprise, à un niveau plutôt régional, national ou bien international. Cette étude amènera à la construction d'une variable et il sera étudié si celle-ci pourrait avoir une importance lors de la modélisation du coût moyen ou de la fréquence des garanties étudiées.

Chapitre 2

Le Fichier des Véhicules Assurés

2.1 La mise en application du FVA

Près de 800 000 personnes roulent dans un véhicule qui n'est pas assuré. On observerait depuis 2014 une augmentation de 30% du nombre de véhicules non assurés. On compte, en 2022, 157 personnes qui sont décédées dans un accident impliquant un véhicule non assuré, ce qui correspond à une hausse de 11,3% depuis 2019 d'après le Fonds de Garantie des Assurances Obligatoires de dommages (FGAO). Le FGAO a la charge d'indemniser les victimes d'accidents de la route provoqués par des conducteurs non assurés ou ayant pris la fuite. Celui-ci a pris en charge près de 24 000 victimes de dommages matériels et corporels au titre de la circulation automobile et a versé plus de 116 millions d'euros d'indemnités. De plus, on note que 80% des véhicules impliqués sont des véhicules légers comme le montre le graphique ci-dessous :

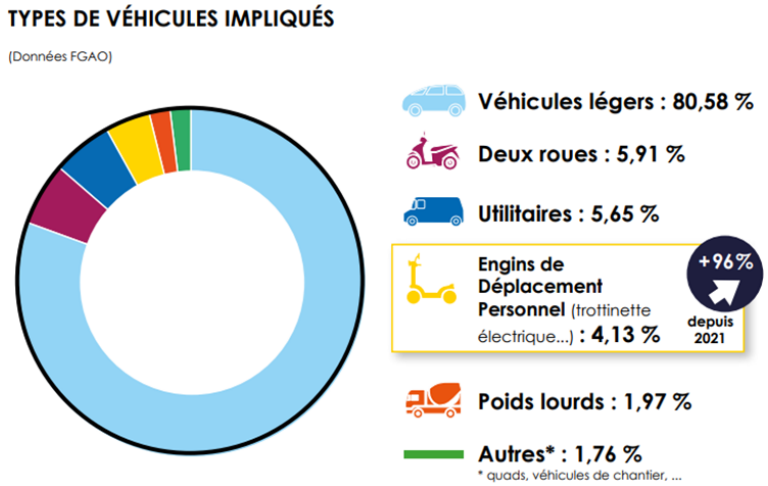


FIGURE 2.1 – Types de véhicules impliqués

D'après le communiqué de presse du 26 juillet 2018 de la délégation à la sécurité routière [1], en 2016, 235 personnes sont décédées dans un accident routier impliquant un véhicule non assuré, ce qui représente 7% de la mortalité routière. Afin de lutter contre la non-assurance, de faciliter l'identification des véhicules en cas de délit de fuite, vol ou bien de trafic, *« décidée lors du Comité interministériel de sécurité routière du 2 octobre 2015, la création d'un fichier des assurés (FVA) a été introduite dans la loi de modernisation de la justice du XXI^{e} siècle du 18 novembre 2016 »*. Cette loi prévoit de répertorier les autos et motos assurés dans un fichier nommé le fichier national des véhicules assurés (FVA). Le décret n°2018-644 du 20 juillet 2018 précise les modalités de constitutions et d'alimentation de ce fichier. Tous les assureurs sont ainsi dans l'obligation d'alimenter leurs véhicules assurés pour une mise en application au 1er janvier 2019.

Les compagnies d'assurances ont l'obligation de communiquer à l'AGIRA (Association pour la Gestion des Informations sur le Risque en Assurance), dans un délai de 72 heures après la souscription ou la résiliation d'une police d'assurance, les données des véhicules. L'AGIRA a la charge de traiter les flux reçus et de constituer et alimenter le FVA dans un délai de traitement de maximum six heures.

Le fichier contient les informations relatives aux contrats d'assurances automobile souscrits, à savoir l'immatriculation du véhicule, qui est une information clé du véhicule, le nom de l'assureur, le numéro de contrat et la période de validité du contrat. Les compagnies d'assurances recensent alors à l'AGIRA leurs véhicules assurés ainsi que leurs opérations (entrées ou sorties, modification de la flotte assurée, entrée en vigueur ou cessation de la garantie responsabilité civile, etc). L'article L.451-1-1 du Code des assurances ainsi que les articles R.451-2 et suivants du même code prévoient les modalités de constitution, d'alimentation et d'interrogation du FVA.

Tous les véhicules immatriculés qui sont assurés sont alors concernés par le FVA, c'est-à-dire tous les véhicules terrestres à moteur. La mise à disposition des données s'est effectuée en deux temps :

— À partir du 1er janvier 2019 : véhicules utilitaires et particuliers, deux-roues, les engins de chantier à caractère routier, les remorques dont le poids est supérieur à 750 kg, les semi-remorques, et les contrats garages, flottes et concessions.

— À partir du 1er janvier 2021, les remorques inférieures à 750 kg et les véhicules agricoles ont été ajoutés.

En cas de défaut d'assurance, on distingue deux types de risques : des risques financiers et des risques judiciaires. Quelque soit ces risques, ils dépendent de la gravité de l'infraction et de ses conséquences.

Pour les risques financiers, en cas d'accident causé par un conducteur sans assurance, les victimes sont indemnisées par le FGAO. Cependant, le responsable de l'accident, qui

1. `https://www.interieur.gouv.fr/Archives/Archives-des-communiques-de-presse/2018-Communiques/Creation-du-fichier-des-vehicules-assures-FVA`

ne dispose donc pas d'assurance, ne sera indemnisé ni pour les dégâts matériels, ni pour ses propres blessures par le FGAO. Celui-ci devra prendre en charge ces coûts, mais il devra également rembourser au FGAO la totalité des sommes engagées par celui-ci pour l'indemnisation des victimes de l'accident. En cas d'accident grave, ces sommes peuvent être très élevées. En fonction des revenus du conducteur sans assurance, il devra rembourser chaque mois une mensualité, cela peut même parfois durer pour le reste de sa vie.

Pour les risques judiciaires, la conduite sans assurance est un délit qui est punie par la loi. Lors d'une première interpellation, la peine encourue pour défaut d'assurance est une amende de 3750 euros, pouvant être accompagnée de peines complémentaires telles que la suspension ou bien l'annulation du permis, avec possibilité d'interdiction de le repasser ainsi que la confiscation du véhicule. La loi prévoit que la conduite sans assurance peut être sanctionnée d'une amende forfaitaire de 750 euros, minorée à 600 euros pour un règlement dans les 15 jours, ou, majorée à 1500 euros au bout de 45 jours, lors de la première interpellation pour un particulier. En cas de récidive, les sanctions pénales sont plus sévères et peuvent aller jusqu'à 7500 euros d'amende, annulation de permis avec interdiction de le repasser et confiscation du véhicule.

Délai de paiement en fonction du mode de paiement		Type amende et montants		
Chèque ou espèces	Télépaiement (internet ou téléphone) ou timbre dématérialisé	Amende forfaitaire	Montant à payer Particulier	Montant à payer Personne morale
15 jours	30 jours	Minorée	600 €	3 000 €
45 jours	60 jours	Normale	750 €	3 750 €
Au-delà de 45 jours	Au-delà de 60 jours	Majorée	1 500 €	7 500 €

TABLE 2.1 – Montant à payer selon le délai et le mode de paiement

2.2 Son apport pour les flottes non dénommées

Pour en revenir aux flottes ouvertes, étant donné que l'article L.211-1 du Code des Assurances impose à tout propriétaire de véhicule terrestre à moteur de souscrire une assurance couvrant au moins sa responsabilité civile en cas d'accident, les véhicules composant la flotte sont alors répertoriés dans le FVA et ont a minima la garantie en responsabilité civile. La mise en place du FVA a impliqué pour les assureurs et leurs délégataires la déclaration obligatoire des véhicules terrestres à moteur immatriculés à l'AGIRA. Les assureurs sont obligés d'être rigoureux lors de la saisie des mouvements de véhicules, notamment lors de l'affaire nouvelle d'un contrat.

Ainsi, le FVA apporte pour AXA France une meilleure connaissance des véhicules et des risques assurés sur les flottes ouvertes, ce qui permet ainsi d'apporter un meilleur service au client, d'optimiser les processus de gestion ainsi que de proposer une tarification plus fine, en particulier lors du renouvellement. La tarification des flottes ouvertes pourrait alors se rapprocher de la tarification du Parc. Les caractéristiques techniques des véhicules peuvent alors être connues grâce à la connaissance des immatriculations via le FVA. En effet, il sera possible de faire des jointures avec différentes bases de données avec comme clé le numéro de contrat ainsi que l'immatriculation des véhicules.

Chapitre 3

Les données du portefeuille

Grâce à la mise en pratique du Fichier des Véhicules Assurés, il est désormais possible d'obtenir l'immatriculation de chaque véhicule assuré. L'immatriculation d'un véhicule est une donnée clé qui permet d'intégrer de nouvelles caractéristiques techniques grâce à des bases de données internes à l'entreprise, mais également des données provenant de sources externes. Ces nouvelles données peuvent ainsi être intégrées aux modèles de tarification afin d'améliorer la rentabilité et le tarif proposé grâce à de nouvelles variables explicatives.

Ainsi, ce chapitre présentera les différentes bases de données à disposition ainsi que la création des bases de données pour les modèles de tarification.

3.1 Les bases de données utilisées

Cette section présente les différentes bases de données utilisées ainsi que leurs provenances.

3.1.1 Bases de données interne à AXA

La base portefeuille

Afin d'obtenir un modèle de tarification pour chaque contrat par véhicule, il convient de récupérer la *base portefeuille* de la branche IARD Entreprise d'AXA France. Depuis cette base, il faut récupérer les différentes informations relatives aux contrats des flottes ouvertes. En effet, lors de la souscription ou lors d'une modification d'un contrat, le souscripteur saisis des données grâce à des outils de souscription et de gestion (AXAPAC, FLOWS, OSE TPC+). Ces données saisies sont alors stockées sur les systèmes d'informations, que l'on appelle Infocentre. Sur cette base, on peut extraire des données à la maille contrat comme le numéro de contrat, le nom de l'affaire, l'activité principale, la prime annuelle TTC, l'intermédiaire et le nom de l'intermédiaire, etc. Ces données sont alors communes aux différents véhicules composant une flotte automobile d'entreprise.

Ces opérations de récupération des données se font sous SAS. Les données récupérées concernent la période de 2019 à 2021.

La base véhicule

La base véhicule d'AXA France Entreprise (nommée base AEVEH) comprend tous les véhicules déclarés dans les outils AXA, hors particulier mais y compris professionnel. On y retrouve une ligne par numéro de contrat, immatriculation du véhicule, date de début de garantie, et date de fin de garantie. Les garanties couvertes par le contrat d'assurance ne sont renseignées que pour le Parc. Pour les flottes ouvertes, les garanties ne sont pas renseignées, uniquement la garantie Responsabilité Civile. Sur cette base, on y retrouve différentes informations comme le genre du véhicule, le nombre de places, le poids, le type, sa valeur, le poids total en charge (PTAC), la puissance fiscale, la marque, l'activité du véhicule (Ambulance, Auto-école, Fourgon funéraire, taxis, Ramassage scolaire, ...), la cylindrée du moteur, des caractéristiques du véhicules qui peuvent être intéressantes dans le cadre d'une tarification automobile.

La base sinistre

La tarification des garanties RCM et RCC s'effectue grâce aux sinistres qui ont eu lieu sur la période de 2019 à 2021. Il faut ainsi récupérer depuis la base source, qui est la base par unité de prestation (noté UP, il s'agit de la typologie des sinistres) du rapport d'expérience, les sinistres rattachés aux couples contrat/immatriculation issus de la base risque sur les survenances 2019, 2020 et 2021. Les données des sinistres sont issues des systèmes de gestion des sinistres et contiennent des renseignements tels que le numéro de sinistre, la date de survenance, la date d'ouverture, le montant du sinistre, la garantie sinistrée, la nature du sinistre (matérielle ou bien corporelle), l'immatriculation du véhicule, ou bien encore le genre du véhicule. Pour chaque sinistre déclaré, on peut y associer un numéro de sinistre avec un numéro de contrat. Il est donc possible de retrouver le nombre de sinistres survenus durant une année ainsi que le montant de ces derniers pour chaque numéro de contrat. Ces informations peuvent alors être ajoutées à la base contrat via le numéro de contrat.

3.1.2 Base de données externe

Pour chacune des bases vues précédemment, le couple numéro de contrat/immatriculation du véhicule est présent et permet ainsi de récolter des informations grâce à des bases de données externes. En effet, il est possible de récupérer des informations sur les véhicules grâce à la carte grise par exemple. Ces informations peuvent être mises à disposition par l'État ou bien par des organismes professionnels, des associations.

La base du Fichier des Véhicules Assurés

Comme décrit dans le chapitre 2, afin de lutter contre la conduite sans assurance, le Comité interministériel de sécurité routière a décidé de la création d'un fichier assurés (FVA) dans le décret n°2018-644 du 20 juillet 2018. Constitué par l'Association pour la gestion des informations sur le risque en assurance (AGIRA), ce fichier contient les informations relatives aux contrats souscrits par les assurés : l'immatriculation du véhicule, le nom de l'assureur et le numéro du contrat avec sa période de validité. Cette base a pour utilité de compléter et enrichir la base des véhicules des flottes non dénommées. C'est ainsi que ce fichier permet de récupérer l'information véhicule des contrats.

La base du Système d'Immatriculation des Véhicules

L'ancien système d'immatriculation de véhicule, le Fichier National des Immatriculations (FNI) a été remplacé par le Système d'Immatriculation des Véhicules, SIV, le 15 avril 2019. Pour le SIV, son principe majeur est d'associer une immatriculation à un véhicule contrairement au FNI où l'immatriculation était associée au propriétaire. Ce nouveau système a permis aux propriétaires de faire immatriculer leur véhicule directement en ligne ou bien chez un garagiste habilité au SIV plutôt qu'en préfecture.

Le SIV a plusieurs objectifs :

— Simplifier les démarches d'immatriculation

— Sécuriser le certificat d'immatriculation (carte grise)

— Impliquer les professionnels de l'automobile dans le processus d'immatriculation lors de la vente d'un véhicule

— Simplifier la gestion des services de l'État pour la circulation automobile

L'Agence Nationale des Titres Sécurisés (ANTS) est en charge de la mise en oeuvre du SIV en accord avec le Ministère de l'Intérieur, la Délégation à la Sécurité Routière et en partenariat avec la DNUM pour l'hébergement des plateformes et leurs exploitations. [1]

Le Système d'Immatriculation des Véhicules contient les informations suivantes : l'identité du titulaire de la carte grise avec nom, prénom, date et lieu de naissance, adresse, l'autorisation de circuler, ainsi que des informations concernant le véhicule soit le numéro d'immatriculation, le numéro VIN, les caractéristiques techniques du véhicule, les mentions d'usage, la situation du véhicule (retrait, suspension, destruction par exemple), la date de la première immatriculation, ainsi que l'identité des professionnels habilités à transmettre les données au système.

1. `https://ants.gouv.fr/nos-missions/les-solutions-numeriques/siv`

Cette base peut ainsi servir à compléter la base véhicule d'AXA IARD Entreprise grâce aux différentes variables contenant des informations sur les caractéristiques techniques des véhicules comme la marque, la date de mise en circulation, l'énergie, la puissance fiscale, la cylindrée du moteur, la puissance réelle, le nombre de places, le poids à vide, la carrosserie ou encore le nom commercial.

La base Sécurité et Réparation Automobile

Sécurité et Réparation Automobile (SRA) est un organisme professionnel créé en 1977. Toutes les compagnies d'assurances automobiles sont adhérentes à cet organisme. Il possède le statut d'association loi 1901 et a vocation de promouvoir toutes les études au sein de la profession, et, avec les acteurs de l'automobile, de mettre en oeuvre tous les moyens nécessaires à la réalisation d'actions permettant de limiter le nombre et le coût des sinistres. L'une de ses missions principales est de diffuser aux adhérents de l'information sur les véhicules. En effet, SRA met à disposition, via ses bases de données, toutes les caractéristiques techniques et commerciales des véhicules (2, 3 et 4 roues) afin de permettre leur identification. Cependant, ces bases ne concernent que les véhicules particuliers et utilitaire inférieur à 3,5 tonnes ainsi que les 2/3 roues et quad motorisés. Ces véhicules ont également un système de classification, déterminé à partir des données objectives et uniquement liées à la conception du véhicule.

Également, SRA a pour mission d'encourager la sécurité. En effet, travaillant avec les services d'études et de recherche des principaux constructeurs automobiles, les performances de sécurité active et passive des véhicules sont intégrées afin d'assurer une veille technologique et de pouvoir percevoir les évolutions techniques, de conception ou d'équipement intervenant sur les véhicules. Une autre de ses missions est de participer à la protection contre le vol des véhicules. Même si sur les véhicules particuliers et utilitaires légers, les équipements d'anti-démarrage sont généralisés, le classement SRA reste nécessaire pour valider le niveau de protection exigé par le cahier des charges. Enfin, une dernière des missions de SRA est de maîtriser les coûts de réparation en s'appuyant sur des indicateurs tels que le suivi de l'évolution du prix des pièces de rechange chez les différents constructeurs, le suivi du taux horaires de main-d'oeuvre carrosserie et équivalents horaires des prix des ingrédients de peinture. De plus, SRA participe à l'amélioration de la réparabilité des véhicules en analysant leur conception avec les constructeurs automobiles.

3.2 Traitement/Construction des bases de données

Maintenant, que les différentes bases ont été présentées, il sera exposé la méthode de construction du modèle finale. Il est à remarquer que sur ces différentes bases, l'immatriculation du véhicule est une information toujours présente.

Tout d'abord, la base véhicule (AEVEH) a été enrichi grâce à l'ajout de variables provenant des bases FVA et SIV dans un premier temps, puis vint l'ajout des variables SRA dans un second temps. Afin de joindre la base FVA avec la base AEVEH, il faut appliquer le même formatage pour la variable Police FVA_NumeroPoliceAXA et NUMCNT et pour la variable de l'immatriculation du véhicule FVA_NumeroImmatriculationNorme et VHIMMAT. Il a également été nécessaire de formater et retraiter quelques autres variables provenant de la base FVA. La jointure se fait alors avec comme clé l'immatriculation du véhicule. D'autres retraitements sont alors faits comme la suppression des polices/immatriculations dont l'immatriculation est vide ou encore dont la date de début de garantie est supérieure ou égale à la date de fin.

Un travail similaire a été fait sur la base SIV avec comme clé de jointure l'immatriculation du véhicule également. Certaines variables ont déjà été discrétisées, comme par exemple la variable concernant le poids à vide, afin de pouvoir interpréter plus facilement les modèles. Les variables SIV qui sont ajoutées sont les suivantes :

Variables SIV	**Signification**
puiss_max_siv_tranche	Puissance maximale du moteur discrétisée
carrosserie_siv	Carrosserie du véhicule
Tranche_Cylindre_SIV	Cylindrée du moteur (en litres) discrétisée
Tranche_Vit_mot_siv	Vitesse du moteur (en tours par minute) discrétisée
Tranche_poids_vide	Poids à vide du véhicule discrétisé
marque_SIV	Marque du véhicule
nom_com_siv	Nom commercial du véhicule
genre_siv	Genre du véhicule
energie_siv	Type d'énergie du véhicule
nb_places_siv	Nombre de places du véhicule
co2_siv	Taux d'émission de CO2 du véhicule

Table 3.1 – Listes des variables SIV

La base véhicule SIV est alors créée. Par la suite, la base de donnée SRA a également dû être traité afin de garder uniquement les variables qui peuvent être utiles pour la modélisation. Il y a ainsi une base véhicule comprenant des variables provenant des bases FVA, SIV et SRA. Les variables provenant de la base SRA sont celles-ci :

Variables SRA	Signification
code_sra	Code SRA
GROUP_SRA	Groupe SRA
CLASS_PRIX_sra	Classe de prix SRA
CLASS_REPAR_sra	Classe de réparation SRA
GENRE_SRA	Genre du véhicule
marque_SRA	Marque du véhicule
nom_com_SRA	Nom commercial du véhicule
modele_sra	Modèle du véhicule
BOIT_VIT_sra	Boite de vitesse du véhicule
PRIXACT_sra	Prix actuel du véhicule
PRIXORI_siv	Prix d'origine du véhicule
poids_vide_sra	Poids à vide du véhicule
PTAC_sra	Poids total autorisé en charge du véhicule
CARROSS_sra	Carrosserie du véhicule
SEGMENT_sra	Segment automobile du véhicule
TRANS_sra	Transmission du véhicule
TYP_ALIM_sra	Type d'alimentation du véhicule
energie_SRA	Type d'énergie du véhicule

TABLE 3.2 – Listes des variables SRA

Après avoir joint la base SRA, la base véhicule finale comprend beaucoup de variables avec les caractéristiques techniques des véhicules. Certaines variables comme le poids à vide sont communes entre les variables provenant de SIV et celles de SRA. Pour la modélisation, il ne faudra alors pas prendre les deux variables pour ne pas avoir de surapprentissage. Sur les variables SRA, aucune n'a été discrétisée pour la modélisation, car ce travail ainsi que d'autres retraitements seront expliqués dans le chapitre suivant.

De plus, grâce à l'immatriculation du véhicule, il est possible de récupérer des données liées à la base sinistre. Ce sont des données importantes pour la base de modélisation, notamment le nombre de sinistres et la charge des sinistres. Ainsi, la base de modélisation contient :

— Des données véhicules provenant de la base véhicule AXA, avec des informations sur les véhicules présents dans les flottes

— Des données contrats provenant de la base portefeuille, avec des informations telles que l'activité de l'entreprise, les cotisations, le code NAF, ...

— Des données sinistres provenant de la base sinistre, avec le nombre de sinistres, la charge de sinistres, la garantie du sinistre, ...

— Des données externes qui viennent compléter la base véhicule avec des informations SIV, FVA et SRA

Ci-dessous, un schéma récapitulatif des bases qui ont enrichi la base véhicule :

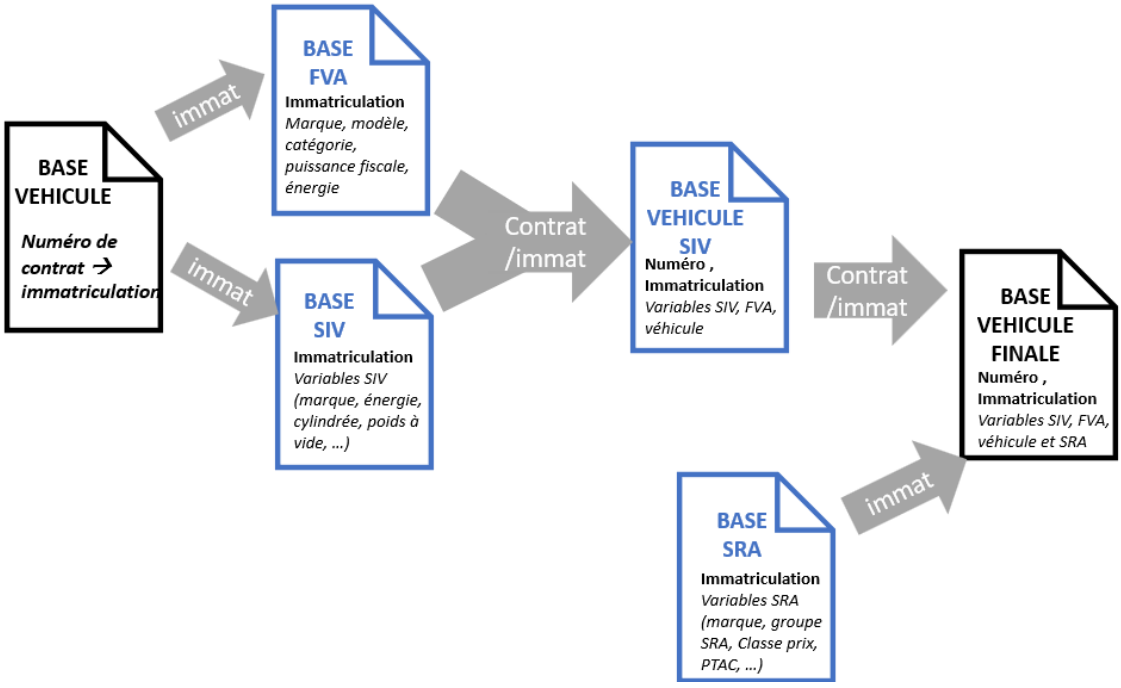


FIGURE 3.1 – Schéma synthèse formation base véhicule

Une fois la base de modélisation créée, il est possible de créer les bases de modélisation qui seront utilisées pour ce mémoire, à savoir les bases de modélisation suivante :

- Une base pour modéliser le coût moyen des sinistres de la garantie RCM
- Une base pour modéliser le coût moyen des sinistres de la garantie RCC
- Une base pour modéliser la fréquence des sinistres de la garantie RCC
- Une base pour modéliser la fréquence des sinistres forfaitaires de la garantie RCM
- Une base pour modéliser la fréquence des sinistres non forfaitaires de la garantie RCM

Il y a une distinction entre les sinistres forfaitaires et non forfaitaires. En effet, certains sinistres peuvent engendrer des recours entre assureurs selon la responsabilité de leur assuré, à savoir responsable à 0%, 50% ou 100%. En effet, il existe deux conventions : la convention IRCA et la convention IRSA.
La **convention IRCA** est la convention d'Indemnisation et de Recours Corporel Automobile, qui fait partie des conventions d'indemnisation que les compagnies d'assurances signent entre elles afin de "faciliter l'indemnisation des victimes". Elle s'applique pour tous les accidents de la route survenus sur le territoire français (ou à Monaco) dès lors

qu'au moins deux véhicules sont impliqués. En vertu de cette convention, tous les dommages corporels sont pris en charge par l'assureur dès lors que le taux d'Atteinte à l'Intégrité Physique et Psychique (AIPP) est inférieur à 5%, au-delà, l'assurance adverse indemnise l'assuré. Après avoir procédé à l'indemnisation de la victime et au règlement des tiers payeurs, l'assureur exerce son recours contre le ou les autres assureurs des véhicules impliqués dans l'accident. En cas d'AIPP nulle, le recours est forfaitaire, avec un forfait qui est fixé chaque année par la commission d'application de la convention. Au 1^{er} janvier 2023, le forfait est fixé à 1 200 €, contre 1 254 € en 2022.
Pour les dommages matériels, il y a la **convention IRSA** qui est la convention d'Indemnisation et de Recours entre Sociétés d'Assurance. Elle concerne les accidents matériels, c'est-à-dire sans aucune conséquence corporelle même minime, survenus en France métropolitaine et dans les DOM (ainsi qu'à Monaco) impliquant au moins deux véhicules dont les assureurs sont signataires de la convention. Tout comme l'IRCA, après indemnisation de l'assuré, l'assureur du conducteur non responsable présente à l'assureur du conducteur responsable son recours. Si le montant des dommages est inférieur au plafond fixé à 6 500 €, le recours est forfaitaire dans la limite de 1 776 €, contre 1 706 € en 2022. Le recours exercé est proportionnel au niveau de responsabilité du conducteur responsable.

Finalement, dans les bases de modélisation pour la garantie RCM, le coût moyen ne prend en compte que les sinistres hors forfaits et la distinction pour les sinistres forfaitaires se fait pour le modèle de fréquence des sinistres.

3.3 Quelques statistiques descriptives

Avant de séparer les bases de modélisation pour les garanties RCM et RCC, quelques statistiques descriptives ont été réalisées sur la base FND de 2019 à 2021. Pour les sinistres entre 2019 et 2021, 139 462 sinistres sont comptés. Il y a également 3282 contrats et 92 199 véhicules différents. Le tableau suivant fait le split par garantie :

Garantie	Nombre contrats*	Nombre véhicules*	Nombre sinistres
BDG (Bris de glace)	2 382	31 394	37 224
CLIM (Climatique)	432	1 132	1 144
DOM (Dommages)	2 457	30 979	35 804
DR (Défense Recours)	500	1 048	1 058
INC (Incendie)	191	254	257
RCC (RC Corporelle)	947	2 996	3 028
RCM (RC Matérielle)	2 778	49 612	58 134
SDC (Sécurité du Conducteur)	221	355	37 224
VOL (Vol)	732	2 349	2 458

TABLE 3.3 – Sinistres par garanties sur les années 2019 à 2021

* Nombre de contrats ou de véhicules ayant eu au moins un sinistre entre 2019 et 2021.

La charge moyenne par garantie pour les sinistres de 2019 à 2021 a été regardée :

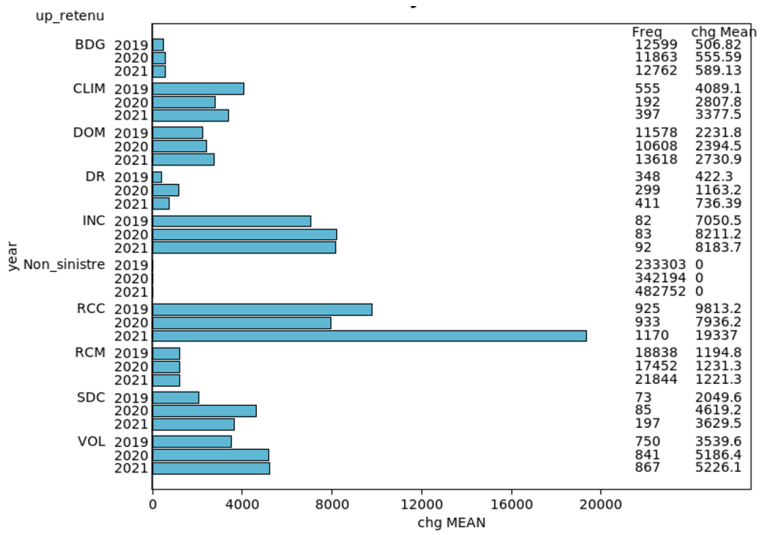


FIGURE 3.2 – Charge moyenne des sinistres par garanties par année

La garantie avec la charge moyenne des sinistres la plus élevée est la garantie responsabilité civile corporelle. Ceci s'explique par le fait qu'un sinistre impliquant du corporel devient plus important que n'importe quel autre sinistre car les frais liés au sinistre peuvent vite avoir un montant élevé. Les frais peuvent être médicaux voir judiciaires. Cependant, les sinistres intervenant sur la garantie RCC ne sont pas les plus fréquents :

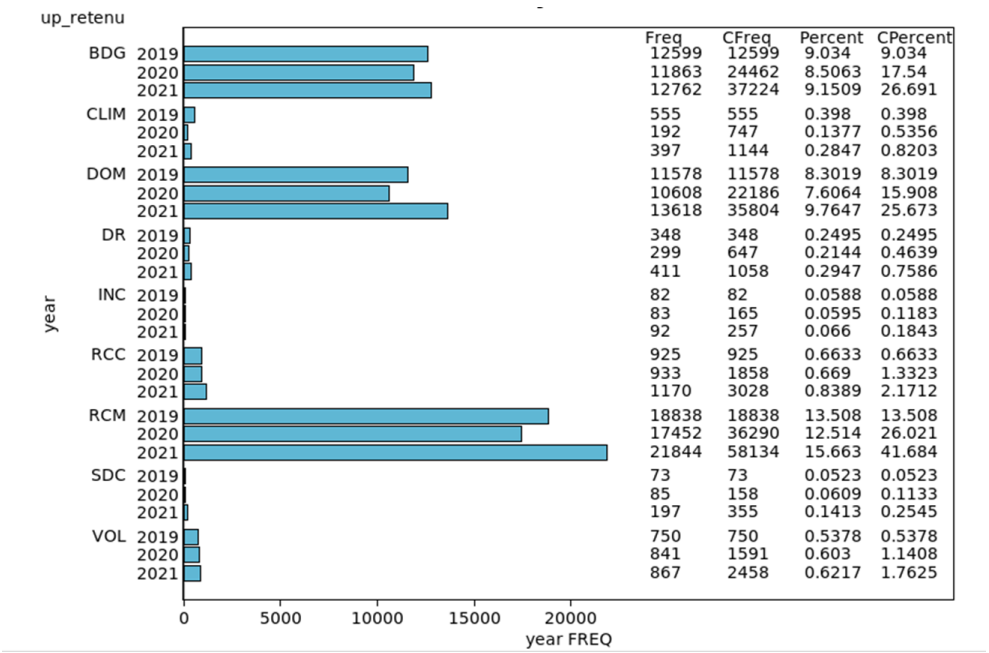


FIGURE 3.3 – Fréquence des sinistres par garanties par année

En effet, les sinistres sont plus fréquemment liés à la garantie RCM, suivis des garanties Bris De Glace (BDG) et dommages accidentels (DOM). Avec ces statistiques ci-dessus, il est facile de voir que la modélisation des garanties RCM et RCC est importante dans le cadre des flottes ouvertes au vu du montant des sinistres ainsi que de leurs fréquences.

Chapitre 4

Retraitement et analyse des données

La construction des bases de données, ainsi que les retraitements des données sont des étapes primordiales lors de la modélisation. En effet, pour éviter d'avoir des résultats et conclusions erronées, il est nécessaire de traiter les anomalies et d'analyser les variables des bases de données. Ce travail a été réalisé sur les variables provenant de la base SRA qui ont été ajoutées pour cette version de modélisation, ainsi que sur certaines variables provenant de la base SIV.

4.1 Le traitement des valeurs manquantes

Dans les bases de données, il est possible de trouver des valeurs sans renseignements du fait d'erreurs de saisie, d'information non disponibles ou encore d'omissions intentionnelles. Il est donc important d'aborder de manière adéquate les valeurs manquantes. Une fois les valeurs manquantes identifiées et caractérisées, il faut choisir la meilleure approche de traitement en fonction du contexte et de la nature des données. Plusieurs stratégies sont possibles :

- **Imputation des Données** : Cette méthode consiste à estimer et à remplacer les valeurs manquantes calculées à partir des données existantes. L'imputation des données par la moyenne, la médiane, le mode ou encore l'imputation basée sur des modèles plus complexes. Cette approche nécessite cependant une réflexion soignée pour ne pas biaiser les résultats.
- **Suppression sélective** : Pour les données où les valeurs manquantes sont peu nombreuses, et ne représentent qu'une petite proportion des données, la suppression des observations concernées peut être une possibilité. Cependant, cette approche réduit la taille de l'échantillon, il ne faut donc pas que cette réduction soit significative.
- **Modélisation prédictive** : Pour prédire les valeurs manquantes, il est possible d'utiliser des techniques d'apprentissage automatique en se basant sur d'autres variables des données. Il faut tout de même surveiller les performances du modèle pour éviter un sur-ajustement.

Dans le cadre de ce mémoire, le pourcentage de valeurs manquantes ne représentait que $4,8\%$ soit moins de 5%. Le taux de manquant n'est donc pas significatif. En effet, il n'est pas possible de supprimer les valeurs manquantes car leur absence pourrait influencer les résultats de l'analyse. De plus, pour les variables provenant de SRA s'il y a une variable avec des variables manquantes alors toutes les autres variables provenant de SRA sont également manquantes. Les valeurs manquantes peuvent avoir un impact sur la performance des modèles. Enfin, sur le modèle AKUR8 (plateforme utilisée pour la modélisation, cf. chapitre 7), il est possible d'indiquer les valeurs manquantes afin que la plateforme puisse gérer cette information dans les modèles. Sur les variables catégorielles, les manquants sont attribués à une modalité spéciale comme par exemple "00-NR", "NR" ou encore "0.Manquants". Cette approche est courante dans le traitement des données et permet de conserve l'information sur l'absence de valeur tout en évitant de supprimer les observations comportant des valeurs manquantes, ce qui pourrait entraîner une perte de données importante. La plateforme AKUR8 est capable de gérer sans problème les modalités manquantes, ce qui facilite ainsi l'interprétation des résultats.

4.2 Retraitement des variables

Le retraitement des variables est une étape essentielle visant à transformer, combiner, simplifier ou même créer des variables afin de faciliter l'analyse et d'extraire des informations plus pertinentes. Parmi les techniques de retraitement les plus utilisées, il est possible de retrouver la discrétisation des variables quantitatives et le regroupement de modalités.

La discrétisation d'une variable numérique consiste à diviser une plage continue de valeurs en intervalles discrets ou catégories distinctes. Cette technique offre plusieurs avantages dans l'analyse des données et la modélisation, notamment :

— La simplification de la modélisation et interprétation facilitée : en effet, la discrétisation rend le modèle moins complexe en regroupant des valeurs similaires dans des catégories. Cela peut ainsi faciliter l'analyse et l'interprétation des données. Les variables discrètes sont souvent plus faciles à interpréter, ce qui rend les conclusions plus compréhensibles. Cependant, la discrétisation peut présenter des inconvénients comme la perte d'information et ainsi la possibilité d'introduire des biais. On peut citer en contre exemple, la discrétisation de l'âge des assurés : une tranche d'âge pour les conducteurs, dits "jeunes", de 18 à 30 ans ne serait pas correcte. En effet, les risques pour un assuré de 18 ans et un assuré de 29 ans ne sont pas les mêmes. De plus, les conducteurs de 18 à 25 ans sont plus à risque que les conducteurs de 26 à 30 ans.

— Traitement des valeurs manquantes : la discrétisation peut fournir une approche pour gérer les valeurs manquantes en les regroupant dans une catégorie spéciale. On évite ainsi d'ignorer les données manquantes et cela peut être utile certaines analyses.

— Réduction de la sensibilité pour les valeurs aberrantes : en effet, elles peuvent avoir moins d'impact que sur les valeurs continues. Cela peut rendre les modèles plus robustes face aux valeurs aberrantes.

Certaines variables provenant du SIV ont également été discrétisées à nouveau afin de challenger la discrétisation faite en 2022 par les membres de l'équipe APAT.

Deux types de méthodes ont été utilisés afin de faire la discrétisation des variables numériques SRA et SIV. La méthode quantile et la méthode utilisant XGBoost (eXtreme Gradient Boosting).

1. *La méthode basée sur les quantiles* a pour but d'avoir une répartition égale des effectifs, c'est-à-dire chacune des classes créées contiendra le même nombre de données. C'est une approche couramment utilisée qui consiste à diviser les données en intervalles en utilisant les valeurs des quantiles. Les quantiles sont des valeurs qui divisent un ensemble de données en portions égales. Par exemple, les quartiles divisent les données en quatre parties égales : le premier quartile (Q1) qui correspond au 25ème percentile, le deuxième quartile qui correspond à la médiane (50ème percentile), et le troisième quartile (Q3) qui correspond au 75ème percentile. Dans chacun des intervalles de l'exemple précédent, il y a 25% des données. Cette méthode prend en compte la distribution réelle des données, les intervalles ne sont pas faits de façon arbitraire. Cependant, avec cette méthode il est possible d'obtenir des intervalles de tailles inégales ce qui peut ne pas refléter correctement la structure sous-jacente des données. De plus, il y a une perte d'information en divisant une variable en intervalles basés sur les quantiles ce qui peut rendre les résultats moins précis et nuancer les différences subtiles entre les observations.
2. *L'approche utilisant Extreme Gradient Boosting* est une approche supervisée qui consiste à créer des classes en analysant les valeurs d'une variable réponse Y face aux variations des variables à discrétiser X. XGBoost est un modèle d'apprentissage automatique utilisé pour les problèmes d'apprentissage supervisé dans lesquels est utilisé un ensemble de variables explicatives pour prédire une variable réponse. Cette méthode peut être utilisée pour réaliser une classification en prédisant l'appartenance d'observations à une classe d'une variable qualitative, sur la base de variables explicatives.

Les différentes variables qui ont été discrétisées sont les suivantes :

— PRIXACT_sra : qui représente le prix actuel du véhicule SRA
— PRIXORI_sra : qui représente le prix d'origine du véhicule SRA
— PTAC_sra : qui correspond au poids total avec charge du véhicule SRA
— Poids_vide_sra : qui représente le poids à vide du véhicule SRA
— Poids_vide_siv : qui représente le poids à vide du véhicule SIV
— Cylindree_siv : qui représente la cylindrée du moteur d'un véhicule SIV

— Co2_siv : qui représente le taux d'émission de CO2 (dioxyde de carbone) du véhicule SIV

— Puiss_siv : qui représente la puissance du véhicule SIV

— Puis_fiscale_SIV : qui représente la puissance fiscale du véhicule SIV

La discrétisation s'est effectuée sur chacune des bases de modélisation et le choix entre la méthode quantile ou la méthode XGBoost se fait en comparant les variations du coût moyen ou de la fréquence, entre les différentes classes ainsi que l'exposition. Il faut choisir la méthode pour laquelle le coût moyen ou la fréquence varie fortement et donne une classification assez fine tout en ayant une exposition assez élevée. Ci-dessous, la méthode de discrétisation choisie pour chaque variable à chaque base de modélisation :

Variables		Méthode de discrétisation choisis				
Provenance	Nom	Cout moyen RCM	Fréquence forfaitaire RCM	Fréquence hors forfait RCM	Cout moyen RCC	Fréquence RCC
SIV	Co2_siv	Quantile	XGBoost	Quantile	Quantile	XGBoost
	Cylindree_siv	XGBoost	Quantile	XGBoost	XGBoost	Quantile
	Poids_vide_SIV	XGBoost	Quantile	Quantile	XGBoost	Quantile
	Puiss_SIV	XGBoost	Quantile	XGBoost	XGBoost	Quantile
	Puis_fiscale_SIV	Quantile	Quantile	Quantile	Quantile	Quantile
SRA	PRIXACT_sra	Quantile	XGBoost	XGBoost	Quantile	XGBoost
	PRIXORI_sra	Quantile	XGBoost	XGBoost	Quantile	XGBoost
	PTAC_sra	Quantile	Quantile	Quantile	XGBoost	Quantile
	Poids_vide_SRA	Quantile	Quantile	Quantile	XGBoost	Quantile

FIGURE 4.1 – Choix de la méthode de discrétisation pour chaque variable quantitative

Par exemple pour la garantie RCM sur le coût moyen, pour la variable sur le poids à vide provenant de la base SRA la variable avant discrétisation avait cette allure :

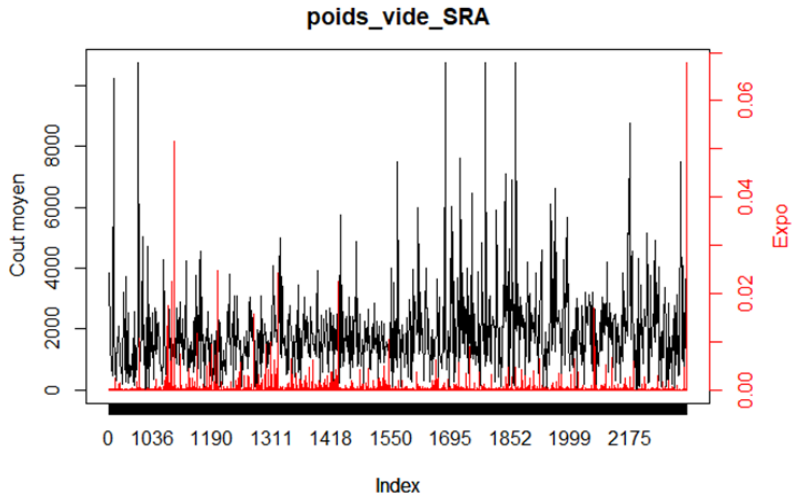


FIGURE 4.2 – Poids à vide du véhicule avant discrétisation

Après la discrétisation utilisant la méthode XGBoost :

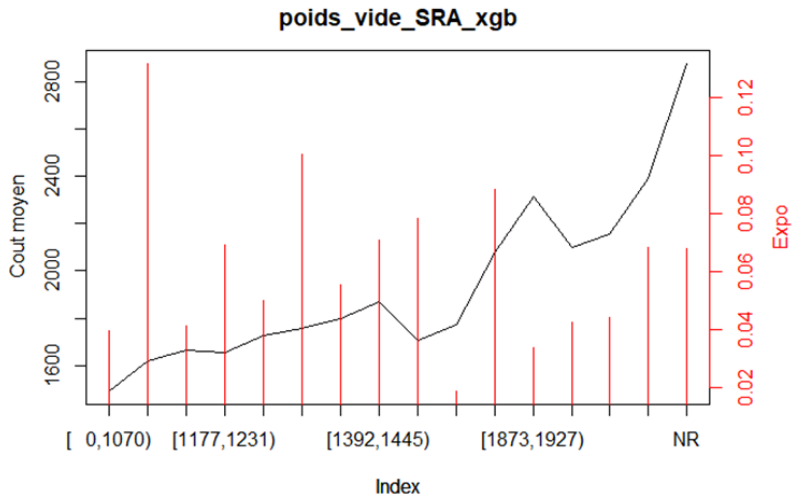


FIGURE 4.3 – Poids à vide du véhicule après discrétisation utilisant la méthode XGBoost

Et après la discrétisation utilisant la méthode des quantiles :

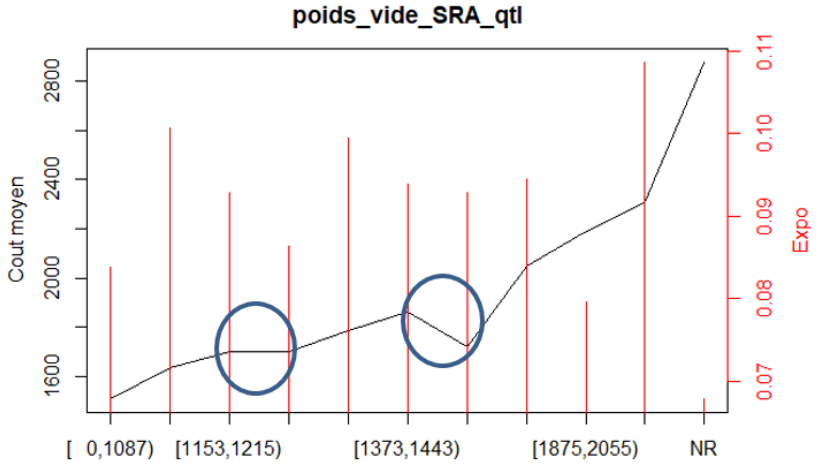


FIGURE 4.4 – Poids à vide du véhicule après discrétisation utilisant la méthode des quantiles

La discrétisation utilisant la méthode des quantiles semble être la plus adéquate. En effet, l'exposition est stable sur la plupart des tranches et le coût moyen varie pour chaque tranche, avec un coût moyen qui augmente plus le poids à vide est important, ce qui est cohérent. Il faut cependant observer que certaines tranches n'ont pas de variations (premier rond sur la gauche du graphique) et, pour d'autres, il y a une baisse du coût moyen (deuxième rond au milieu du graphique). Il n'est pas cohérent que pour certains véhicules, le coût moyen soit moins élevé alors que le poids à vide est plus élevé. En prenant pour exemple un véhicule ayant un poids à vide de 1400 kilogrammes et un autre avec un poids à vide de 1500 kg, il n'est pas logique que le deuxième véhicule ait un coût moyen plus faible que le premier et donc éventuellement un tarif plus faible. Ainsi, il a été décidé de regrouper les tranches entourées sur le graphique afin que la discrétisation soit cohérente et plus compréhensible et facile à analyser. Le graphique de la discrétisation de la variable du poids à vide provenant de la base SRA est finalement celui ci-dessous :

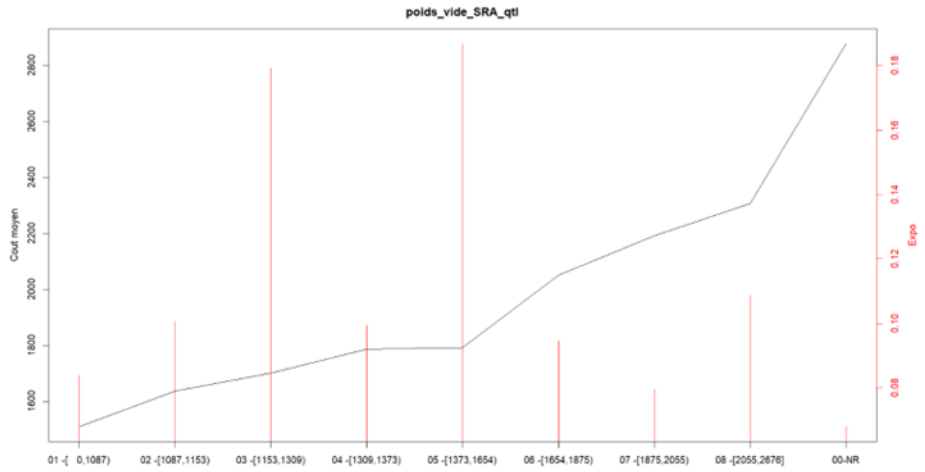


FIGURE 4.5 – Discrétisation finale du poids à vide SRA utilisant la méthode des quantiles

Le regroupement est la technique de retraitement qui a été utilisé après la discrétisation utilisant la méthode des quantiles. Aussi appelé agrégation ou consolidation, c'est une technique qui consiste à combiner des modalités similaires d'une variable en une seule catégorie. Cette approche vise à réduire le nombre de modalités tout en conservant les informations essentielles afin de simplifier les données et les rendre plus compréhensibles.

4.3 Présentation de la base finale

Finalement, plusieurs variables feront partis des différentes bases de modélisation, notamment plusieurs variables SRA. Les différentes variables sont les suivantes :

Variables contrats / Variables Véhicules	Variables SIV	Variables SRA
Age du véhicule	Carrosserie	Boite vitesse
Année	Cylindrée	Classe de prix
Catégorie x Tonnage	Genre	Classe réparation
Code genre	Poids à vide	Energie
Code NAF	Puissance fiscale	Genre
Marque	Puissance maximale	Groupe SRA
Nombre de véhicule à l'affaire nouvelle	Taux de co2	Marque
	Vitesse du moteur	Poids à vide
		Prix actuel
		Prix d'origine
+ Tranche distance moyenne des sinistres (expliquée dans le Chapitre 5)		PTAC
		Segment
		Transmission
		Type Alimentation

FIGURE 4.6 – Variables disponible dans les bases de modélisation

Il est possible de remarquer que certaines variables sont redondantes entre les variables SIV et SRA ainsi que les variables contrats/véhicules et SRA. Par exemple, la marque véhicule et la marque SRA, le poids à vide SIV et SRA ou encore le genre SIV et le genre SRA. Il faudra ainsi prêter attention à ces variables afin de voir lesquelles sont les plus importantes lors de la sélection des variables (cf. Chapitre 6).

4.4 Études des variables

Parmi les différentes études à faire avant de passer à la modélisation de la prime pure, il y a l'étude des valeurs extrêmes. En effet, la théorie des valeurs extrêmes s'applique dans de nombreux domaines, y compris la détermination du seuil des sinistres grave. Le seuil d'écrêtement fait référence) un niveau au-delà duquel les sinistres sont considérés comme extrêmes, aussi appelé sinistres graves. Dans le cadre de ce mémoire, la théorie des valeurs extrêmes permettra la détermination de ce seuil.

La théorie des valeurs extrêmes

Cadre mathématique :

Avant de déterminer le seuil, il faut tout d'abord étudier la loi du maximum de la distribution. On a $X_1, \ldots, X_n$ une suite de variables aléatoires indépendantes et identiquement distribuées. On note $M_n = \max(X_1, \ldots, X_n)$ et $F_X(x) = P(X \leq x)$. On a ainsi la loi de M_n :

$$P(M_n \leq x) = P(\max(X_1, \ldots, X_n) \leq x) = P(X_1 \leq x, \ldots, X_n \leq x) = \prod_{i=1}^{n} P(X_i \leq x) = [F_X(x)]^n$$

Cependant, nous ne connaissons pas F et nous souhaitons étudier le comportement asymptotique de la distribution. On peut montrer que la loi limite de la suite des maxima renormalisés appartient nécessairement à une des trois familles de lois suivantes : Loi de Weibull, loi de Gumbel et loi de Fréchet.

Théorème de Fisher-Tippett
Soit $(X_i)_{i\geq 1}$ une suite de variables i.i.d. On suppose qu'il existe des suites (a_n) avec $a_n > 0$ et (b_n) telles que $\frac{M_n - b_n}{a_n}$ converge en loi vers une variable Y de fonction de répartition H. Alors H (ou $\tilde{H}(x) = H\left(\frac{x-b}{a}\right)$) appartient à une des trois familles caractérisées par les fonctions de répartition suivantes :

$$\text{Weibull} : \psi_\alpha(x) = \begin{cases} e^{-(-x)^\alpha}, & x \leq 0, \alpha > 0, \\ 1, & x > 0, \end{cases}$$

$$\text{Gumbel} : \Lambda(x) = e^{-e^{-x}}, \quad x \in \mathbb{R},$$

$$\text{Fréchet} : \Phi_\alpha(x) = \begin{cases} 0, & x \leq 0, \\ e^{-(x^{-\alpha})}, & x > 0, \alpha > 0. \end{cases}$$

Ces trois lois ont une fonction de répartition de la forme (distribution GEV, Generalized Extreme Value) suivante :

$$H_\xi(x) = \begin{cases} e^{-(1+\xi x)^{-1/\xi}}, & \xi \neq 0, \\ e^{-e^{-x}}, & \xi = 0. \end{cases}$$

Où ξ est l'indice de valeurs extrêmes, le paramètre de forme, également appelé paramètre de queue.

On a alors :

$$\begin{aligned}
\text{Weibull } \xi &= -\frac{1}{\alpha} < 0 \\
\psi_\alpha(x) &= H_{-\frac{1}{\alpha}}(\alpha(x+1)) \\
\text{Gumbel } \xi &= 0 \\
\Lambda(x) &= H_0(x) \\
\text{Fréchet } \xi &= \frac{1}{\alpha} > 0 \\
\Phi_\alpha(x) &= H_{\frac{1}{\alpha}}(\alpha(x-1))
\end{aligned}$$

Le signe de ξ permet d'identifier la famille de loi.

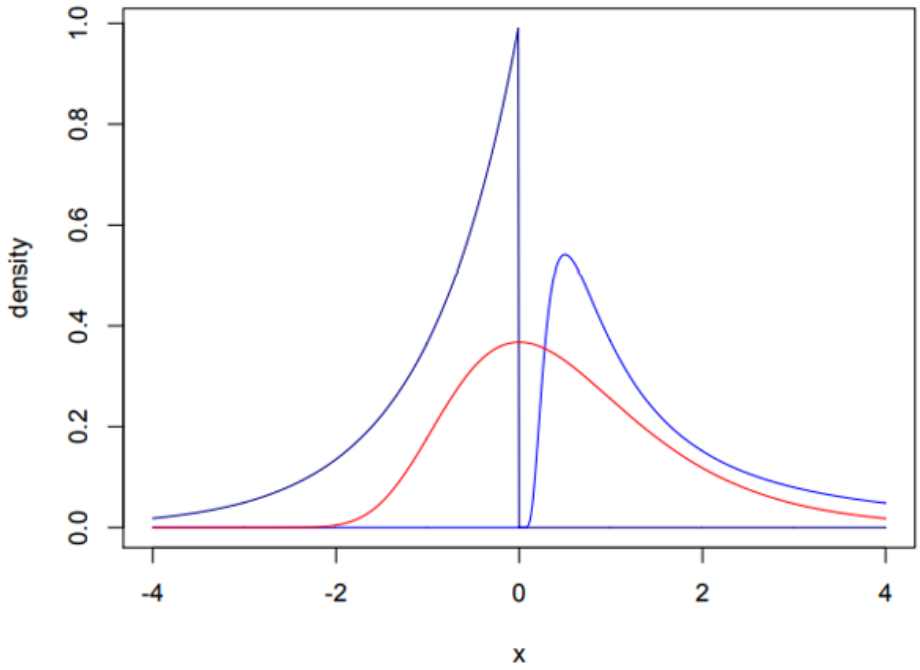


FIGURE 4.7 – Densité GEV : Weibull ($\xi = -1$, darkblue), Gumbel (red) et Fréchet ($\xi = 1$, blue)

Une distribution GPD, Generalized Pareto Distribution, est caractérisée par la fonction de répartition

$$G_\xi(x) = \begin{cases} 1 - (1+\xi x)^{-1/\xi}, & \text{si } \xi \neq 0,\ 0 < x < x^*, \\ 1 - e^{-x}, & \text{si } \xi = 0,\ x > 0, \end{cases}$$

Où $x^* = +\infty$ si $\xi > 0$ et $x^* = -1/\xi$ si $\xi < 0$.

Si X suit une loi G_ξ, alors βX suit la loi $G_{(\xi,\beta)}$ avec

$$G_{\xi,\beta}(x) = G_\xi\left(\frac{x}{\beta}\right)$$

Théorème : F appartient au domaine d'attraction de H_ξ si et seulement s'il existe une fonction a strictement positive telle que pour $x > 0$

$$\lim_{u \to x_F} \mathbb{P}\left(\frac{X-u}{a(u)} > x | X > u\right) = \begin{cases} (1+\xi x)^{-1/\xi}, & \text{si } \xi \neq 0, \\ e^{-x}, & \text{si } \xi = 0 \end{cases}$$

Si on s'intéresse au dépassement du seuil, POT peak other threshold, pour un niveau de seuil u élevé, le théorème précédent montre que, si F la fonction de réparation de X appartient au domaine d'attraction de H_ξ alors la distribution $X - u | X > u$ est proche d'une distribution $G_{\xi,a(u)}$.

On note $e(u)$ la moyenne du dépassement de seuil, aussi appelé mean excess function :

$$e(u) = \mathbb{E}(X - u | X > u)$$

Proposition : Si X suit une loi GPD $G_{\xi,a}$, avec $\xi < 1$ alors pour $u < x_F$, on a

$$e(u) = \frac{a + \xi u}{1 - \xi}$$

Mean excess plot : En pratique pour identifier a et ξ, l'idée est de tracer le graphe $\big(u, e(u)\big)$ et de voir à partir de quel seuil ce graphe devient linéaire. On estime la fonction $e(u)$ par

$$e_n(u) = \frac{\sum_{i=1}^n (X_i - u) 1_{X_i > u}}{\sum_{i=1}^n 1_{X_i > u}}$$

et on trace le graphe $\big(X_{(n-k,n)}, e(X_{(n-k,n)})\big)$, avec $0 \leq k \leq n-1$ sachant que :

$$e(X_{(n-k,n)}) = \frac{1}{k} \sum_{j=1}^{k} \big(X_{(n_j+1,n)} - X_{(n-k,n)}\big)$$

Pour rappel, on appelle statistique d'ordre, l'échantillon ordonné en croissant

$$X_{(1,n)} \leq ... \leq X_{(i,n)} \leq ... \leq X_{(n,n)}$$

avec $X_{(1,n)} = \min_{1 \leq i \leq n} X_i$ et $X_{(n,n)} = \max_{1 \leq i \leq n} X_i$.

On peut ainsi choisir le seuil u à partir du graphe "mean excess plot". Des autres méthodes peuvent être utiliser pour estimer le paramètre ξ :

L'estimateur de Pickands est une méthode qui permet d'estimer le paramètre ξ. Soient $(X_i)_{1\leq i\leq n}$ des variables i.i.d. de loi F, l'estimateur de Pickands est défini à partir de la statistique d'ordre par

$$\hat{\xi}^P_{(k(n),n)} = \frac{1}{\log(2)} \log\left(\frac{X_{(n-k(n+1,n)} - X_{(n-2k(n)+1,n)}}{X_{(n-2k(n)+1,n)} - X_{(n-4k(n)+1,n)}}\right)$$

où $k(n)$ est un entier tel que $1 \leq k(n) \leq n/4$ (cf. Annexe B 7.2).

L'estimateur de Hill est une autre méthode qui permet d'estimer le paramètre ξ. Soient $(X_i)_{1\leq i\leq n}$ des variables i.i.d. de loi F, l'estimateur de Hill est défini à partir de la statistique d'ordre par

$$\hat{\xi}^H_{(k(n),n)} = \frac{1}{k(n)} \sum_{i=n-k(n)+1}^{n} (\log(X_{(i,n)}) - \log(X_{(n-k(n)+1,n)}, 1 \leq (n) \leq n$$

En pratique, pour les deux estimateurs de Hill et Pickands, il est conseillé de tracer la courbe $k \to \hat{\xi}_{(k,n)}$ pour avoir une idée du choix de k, c'est-à-dire la valeur où la courbe est horizontale (cf. Annexe C 7.2).

La méthode des quantiles, n'est pas un estimateur, mais une approche couramment utilisée pour estimer le seuil d'écrêtement dans la théorie des valeurs extrêmes. Cette méthode est basée sur l'idée de choisir un seuil au niveau d'un certain quantile élevé de la distribution des données de telle façon que seuls les sinistres extrêmes soient au-delà de ce seuil. En général, le niveau de quantile est élevé, aux alentours du 95e, 99e quantile. Il est également possible de déterminer grâce à la méthode des quantiles le taux de grave en moyenne de chaque contrat à partir duquel les sinistres sont considérés comme extrêmes, comme par exemple 2% de sinistres graves.

Dans le cadre de ce mémoire, il a été décidé de vérifier si les seuils choisis par l'équipe APAT en 2022 pour la modélisation du coût moyen des garanties RCM et RCC sont cohérents. Pour la garantie RCM, le seuil d'écrêtement a été fixé à 10 000 € et pour la garantie RCC, à 50 000 €. Il sera expliqué la vérification du seuil pour la garantie RCM ci-dessous.
Pour l'estimateur de Pickands, celui-ci a été calculé manuellement sur R Studio ainsi que grâce à la fonction *pickandsplot* présente sur le logiciel. Le seuil à partir duquel la courbe $k \to \hat{\xi}_{(k,n)}$ se stabilise se situe autour de $k = 1500$, ce qui indique un seuil à 4538.75. Pour la fonction R, le seuil u s'établit à 4300, avec $\xi = 0.16$.

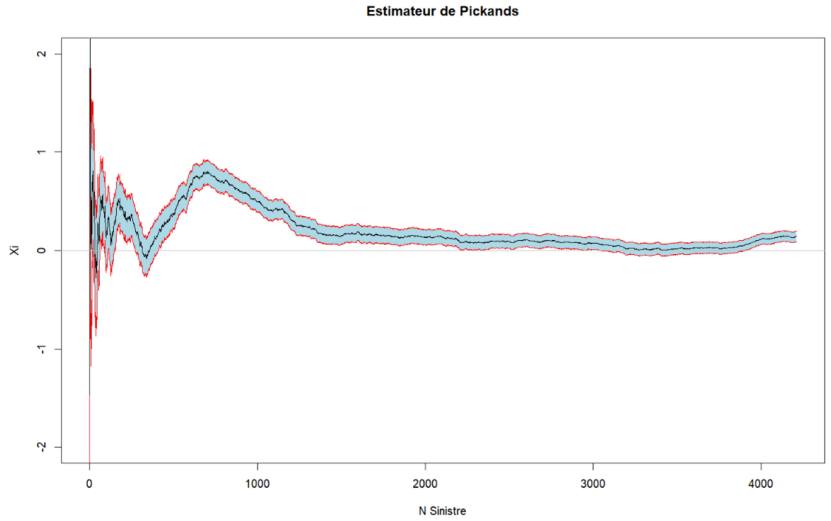


FIGURE 4.8 – Estimateur de Pickands, construit manuellement

Concernant l'estimateur de Hill, la fonction *hillplot* est également disponible sur R Studio et indique également un estimateur à 4300.

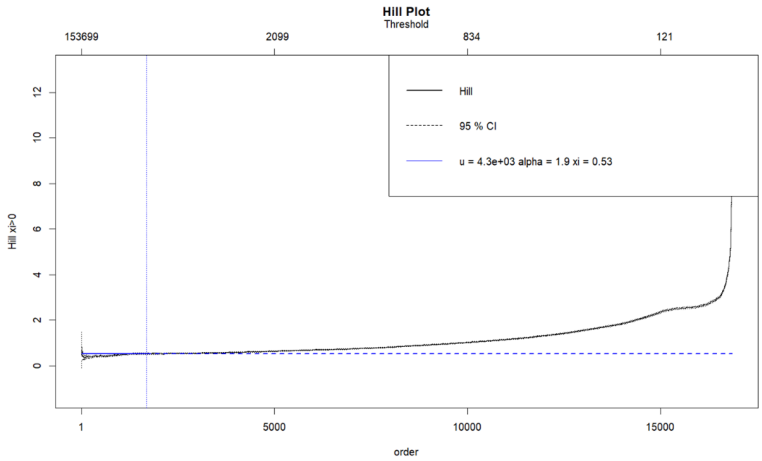


FIGURE 4.9 – Estimateur de Hill, construit grâce à la fonction *hillplot*

Il faut prendre les résultats de ces estimateurs avec précautions. En effet, pour l'estimateur de Hill, les queues de distribution sont lourdes en principe, et peut ne pas bien fonctionner pour des distributions avec des queues légères. De plus, étant donné qu'il n'y a pas de règles strictes pour déterminer le seuil, le choix du seuil peut être sensible et biaisé ou inefficace. C'est pourquoi il est judicieux de regarder d'autres méthodes d'estimation.

Tout comme pour les estimateurs de Hill et de Pickands, la fonction moyenne des

excès, mean excess function, peut être tracée sur une graphique grâce à la fonction *MeanExcessPlot* mais peut également être tracé manuellement sur le logiciel R Studio. Dans ce cas, le seuil à partir duquel le graphique montre une tendance à la stabilisation serait autour de 15 000.

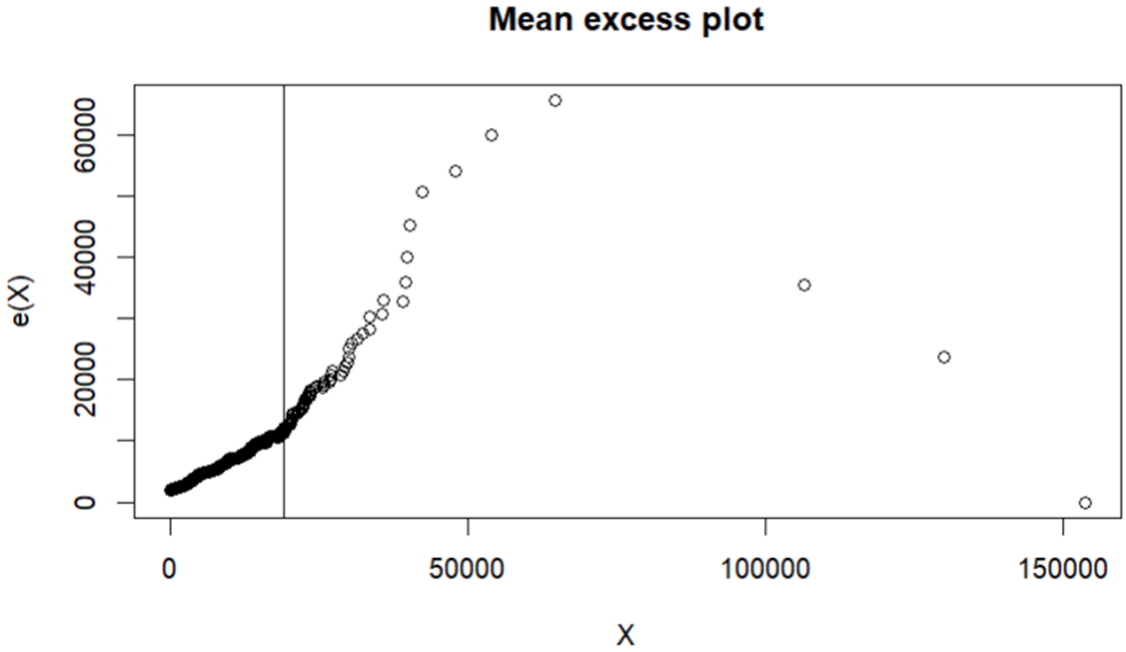


FIGURE 4.10 – Mean Excess Plot, construit manuellement

Enfin, la méthode des quantiles a été appliquée également pour la détermination du seuil d'écrêtement. Le seuil est déterminé à partir des quantiles 98%, 98,2%, 98,4%, 98,6%, 98,8% et 99%. Il est également regardé le nombre de sinistres grave, la somme totale des charges graves, la moyenne des charges graves, le taux de grave moyen par contrat, le pourcentage du nombre de grave et le pourcentage de charge grave. Les résultats pour les quantiles cités précédemment sont ainsi les suivants :

	seuil	nb	chg	Charge_moyenne	Taux_grave_contrat	Nb_seuil	Charge_seuil
1	10000	335	5689015	16982.13	0.020191756	0.019871871	0.1710313
2	10500	311	5444115	17505.19	0.019057579	0.018448214	0.1636688
3	11500	255	4827643	18931.93	0.015604733	0.015126350	0.1451355
4	12000	231	4544795	19674.44	0.014598814	0.013702693	0.1366321
5	12500	212	4312808	20343.43	0.012859878	0.012575632	0.1296578
6	13500	167	3726475	22314.22	0.009976996	0.009906276	0.1120306

FIGURE 4.11 – Seuils déterminé avec la méthode des quantiles

Finalement, au vu des différents résultats, le seuil d'écrêtement déterminé par l'équipe APAT en 2022 de 10 000 pour le coût moyen des sinistres de la garantie RCM est cohé-

rent au vu des résultats des différentes méthodes utilisées précédemment.

Une autre étude primordiale qui doit être réalisée avant la modélisation de la prime pure est la détermination de la loi de la variable réponse. Cette étude vise à comprendre la distribution des valeurs de la variable réponse, ce qui est crucial pour modéliser correctement les sinistres.

Concernant le coût moyen des sinistres, différentes lois peuvent être utilisées :

- La loi Gamma : souvent utilisée pour modéliser les variables continues positives, comme le coût des sinistres. Elle est adaptée lorsque la distribution des coûts est asymétrique et possède une queue lourde.
- La loi Log-Normale : Si les logarithmes des coûts suivent une distribution normale, alors la loi log-normale peut être utilisée pour modéliser le coût des sinistres. Elle est utilisée lorsque les coûts sont positifs et ont une distribution asymétrique.
- La loi Weibull est couramment utilisée pour modéliser des phénomènes de durée de vie ainsi que des coûts de sinistres. Elle est utile pour les distributions ayant des queues lourdes et des caractéristiques de croissance.

Par exemple, pour le coût moyen des sinistres de la garantie RCM, ces trois lois ont été testées :

Graphiquement, la loi log-normale qui semble le moins coller à la distribution du coût moyen par rapport aux autres lois. Afin de savoir quelle loi évalue le mieux la distribution du coût moyen, le test de Kolmogorov-Smirnov (KS test) a été utilisé. C'est un test statistique non paramétrique utilisé pour évaluer si un échantillon de données suit une distribution particulière. Il est largement utilisé pour déterminer si une distribution empirique (basée sur les données observées) diffère significativement d'une distribution théorique (ici, la distribution Gamma, Log-normale ou Weibull). La statistique de test Kolmogorov-Smirnov est basée sur la plus grande différence verticale entre les deux fonctions de répartition. Autrement dit, la statistique de Kolmogorov-Smirnov est la suivante :

$$D_n = \sup_x |F_n(x) - F(x)|$$

Ainsi, après avoir appliqué le test de Kolmogorov-Smirnov, il a fallu regarder la statistique pour chacun des tests pour chacun des lois. La statistique de la loi Gamma est de 0.0408708, pour la loi log-normale, elle est de 0.07290307 et pour la loi Weibull, elle est de 0.5. Ainsi, la loi Gamma est la loi candidate pour la modélisation du coût moyen de la garantie RCM. C'est également cette loi pour le coût moyen des sinistres de la garantie RCC.

Pour la fréquence des sinistres forfaitaires et non forfaitaires de la garantie RCM, ainsi que pour la fréquence des sinistre la garantie RCC, la loi retenue est la loi Poisson. Les lois qui ont été testés pour la fréquence sont les suivantes :

- Loi Poisson : c'est l'une des distributions les plus couramment utilisées pour modéliser la fréquence des évènements rares et indépendants.

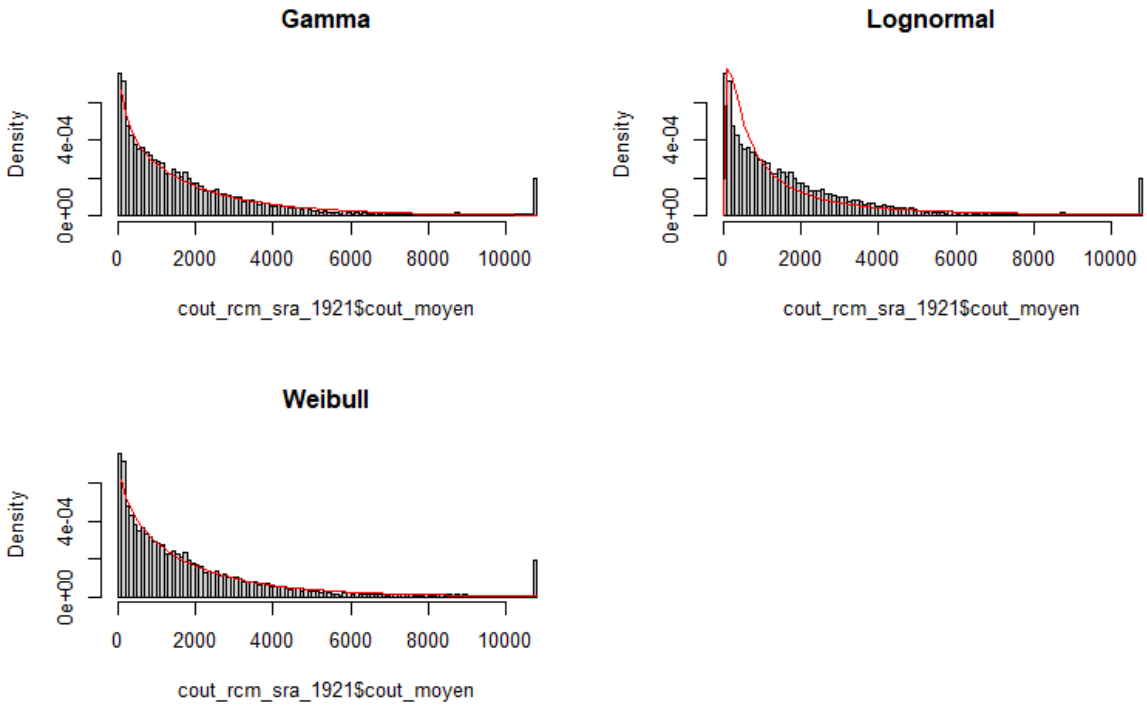


FIGURE 4.12 – Loi du coût moyen de la garantie RCM

- Loi Binomiale Négative : elle est utilisée pour modéliser les sinistres qui surviennent avec une variabilité plus élevée que celle attendue dans une distribution de Poisson. Cette loi est choisie lorsque la variance est plus forte que la moyenne, les observations sont sur-dispersées par rapport à la loi de Poisson.

Une dernière étape avant de passer à la modélisation est de regarder pour chaque nouvelle variable ajoutée à la base de modélisation (provenant de la base SRA) l'exposition et le coût moyen (ou la fréquence) afin d'avoir une idée de la tendance et des caractéristiques de chacune des variables. Par exemple, pour la garantie RCM pour la variable *SEGMENT_SRA* qui représente le segment automobile, il s'agit des différentes catégories de voiture que les constructeurs proposent sur le marché, les graphiques du coût moyen et de la fréquence des sinistres sont les suivants :

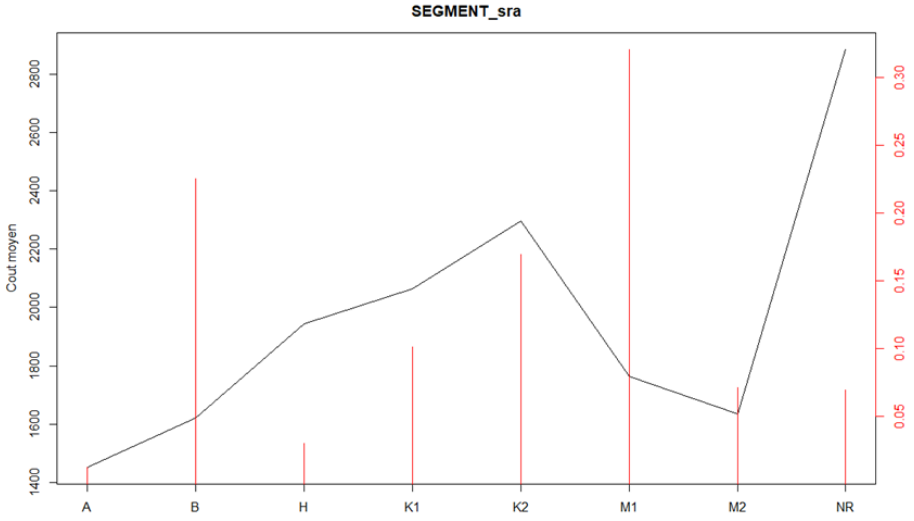


FIGURE 4.13 – Coût moyen des sinistres de la garantie RCM

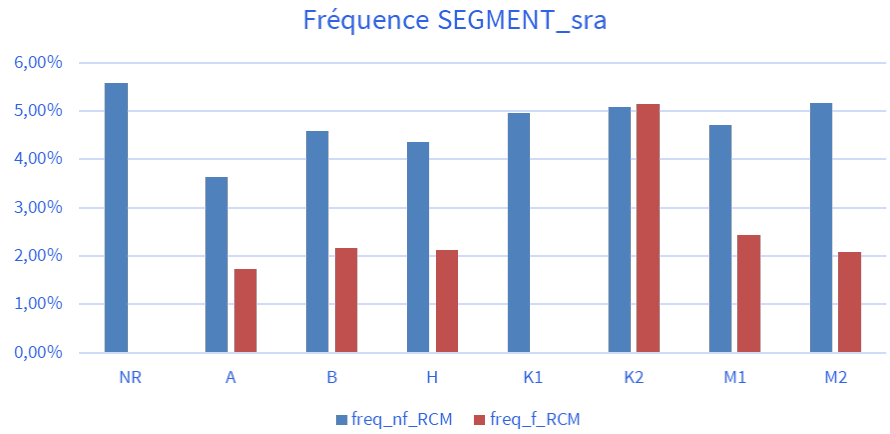


FIGURE 4.14 – Fréquence des sinistres forfaitaires et non forfaitaires de la garantie RCM

Modalités	Description
A	Petites citadines
B	Urbaines
M1	Familiales compactes
M2	Familiales
H	Haut de gamme
K1	Petits fourgons / Fourgonnettes
K2	Fourgons
NR	Unknown

TABLE 4.1 – Définition de la variable *SEGMENT_SRA*

Des variations sont observées entre chaque modalité, avec une exposition globalement du même ordre. Cette variable pourrait ainsi être intéressante dans la modélisation, que ce soit pour le coût moyen des sinistres, pour la fréquence des sinistres forfaitaires ou non forfaitaires.

A l'inverse, la variable *modele_SRA* ne paraît pas intéressante pour la modélisation du fait du nombre trop important de modalités :

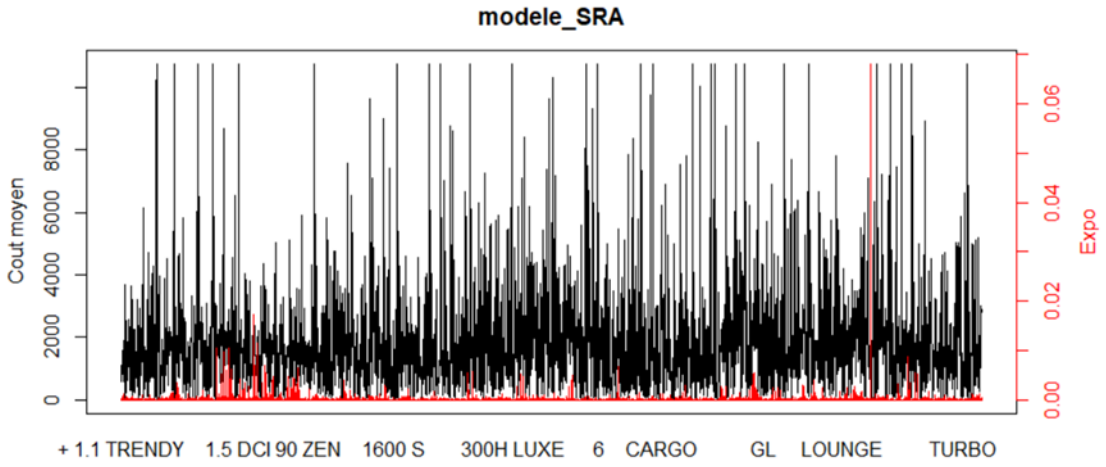


FIGURE 4.15 – Coût moyen des sinistres de la variable concernant le modèle des véhicules SRA

Chapitre 5

Étude sur la géolocalisation des sinistres

5.1 Contexte et objectifs

Une étude sur la géolocalisation des sinistres a été menée afin d'en savoir plus sur l'activité de l'entreprise, à savoir si ses véhicules circulent de manière locale, départementale, régional, nationale, voire même international. Cette étude a pour principal objectif de voir si cette information pourrait être intéressante dans le cadre de la modélisation des garanties RCM et RCC, c'est-à-dire si elle pourrait affiner la tarification.

Il serait ainsi intéressant de regarder pour la garantie RCM, le coût moyen, la fréquence de sinistres forfaitaires et hors forfaits, et pour la garantie RCC, le coût moyen et la fréquence, si la variable contenant l'information sur la distance moyenne entre l'adresse du souscripteur et les lieux de sinistres.

5.2 Étude sur la distance lieu de sinistres / adresse souscripteur

Pour mener cette étude, il a fallu récupérer les adresses des lieux de sinistres et les adresses des souscripteurs. Ces informations sont disponibles sur des bases AXA France grâce aux différents outils de gestion de sinistres et de souscription.

Pour l'adresse des sinistres, il a fallu regarder le taux de manquants sur la base de donnée 2019-2021. Certains sinistres n'ont aucune information sur le lieu de sinistres, et d'autres ont une information comme « 99999 », « 99999 NON PRECISE », « NC », « ILLISIBLE », ou encore « ILLISIBLE SUR LE CONSTAT ». Ces informations sont considérées comme des informations manquantes, il faut donc les compter dans le taux de manquants. Pour les variables contenant des informations sur le lieu des sinistres, trois variables sont présentes, à savoir le code postal du sinistre, l'adresse du sinistre et

la ville du sinistre. Cependant, il se peut que certains sinistres n'aient qu'une seule des trois informations. Une variable a alors été créée qui regroupe ces trois variables et c'est donc sur celle-ci que le taux de manquants est calculé. Le taux de manquants est calculé pour chaque contrat et est le rapport entre le nombre de sinistres ayant une information sur la localisation des sinistres et le nombre total des sinistres du contrat, sur la période de 2019 à 2021.

$$\text{Taux manquants} = \frac{\text{Nombre de sinistres avec une information sur le lieu de sinistre}}{\text{Nombre de sinistres du contrat}}$$

Des tranches de taux de manquants ont alors été créées allant de 0 à 100, de 10% en 10% avec 0 et 100 isolés, afin de savoir le pourcentage de contrat par tranche.

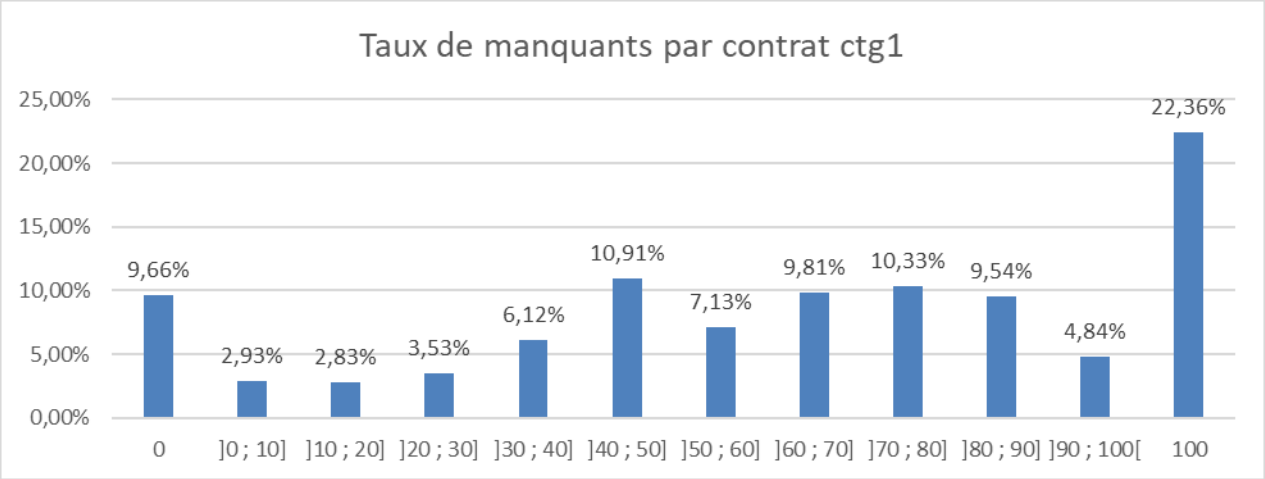


FIGURE 5.1 – Taux de manquants par contrat sur la localisation des sinistres

On peut voir dans le graphique ci-dessous que 22,36% des contrats de la base FND catégorie 1 des années 2019 à 2021 ont un taux de manquants de 100%, soit aucune information sur le lieu de tous leurs sinistres. Ce taux est assez élevé, mais il y a presque 80% des contrats qui ont au moins un lieu de sinistre.

Par la suite, il faut récupérer l'adresse des entreprises (des souscripteurs). Contrairement au lieu de sinistres qui peut ne pas être renseigné, ou être mal renseigné, dans les outils de gestions, l'adresse du souscripteur doit être renseignée. Il a été possible d'obtenir l'adresse grâce à deux bases de données différentes qui à elles deux permettent d'avoir toutes les adresses des souscripteurs.

La prochaine étape consiste à obtenir les coordonnées GPS (x,y) des lieux de sinistres et des adresses des entreprises. Pour ce faire, deux sites ont été utilisés, à savoir un site interne à AXA, où l'on peut y déposer un fichier en format CSV et récupérer ce

même fichier avec les informations des coordonnées GPS, et un site du gouvernement (`https://adresse.data.gouv.fr/csv`), où le principe est le même. Cependant, certaines adresses comportaient des codes CEDEX qui n'étaient pas retrouvées sur ces sites, et d'autres se situait en Corse. Il a ainsi été décidé d'ajouter des bases de données provenant de l'extérieur, notamment de l'INSEE, afin d'avoir des informations supplémentaires :

— Une base comportant les codes CEDEX avec les coordonnées GPS

— Une base comportant les codes postaux, avec coordonnées GPS

— Une base associant chaque département à sa région

Une fois ces étapes réalisées, il faut calculer la distance à vol d'oiseau, en kilomètres, entre les lieux de sinistres et les adresses des entreprises. Pour ce faire, la formule utilisée est la suivante :

$$\text{Distance} = \text{R} \times \arccos[\sin(a)\sin(b) + \cos(a)\cos(b)\cos(c-d)]$$

Avec R le rayon de la terre (6 371 km), a : la latitude du point A, b : la latitude du point B, c : la longitude du point A et d la longitude du point B. À noter : les latitudes et longitudes doivent être en radians. Pour passer de degré en radian, la formule suivante est utilisée :

$$X^\circ = \frac{X^\circ \times \pi}{180^\circ}\text{radians}$$

Une fois la distance calculée pour chaque sinistre, pour chaque contrat, la moyenne des distances de chaque sinistre a été calculée afin de savoir à quelle distance les véhicules d'un contrat se déplacent (et ont des sinistres) en moyenne. Il y a alors pour chaque contrat de la base 2019 à 2021, une valeur égale à la distance moyenne des sinistres. De plus, il a été ajouté un topage si l'adresse de l'entreprise se situe en Île-de-France, étant donné la particularité de la région qui est assez vaste et contient une très grande population et donc un nombre important de véhicules circulant. Ainsi, pour chaque contrat, il y a une distance moyenne de sinistres avec un topage 1 si l'entreprise est en Île-de-France et 0 sinon. Grâce à cela, il est possible de construire une variable avec différentes tranches de distance moyenne : <150 km, [150, 300 km] et >300 km ; avec, pour chacune des tranches, un topage IDF (Île-de-France) et une modalité « International » qui représente les sinistres à plus de 1000 km. Cette nouvelle variable sera nommée "Tranche_dist_moy.

Par la suite, une analyse pour le coût moyen RCM, RCC et la fréquence RCC, RCM hors forfait et forfait a été réalisée afin de voir si d'une tranche à une autre le coût moyen ou la fréquence varie et donc si cette variable pourrait être intéressante dans la modélisation.

5.3 Résultats et conclusion

Les résultats sont l'étude et de la création de la variable Tranche_dist_moy sont les suivants :

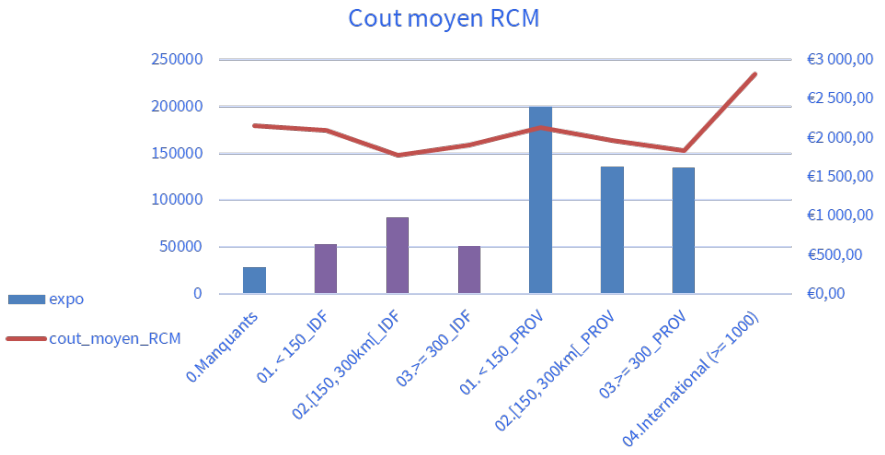


FIGURE 5.2 – Coût moyen des sinistres sur la garantie RCM en fonction de la variable Tranche_dist_moy

Pour le cas ci-dessus, il y a une exposition plus importante hors Île-de-France, mais il n'y a pas de différences entre le coût moyen en Île-de-France ou Province, que ce soit pour les sinistres en moyenne proche de l'entreprise ou ceux plus loin. La variable Tranche_dist_moy ne semble pas intéressante ici.

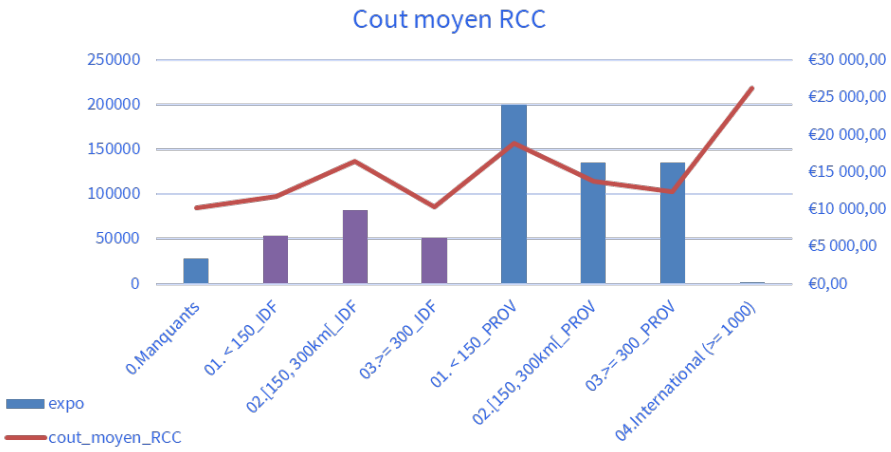


FIGURE 5.3 – Coût moyen des sinistres garantie RCC en fonction de la variable Tranche_dist_moy

Pour la garantie RCC, la variable pourrait être intéressante dans la modélisation du coût moyen. En effet, il y a plus de variations d'une tranche à une autre et la tendance n'est pas la même entre la région Île-de-France, où il y a une hausse entre les deux premières tranches puis une baisse entre les deux secondes, et la Province, où le coût moyen est plus faible plus la distance moyenne des sinistres est grande.

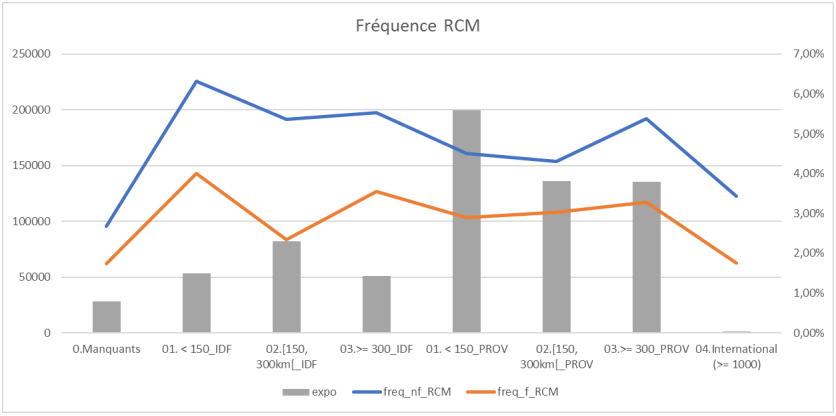


FIGURE 5.4 – Fréquence des sinistres forfaitaires (f) et hors forfait (nf) - garantie RCC en fonction de la variable Tranche_dist_moy

La variable Tranche_dist_moy parait intéressante pour la modélisation de la fréquence des sinistres forfaitaires et hors forfaits de la garantie RCC. En effet, d'une tranche à une autre, les fréquences varient, ce qui indique que la variable pourrait avoir un impact significatif dans la modélisation.

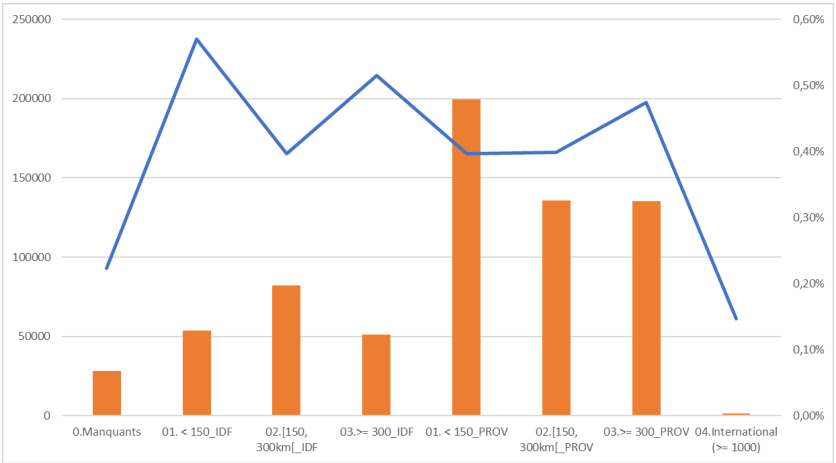


FIGURE 5.5 – Fréquence des sinistres - garantie RCC en fonction de la variable Tranche_dist_moy

Tout comme pour la fréquence de sinistres forfaitaires et non forfaitaires de la garantie RCM, la fréquence de sinistres de la garantie RCC varient pour chaque tranche de distance moyenne, exceptés pour la première et deuxième tranche de distance pour la Province. La variable pourrait donc avoir une importance lors de la modélisation.
Pour résumer, les choix qui ont été faits pour les différentes modélisations sont les suivants :

Garanties	Fréquence	Coût moyen
RCC	Oui	Non
RCM forfaitaire	Oui	Oui
RCM hors forfait	Oui	

TABLE 5.1 – Récapitulatif des choix de prise en compte de la variable Tranche_dist_moy

A noter : cette variable est construite en fonction des adresses de sinistres ce qui peut provoquer un biais sur la modélisation de la prime pure étant donné que les contrats n'ayant pas de sinistres auront ainsi la valeur "0.Manquants", ce qui peut ainsi introduire un biais. Cependant, cette variable reste très importante afin d'avoir une première approche sur l'introduction d'une variable ayant une indication sur la circulation de la flotte. En effet, si celle-ci ressort des modèles il faudrait utiliser d'autres variables, qui peuvent être issues des bases contrats, véhicule, SIV ou bien SRA, afin de pouvoir prédire les tranches de la variable *Tranche_dist_moy*. Cette variable pourra par la suite évoluer vers de nouvelles modalités, comme par exemple "circulation locale", "circulation départementale", "circulation régionale", "circulation nationale" ou encore "circulation internationale".

Chapitre 6

Théorie - Modélisation de la prime pure

6.1 Les modèles de tarification

Les modèles de tarification en assurance automobile sont en général des modèles de Fréquence × Coût moyen ou bien des modèles de Prime pure. L'objectif de ces modèles est d'ajuster au mieux la prime pure. La prime pure fait référence à la partie du montant total d'une prime d'assurance qui est destinée à couvrir les risques associés à une police d'assurance. Elle est basée sur l'espérance du coût des sinistres. Elle est calculée en tenant compte de la moyenne attendue des coûts des sinistres. C'est le reflet direct du risque que l'assureur prend en charge en fournissant une couverture à l'assuré.

Le modèle Fréquence × Coût Moyen est le modèle qui est retenu dans le cadre de la tarification des garanties responsabilité civile matérielle (RCM) et corporelle (RCC). En effet, ce modèle permet de prendre en compte à la fois la fréquence des sinistres et le coût moyen de chaque sinistre, cela est utile dans le cas où les sinistres peuvent varier considérablement en termes de coût. De plus, ce type de modèle permet une meilleure adaptation aux différents profils de risque.

La prime pure d'un contrat (noté PP) se calcule selon la formule suivante :

$$\text{PP} = \text{Coût moyen (noté CM)} \times \text{Fréquence}$$

Soit N une variable aléatoire à valeur dans N représentant le nombre total des sinistres, $Y = (Y_i)_{i>0}$ une suite de variables aléatoires à valeurs dans $\mathbb{R}^+$ telle que pour tout i, Y_i représente le coût du sinistre i. On a alors la charge totale de sinistres qui s'écrit de la manière suivante :

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

On note les hypothèses suivantes :

— Pour tout i>0, les variables aléatoires N et Y_i sont supposées indépendantes

— La suite de variables aléatoires $Y = (Y_i)_{i>0}$ est supposée i.i.d. (indépendante et identiquement distribuée)

Il est ainsi possible de montrer que :

$$\mathbb{E}(S) = \mathbb{E}(\sum_{i=1}^{N} Y_i) = \mathbb{E}(Y_1)\mathbb{E}(N)$$

Où $\mathbb{E}(Y_1)$ représente le coût moyen et $\mathbb{E}(N)$ représente la fréquence.

Les hypothèses précédentes sont de fortes hypothèses qui peuvent être contestées, car en pratique l'indépendance entre les coûts moyen et la fréquence des sinistres n'est pas toujours vérifiée, mais cela est négligeable par rapport aux incertitudes de la tarification.

Le coût moyen des sinistres représente la moyenne des montants payés pour chaque sinistre survenu dans la période considérée (notée P). Le coût moyen peut varier en fonction de facteurs tels que la gravité des dommages, le type de véhicule impliqué et les coûts de réparation. Il se calcule de la manière suivante :

$$\text{Coût moyen période P} = \frac{\sum_{\text{en montant}} \text{sinistres période P}}{\sum_{\text{en nombre}} \text{sinistres période P}}$$

La fréquence des sinistres correspond au nombre total de sinistres enregistrés au cours de la période considérée (P). Cette fréquence est généralement exprimée en pourcentage ou en nombre par unité de temps (par exemple 1 an). Elle se calcule de la manière suivante :

$$\text{Fréquence période P} = \frac{\sum_{\text{en nombre}} \text{sinistres période P}}{\sum \text{exposition période P}}$$

Cependant, ce type de modèle peut présenter certaines limites, comme le manque de précision si les données historiques ne sont pas représentatives des tendances futures, ou encore la variation des coûts, qui peut beaucoup varier en fonction de la nature des accidents et des caractéristiques des véhicules impliqués.

La tarification des flottes ouvertes chez AXA France Entreprise

Dans le cadre des flottes ouvertes, les souscripteurs établissent un tarif rapidement grâce aux outils de souscriptions. Trois méthodes de tarification sont proposées :

- Méthode 1 : la prime pure est calculée sur l'observée du client. On calcule la fréquence de sinistres et le coût moyen des sinistres sur l'observée du client. Cette méthode peut s'appliquer pour les garanties RC, BDG, DTA, INC et VOL. La responsabilité civile totale est calculée en faisant le produit de la fréquence du contrat (comprenant la fréquence de sinistres des garanties RCM et RCC) par le

coût moyen des sinistres (RCM et RCC), qui est calculé à partir de la charge écrêtée puis majorée d'un coefficient de mutualité afin de prendre en compte la charge des sinistres graves. Cette méthode est la plus adéquate pour les flottes ayant un nombre de véhicules important.

- Méthode 2 : la prime pure est calculée depuis juillet 2023 en faisant la moyenne pondérée des primes pures calculée avec la méthode 1 et méthode 3 en fonction de la taille de flotte. En effet, avant, la méthode 2 prenait en compte la fréquence du client mais le coût moyen se faisait sur le théorique AXA. Le coût moyen basé sur le théorique AXA est calculé par catégorie de véhicule et par garantie et se base sur l'observé du portefeuille Parc et FND d'AXA France. Cette méthode était préconisée pour les flottes ayant un nombre de véhicules faible. Cependant, le taux d'utilisation de cette méthode étant très faible, et, le taux d'utilisation de la méthode 1 trop élevé malgré un faible nombre de véhicules / sinistres, il a été décidé de changer cette méthode en une moyenne des méthodes 1 et 3 pondérée par le nombre de véhicules. Ce poids ω est borné entre 0.1 et 0.9 et dépend des garanties ainsi que du nombre de véhicules présent dans la flotte.
- Méthode 3 : la prime pure est calculée uniquement sur le théorique AXA, en fréquence et en coût moyen. Cependant, cette méthode n'est pas proposée pour les garanties en responsabilité civiles car pour la RC les données du client sont primordiales. Cette méthode s'applique pour les garanties BDG, DTA, INC, VOL et SDC, particulièrement pour les contrats sur lesquels les antécédents des sinistres ne sont pas connus.

Ainsi, pour la garantie RCC et RCM, la méthode 1 est la méthode la plus utilisée étant donné que la méthode 3 est interdite.

6.2 Les modèles linéaires

La régression est utilisée pour modéliser la relation entre une variable aléatoire Y, dite variable réponse, variable dépendante, variable expliquée ou variable ajustée, et les prédicteurs $X_1, ..., X_p$, aussi appelés variables explicatives, variables indépendantes, ou même régresseurs. Ces variables explicatives sont supposées non aléatoires et mesurables sans erreurs. Le modèle linéaire gaussien est un modèle statistique utilisé pour modéliser la relation entre une variable dépendante continue et un ensemble de variables indépendantes dans un cadre linéaire, en supposant que les résidus (erreurs de prédiction) suivent une distribution gaussienne. Le modèle peut être exprimé de la manière suivante :

$$y = \beta^T X + \epsilon$$

Avec y la variable réponse, dépendant continue, que l'on cherche à prédire ; X est un vecteur de variables indépendantes qui influencent y (variables explicatives) ; et, β est un vecteur de coefficients associés à chaque prédicteur, c'est l'impact de chaque variable explicative sur la variable réponse. ϵ est le terme d'erreur, qui identifie les écarts entre

les valeurs prédites et les valeurs observées de y, i.i.d. de moyenne nulle et de variance constante.

Ce modèle linéaire implique que les résidus sont gaussiens. Cependant, dans la pratique, de nombreux types de données contiennent des erreurs non-normales : les erreurs peuvent être asymétrique, comme une densité Gamma avec une queue de distribution significativement épaisse à droite, elles peuvent être plus aplaties que la distribution normale, avoir des valeurs négatives, ou encore être délimitées. Pour faire face à ces problèmes, par le passé, il fallait transformer la variable réponse ou adopter des méthodes non paramétriques.

Si la variable réponse n'est pas normale, mais d'une autre famille de la classe exponentielle, ces problèmes peuvent être résolus grâce aux modèles linéaires généralisés, Generalized Linear Model (GLM), introduits par John Nelder et Robbert Wedderburn en 1972. Ces modèles constituent une extension des modèles linéaires classiques et sont particulièrement utilisés afin de modéliser la fréquence et le coût moyen des sinistres. La famille exponentielle est importante en statistiques et en probabilité, notamment grâce à ces propriétés algébriques et ces distributions naturelles à considérer. Dans le cas des GLM, les résidus peuvent être d'une autre distribution que la loi normale, les erreurs peuvent être :

- Poisson, cette loi est utile pour les données de comptage (nombre de sinistres pour la fréquence par exemple)
- Binomiale, cette loi est utile sur les proportions
- Gamma, cette loi est utile pour les données avec un coefficient de variation constant
- Exponentielles, pour des données de mortalité par exemple

Les modèles linéaires généralisés possèdent trois caractéristiques :

1. **Une composante aléatoire**, qui précise que les observations sont des variables aléatoires indépendantes Y_i, $i = 1, ..., n$ avec une densité appartenant à al famille de dispersion exponentielle. Il s'agit de la loi de probabilité adossée à la variable cible Y, qui doit faire partie de la famille exponentielle. La densité d'une loi appartenant à la famille exponentielle peut s'écrire sous la forme suivante :

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

   Avec $a(.)$, $b(.)$ et $c(.)$ des fonctions spécifiques, θ le paramètre naturel de la famille exponentielle et ϕ le paramètre de dispersion (cf. Annexe D 7.2).
   Il est à noter qu'il est possible d'affecter un poids connu ω_i aux observations tel que $a(\phi) = \frac{\phi}{\omega_i}$. Pour simplification, on suppose que $\omega_i = 1$ Les exemples les plus importants sont les suivants :

Loi	θ	ϕ	$a(\phi)$	$b(\theta)$	$c(y_i, \phi)$
Normale(μ, σ^2)	μ	σ^2	ϕ	$\frac{\theta^2}{2}$	$-\frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)$
Gamma(μ, λ)	$-\frac{\mu}{\lambda}$	$\frac{1}{\mu}$	ϕ	$-ln(-\theta)$	$\frac{ln(\frac{y}{\phi})}{\phi} - ln(y) - ln\left(\Gamma\left(\frac{1}{\phi}\right)\right)$
Poisson(μ)	$ln(\mu)$	1	ϕ	$\exp(\theta)$	$-ln(y!)$

TABLE 6.1 – Famille exponentielle

2. **Une composante déterministe** qui attribue à chaque observation un prédicteur linéaire noté η_i telle que

$$\forall i \in \{1, ...n\}, \eta_i = \beta_0 + \beta_1 \cdot x_{i,1} + ... + \beta_p \cdot x_{i,p}$$

Il est également possible de noter sous forme matricielle :

$$\eta = X\beta$$

où X est la matrice du modèle avec les valeurs X_{ij} des variables explicatives et β est le vecteur des paramètres du modèle.

3. **Une fonction de lien g**, qui connecte les deux premiers éléments. L'espérance μ de Y est liée au prédicteur linéaire η_i par une fonction de lien. Elle permet de relier la variable cible à l'ensemble des variables explicatives telle que

$$g\big(\mathbb{E}(Y|X)\big) = \beta_0 + \beta_1 \cdot X_1 + ... + \beta_p \cdot X_p = g(\mu(x)$$

Définie sur $\mathbb{R}$, elle est appelée fonction de lien. Elle est monotone, différentiable et donc inversible. La fonction de lien canonique qui associe la moyenne μ_i au paramètre naturel de la famille exponentielle θ est défini par la relation suivante :

$$\mu_i = g^{-1} \Leftrightarrow \eta_i = \theta_i$$

Le choix de la fonction de lien est d'une importance cruciale car elle détermine la manière dont la combinaison linéaire des prédicteurs est transformée pour modéliser la relation entre les prédicteurs et la variable réponse. Elle a un impact direct sur la structure du modèle, les interprétations des coefficients ainsi que la forme des prédictions. Les fonctions de liens les plus fréquemment utilisées sont les suivantes :

Liens	Fonctions de lien
Log	$g(\mu_i) = \log \mu_i$
Logit	$g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right)$
Probit	$g(\mu_i) = \Phi^{-1}(\mu_i)$, où $\Phi(.)$ est la fonction de répartition de la loi normale centrée réduite
Inverse	$g(\mu_i) = \mu_i^{-1}$

TABLE 6.2 – Fonctions de lien

L'équation générale du modèle s'écrit ainsi de la manière suivante (avec g la fonction de lien) :

$$g\Big(\mathbb{E}(Y|X)\Big) = \beta_0 + \beta_1 \cdot X_1 + ... + \beta_p \cdot X_p + \epsilon$$

L'estimation des paramètres β se fait par l'utilisation du maximum de vraisemblance. On peut ainsi maximiser la log-vraisemblance :

$$\log\Big(L(y_1, ..., y_n; \theta, \phi)\Big) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)$$

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

Il faut résoudre les équations de vraisemblance suivante :

$$\frac{\partial \log(L)}{\partial \beta_j} = 0$$

Loi	Support	Fonction de lien	Expression du modèle
Normale	$\mathbb{R}$	$g(x) = x$	$y_i = \beta_0 + \beta_1 x_{i,1} + ... + \beta_p x_{i,p} + \epsilon$
Poisson	$\mathbb{N}$	$g(x) = ln(x)$	$y_i = \exp(\beta_0 + \beta_1 x_{i,1} + ... + \beta_p x_{i,p} + \epsilon)$
Gamma	$\mathbb{R}^+$	$g(x) = \frac{1}{x}$	$y_i = (\beta_0 + \beta_1 x_{i,1} + ... + \beta_p x_{i,p} + \epsilon)^{-1}$

TABLE 6.3 – Expression des modèles en fonction des lois

Pour la modélisation du nombre de sinistres, la loi Poisson, aussi appelée loi des évènements rares, peut être utilisée. Exemple du modèle log-Poisson :

$$Y \sim P\Big(\lambda\big(X_1 + ... + X_p\big)\Big)$$

$$ln\left(\frac{\text{Nombre de sinistres}}{Exposition}\right) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$

$$ln(\text{Nombre de sinistres}) = \sum_i \beta_i X_i + ln(\text{Exposition})$$

$$\text{Nombre de sinistres} = \text{Exposition}, \prod_i \exp(\beta_i X_i)$$

Pour la modélisation du montant des sinistres, la loi Gamma peut être utilisée. Exemple du modèle log - Gamma :

$$Y \sim \Gamma\Big(\alpha, \beta\big(X_1 + ... + X_p\big)\Big)$$

$$\log\left(\frac{\text{Coût des sinistres}}{\text{Nombre de sinistres}}\right) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$

$$\text{Coût moyen} = \prod_i \exp(\beta_i X_i)$$

Bien que la fonction de lien canonique de loi Gamma soit la fonction inverse, la fonction de logarithme est généralement privilégiée car elle permet d'obtenir un modèle multiplicatif.

6.3 Les modèles de Machine Learning

Les modèles de Machine Learning ont révolutionné la manière dont les entreprises abordent la résolution de problèmes complexes et l'automatisation de processus. Ils sont utilisés dans une multitude de domaines, tels que la finance, la santé, la gestion des ressources humaines, la recommandation de produits et bien d'autres.

Un modèle de Machine Learning est une représentation mathématique d'une relation entre les données d'entrée et les résultats souhaités. En d'autres termes, c'est un algorithme qui apprend à partir de données passées pour faire des prédictions ou prendre des décisions sur de nouvelles données. Plutôt que de programmer explicitement des règles pour résoudre un problème, on fournit au modèle des exemples et il apprend par lui-même à généraliser et à faire des inférences sur de nouvelles données.

Il en existe plusieurs qui sont utilisés dans divers. Parmi les plus connus sont les suivants : la régression linéaire, la régression logistique, les arbres de décisions, le Random Forest, les réseaux de neurones artificiels, les réseaux neuronaux convolutionnels (CNN), les K-plus proches voisins (KNN), le XGBoost, le Gradient Boosting Machine (GBM) ou encore les autoencodeurs. Tous ces modèles ont la capacité d'apprendre à partir des données, de prendre des décisions basées sur des modèles complexes et des relations complexes.

Dans le cadre de ce mémoire, les modèles de Machine Learning qui ont été utilisés sont le Random Forest, aussi appelé forêt aléatoire, et le Gradient Boosting Machine, aussi appelé machine à gradient boosté. Ces modèles ont particulièrement été utilisés dans le cadre de sélection de variable, lors de la phase préliminaire de modélisation sur la plateforme AKUR8 (qui sera présentée dans le Chapitre 7).

Le **Random Forest** est un algorithme d'apprentissage automatique qui combine plusieurs arbres de décision pour obtenir des prédictions plus robustes et précises. Une forêt aléatoire est un ensemble d'arbre de décision, où chaque arbre est construit à partir d'un échantillon aléatoire de données d'entraînement (train) et utilise une sélection aléatoire de variables lors de la construction des arbres. Pour évaluer l'importance des différentes variables, trois grandes étapes sont à effectuer :

1. Bootstrap Aggregating (Bagging) : Entraînement de la forêt aléatoire avec un ensemble de données historiques comprenant la variable réponse (coût moyen ou fréquence) et les variables explicatives. Le Random Forest construit plusieurs arbres de décision en utilisant différents échantillons aléatoires de données et de variables.
2. Calcul de l'importance des variables : analyse de comment chaque variable a contribué à la réduction de l'erreur de prédiction lors de la construction des arbres. Celles qui sont souvent utilisées dans la construction des arbres et qui ont réduit de manière significative l'erreur de prédiction sont considérées comme des variables importantes.
3. Sélection des variables : les variables ayant une importance élevée sont les plus pertinentes dans la prédiction et donc auront un impact significatif sur la modélisation des tarifs.

C'est un modèle robuste, performant sur des données complexes et est capable de gérer un grand nombre de caractéristiques. Cependant, il est plus complexe à interpréter et nécessite des ressources computationnelles plus importantes.

Le **Gradient Boosting Machine** est une autre technique d'apprentissage automatique qui construit plusieurs modèles de manière séquentielle, en corrigeant les erreurs du modèle précédent. Le processus de construction d'un GBM est le suivant :

1. Construction du premier modèle, généralement un arbre de décision, sur les données d'entraînement
2. Calcul des résidus du modèle précédent pour chaque point de données
3. Construction d'un nouveau modèle pour prédire les résidus du modèle précédent afin de réduire les résidus
4. Mise à jour du modèle en ajoutant les prédictions du nouvel arbre de décision
5. Répétition des étapes 2 à 4, ce qui permet d'améliorer graduellement la performance du modèle

Cette méthode peut être utilisée pour évaluer l'importance relative des différentes variables dans la prédiction des tarifs. Après avoir entraîné un modèle GBM sur un ensemble de données historiques, il est possible d'analyser les scores d'importance attribués

à chaque variable par le modèle. Les variables ayant des scores d'importance élevés sont considérées comme plus influentes pour la prédiction des tarifs. L'identification des variables les plus importantes peut guider dans la sélection des variables tarifaires.

6.4 Les critères de sélection des modèles et méthode de sélection des variables tarifaires

Une étape importante dans la construction d'un modèle consiste à sélectionner les variables qui seront incluses dans le modèle final. En effet, toutes les variables ne contribuent pas nécessairement de manière significative au modèle de prédiction. Il faut ainsi éliminer les variables non pertinentes et les variables redondantes afin d'améliorer les performances du modèle.

Une méthode qui permet d'éviter la sélection de variables fortement corrélées entre elles, et donc éviter un risque de surapprentissage, est l'**étude des corrélations**. Cette étape est essentielle dans la construction du modèle, et nécessite d'étudier d'une part les variables qualitatives et d'autre part les variables quantitatives.

Variables qualitatives : L'étude des corrélations des variables qualitatives, aussi appelées variables catégorielles, a été réalisée grâce à la méthode du V de Cramer. Cette méthode est une technique statistique utilisée pour mesurer la force et l'association entre deux variables qualitatives dans un tableau de contingence. Cette méthode est basée sur le test du chi-carré, noté χ^2, et quantifie la relation entre les variables en générant un coefficient appelé *V de Cramer*.
Pour rappel, le test du χ^2 est calculée en comparant les fréquences observées dans le tableau de contingence avec les fréquences attendues. Son calcul se fait de la manière suivante :

$$\chi^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k}$$

avec O_k les fréquences observées (Observed frequency) et E_k les fréquences attendues (Expected frequency).

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

$$\begin{cases} \mathrm{H}_0 : & \text{Il n'y a pas de relation entre les variables, les variables sont indépendantes.} \\ \mathrm{H}_1 : & \text{Il existe une relation entre les variables, les variables sont dépendantes.} \end{cases}$$

Aussi connu sous le nom de coefficient de contingence de Cramer, le V de cramer se calcule de la manière suivante :

$$V = \sqrt{\frac{\chi^2}{n \times \min(k-1, r-1)}}$$

Avec χ^2 la statistique du test du chi carré calculée pour les deux variables, n le nombre d'observations et k et r le nombre de niveaux (modalités) pour les variables X et Y. Plus le coefficient est proche de 1, plus la corrélation entre les deux variables est importante, et plus il est proche de 0, plus la corrélation est faible.

En appliquant cette méthode sur les variables qualitatives des données, les relations entre les variables sont établis dans le corrélogramme ci-dessus :

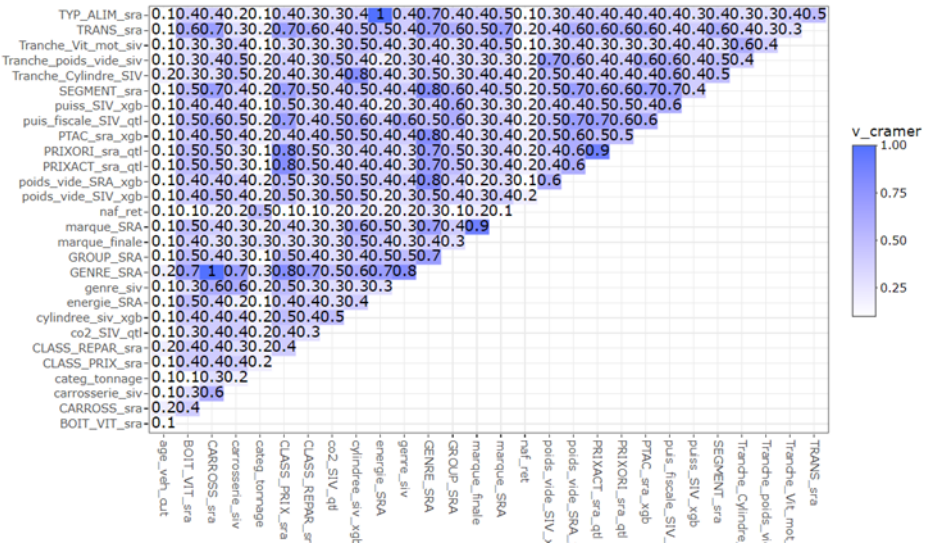


FIGURE 6.1 – Corrélogramme des variables catégorielles

Il est observé qu'il existe une forte corrélation entre la marque dérivée de la base véhicule et la marque SRA, étant donné qu'elles transmettent toutes deux la même information. Afin de déterminer quelle variable parmi ces deux marques revêt une plus grande importance, une analyse de leur corrélation avec la variable réponse (coût moyen ou fréquence) se révèle pertinente. Par ailleurs, il convient de noter que la marque issue de la base véhicule présente un niveau de détail excessif en raison de sa multitude de modalités. Cette abondance de modalités pourrait engendrer une complexité excessive et entraver l'interprétation du modèle. En effet, il serait ardu d'interpréter l'impact spécifique de chaque modalité sur la variable cible. Une corrélation significative entre le genre SRA (véhicule utilitaire ou véhicule particulier) et la carrosserie SRA est également à relever. L'examen approfondi indique que les véhicules utilitaires se composent principalement de fourgons ou de fourgonnettes, ce qui explique la liaison entre ces deux variables. Le type d'énergie et le mode d'alimentation du véhicule sont aussi fortement corrélés, du fait qu'ils transmettent des informations particulièrement similaires. Enfin, il est remarquable que le genre SRA affiche une corrélation marquée avec plusieurs variables, suggérant ainsi l'importance de cette variable qui englobe diverses informations au sein d'une même entité.

Variables quantitatives : L'étude des corrélations des variables quantitatives a été réalisée grâce à la corrélation de Spearman, aussi appelée corrélation de rang de Spearman. C'est une mesure statistique non paramétrique utilisée pour évaluer la relation entre deux variables quantitatives. Le ρ de Spearman mesure cette relation et permet l'étude de la dépendance entre deux variables et évalue la corrélation entre les rangs des observations plutôt que leurs valeurs brutes. Le coefficient de Spearman peut varier de -1 à 1, où -1 indique une corrélation négative parfaite, c'est-à-dire si une variable augmente, l'autre diminue de manière monotone), 0 indique l'absence de corrélation monotone entre les variables et 1 une corrélation positive parfaite. Il est défini de la manière suivante :

$$\rho = \frac{\sum_{i=1}^{n}(r_{Xi} - \overline{r_X})(r_{Yi} - \overline{r_Y})}{\sqrt{\sum_{i=1}^{n}(r_{Xi} - \overline{r_X})^2} \times \sqrt{\sum_{i=1}^{n}(r_{Yi} - \overline{r_Y})^2}}$$

avec r_{Xi} et r_{Yi} les rangs respectifs de la i^e observation de X et Y, et $\overline{r_X}$, $\overline{r_Y}$ les moyennes des rangs.

La représentation sous forme de corrélogramme appliqué sur les variables quantitatives des données est représentée ci-dessous :

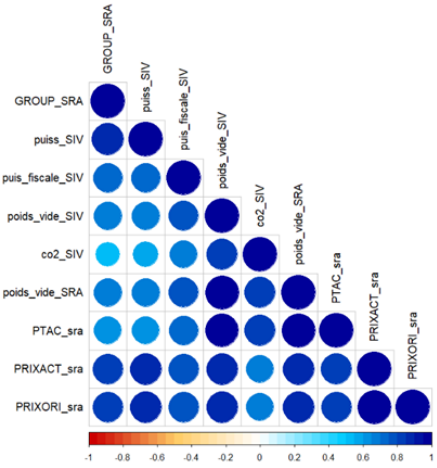


FIGURE 6.2 – Corrélogramme des variables quantitatives

Dans l'analyse des données, il est observé que le prix initial (prix d'origine) du véhicule (désigné par PRIXORI_sra) présente une forte corrélation avec le prix actuel du véhicule (désigné par PRIXACT_sra), ainsi qu'avec diverses autres variables. Ces variables comprennent le groupe SRA auquel appartient le véhicule (représenté par GROUP_SRA), la puissance du véhicule, sa puissance fiscale, son poids à vide, son taux d'émission de CO_2 et son PTAC. Cette corrélation est cohérente et suggère qu'un véhicule doté d'une puissance fiscale élevée, d'un poids à vide important ou d'un PTAC élevé affiche généralement un prix élevé, que ce soit son prix actuel ou son prix initial.

Il est également à noter que le groupe SRA présente une corrélation significative avec diverses caractéristiques telles que la puissance du véhicule (représentée par puis_SIV), les prix du véhicule (PRIXACT_sra, PRIXORI_sra), le poids à vide (poids_vide_SIV, poids_vide_SRA) ainsi que la puissance fiscale (puis_fiscale_SIV). Le groupe SRA reflète la probabilité d'occurrence d'un accident en se basant sur les caractéristiques techniques du véhicule. En effet, il évalue intrinsèquement le niveau de dangerosité du véhicule. Le calcul de ce groupe prend en considération divers éléments tels que la puissance du véhicule, son poids à vide, sa vitesse maximale et son poids total en charge (PTAC). Cette explication justifie ainsi la corrélation entre le groupe SRA et les variables mentionnées précédemment. Par conséquent, cette variable revêt une importance capitale, car elle synthétise plusieurs informations liées au véhicule. Par conséquent, son inclusion dans le processus de modélisation pourrait s'avérer particulièrement pertinente.

Il existe d'autres méthodes dans l'étape de sélection des variables. Il a été présenté les méthodes de Machine Learning utilisées lors de ce mémoire qui sont les méthodes Random Forest et Gradient Boosting Machine. Ces méthodes sont des procédures avancées basées sur des algorithmes, mais il existe également des **méthodes dites pas à pas** qui sont couramment utilisées lors de modèles linéaires généralisés.

La sélection backward : Cette sélection, aussi appelée élimination arrière, est une technique utilisée dans le domaine d'apprentissage automatique (Machine Learning) pour la sélection des variables les plus significatives. Elle consiste à utiliser toutes les variables disponibles, puis supprime de manière itérative les variables qui ont le moins d'importance, qui sont le moins significatif, sur la performance du modèle.
L'idée est de construire un modèle complet puis de le simplifier en éliminant les variables qui ne contribuent pas de manière significative à la précision du modèle. Les variables les moins importantes sont éliminées une par une en utilisant une technique de validation croisée pour évaluer l'impact de la suppression de chaque variable sur les performances du modèle. Cette étape itérative se termine à partir d'un critère d'arrêt prédéfini (par exemple, lorsque la performe du modèle ne s'améliore plus significativement, lorsque le nombre désiré de variables est atteint ou encore lorsque que la p-value soit inférieur à un certain seuil défini).

La sélection forward : Cette sélection, aussi appelée sélection progressive, est également une technique utilisée dans le domaine d'apprentissage automatique pour la sélection des variables. Contrairement à la méthode de sélection backward, celle-ci commence avec un ensemble de vide de variables et les ajoute une par une en évaluant leur impact sur la performance du modèle. Tout comme la méthode backward, elle utilise des techniques de validation croisée pour évaluer la contribution des variables à la performance globale. Elle se poursuit également jusqu'à ce qu'un critère d'arrêt prédéfini soit atteint.

Ces deux méthodes utilisent des tests statistiques comme la déviance, les critères de sélection AIC ou BIC ou encore le R-carré, R^2, ajusté.

Application sur les bases de données

Les quatre méthodes de sélection de variables ont alors été réalisées sur les cinq bases de modélisation, à savoir la base de modélisation du coût moyen des sinistres de la garantie RCM, celle du coût moyen des sinistres de la garantie RCC, celle de la fréquence des sinistres de la garantie RCC, celle de la fréquence des sinistres forfaitaires de la garantie RCM et enfin la base de modélisation de la fréquence des sinistres non forfaitaires de la garantie RCM. Les résultats de la sélection des variables les plus significatives pour la base de modélisation du coût moyen des sinistres de la garantie RCM sont présentés dans le tableau ci-dessous :

	Sélection forward	Sélection backward
Méthodes pas à pas	Tranche_poids_vide_siv CARROSS_sra Tranche_cylindree_SIV genre_siv puis_fiscale_SIV_qtl	puis_fiscale_SIV CARROSS_sra Tranche_poids_vide_siv GROUP_SRA CLASS_REPAR_sra
	Gradient Boosting Machine	Random Forest
Méthodes supervisées	Tranche_poids_vide_siv marque_siv nb_places_SIV CARROSS_sra categ_tonnage	co2_SIV_qtl poids_vide_SRA_qtl PRIXACT_sra_qtl puis_max_siv_qtl CLASS_REPAR_sra

TABLE 6.4 – Sélection des variables pour la base de modélisation du coût moyen des sinitres de la garantie RCM

Concernant les **critères de sélection du modèle**, il en existe plusieurs qui permettent de juger la qualité de l'ajustement du modèle et donc de sélectionner le modèle le plus adapté.

Les critères AIC et BIC : Ces critères sont des méthodes fréquemment utilisées en statistiques pour la sélection du meilleur modèle parmi un ensemble de modèles candidats. Le critère AIC, Akaike's Information Criterion, a été développé par Hirotugu Akaike. Il mesure la qualité d'ajustement d'un modèle en pénalisant la complexité. Deux éléments sont combinés par l'AIC : la capacité du modèle à expliquer les données et la quantité de paramètres du modèle. Il se calcule de la manière suivante :

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

Avec k le nombre de paramètres à estimer du modèle et L le maximum de la fonction de vraisemblance du modèle.
Le modèle avec le plus petit AIC est généralement préféré, car il indique un bon équilibre.

Le critère BIC, Bayesian Information Criterion, a été proposé par Gideon Schwartz en 1978. Il a le même principe que le critère AIC mais inflige une pénalité plus sévère pour la complexité. Le BIC prend en compte la taille de l'échantillon dans la pénalité, ce qui s'avère utile lorsque la taille de l'échantillon est petite. Sa formule est la suivante :

$$BIC = -2\log(L) + k\log(N)$$

Avec k le nombre de paramètres libres du modèle, L la vraisemblance du modèle estimé et N le nombre d'observations dans l'échantillon.
Tout comme pour l'AIC, le modèle avec le plus petit BIC est privilégié.

La déviance : C'est un concept statistique utilisé en analyse de régression et en modélisation statistique, notamment dans le contexte des modèles linéaires généralisés (GLM). La déviance mesure la différence entre la log-vraisemblance du modèle ajusté, autrement dit la probabilité de voir les données observées compte tenu du modèle, et la log-vraisemblance du modèle saturé, qui est un modèle qui ajuste parfaitement les données. Sa formule est la suivante :

$$D = -2\phi \times (\log(L) - log(Ls))$$

Avec L le maximum de la fonction de vraisemblance du modèle, Ls le maximum de la fonction de vraisemblance du modèle saturé et ϕ le paramètre de dispersion.

La déviance standardisée D^* s'exprime de la manière suivante :

$$D^* = -2(log(L) - Log(Ls)) \text{ on a ainsi } D = \phi D^*$$

Ainsi, plus l'écart entre les log-vraisemblances sera faible, autrement dit, plus D sera petit, plus l'ajustement sera bon. De plus, d'après la propriété asymptotique du rapport de vraisemblance :

$$D^* \sim \chi^2_{n-p}$$

avec p le nombre de paramètres du modèle construit et n le nombre de paramètres du modèle saturé.

Le coefficient de détermination, noté R^2, est une mesure statistique utilisée pour évaluer la proportion de la variance de la variable réponse d'un modèle statistique qui est expliquée par les variables explicatives (prédictives) incluses dans le modèle. Le R^2 quantifie et mesure les fluctuations, les variations, présentes dans les données. Il varie entre 0 et 1, où 0 indique que le modèle n'explique aucune variation dans les données et 1 que le modèle explique la totalité de la variation dans les données. Son calcul est le suivant :

$$R^2 = 1 - \frac{\text{Somme des carrés des résidus}}{\text{Somme total des carrés}}$$

Un R^2 élevé ne garantit cependant pas que le modèle est bon. Il ne mesure que la proportion de la variance expliquée par le modèle.

Ces indicateurs sont importants dans le choix des modèles mais ne suffisent pas pour valider un modèle. Il faut analyser les valeurs prédites et les valeurs observées. L'**analyse des résidus** permet ainsi d'évaluer la qualité et l'adéquation d'un modèle statistique aux données observées. Les **résidus** r_i sont définis pour chaque observation i comme l'écart entre l'observation et sa valeur prédit :

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

Dans les critères de sélection du modèle, **l'erreur quadratique moyenne (MSE)** est également utilisée dans l'analyse des écarts du modèle. C'est une mesure d'erreur qui correspond à la somme des carrés des écarts entre les prédictions et les valeurs observées, rapporté au nombre d'observations :

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

On peut de plus parler de **la RMSE (Root Mean Square Error)**, aussi appelée erreur quadratique moyenne. C'est une mesure utilisée pour évaluer la précision d'un modèle de régression ou de prédiction. Elle calcule la racine carrée de la moyenne des carrés des écarts entre les valeurs prédites par le modèle et les valeurs observées. La formule de la RMSE est la suivante :

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

$$RMSE = \sqrt{MSE}$$

où n est le nombre d'observations, y_i est la valeur réelle observée pour l'observation i et $\hat{y}_i$ est la valeur prédite par le modèle pour l'observation i.
La RMSE quantifie ainsi l'écart moyen entre les prédictions et les valeurs réelles. Plus elle est faible, meilleure est la prédiction du modèle. C'est une mesure absolue qui évalue la dispersion moyenne des résidus autour de zéro.

Chapitre 7

Application et résultats des modèles

7.1 Présentation du logiciel AKUR8

Chez AXA France IARD Entreprise, la plateforme AKUR8 est utilisée pour la tarification de nos produits. AKUR8. C'est une plateforme d'automatisation de la tarification dans le domaine de l'assurance. Elle offre la possibilité de prendre des décisions tarifaires plus précises, plus rapides et transparentes en s'appuyant sur les données et l'intelligence artificielle (IA). Elle utilise des algorithmes d'apprentissage automatique avancés pour analyser les données historiques et des données en temps réel des compagnies d'assurances pour générer des tarifs optimaux.

AKUR8 propose des techniques de modélisation statistique et d'apprentissage automatique de pointe pour créer des modèles de tarification très précis. Cela permet ainsi aux compagnies d'évaluer au mieux les risques et de pouvoir ajuster les tarifs proposés en conséquence. De plus, AKUR8 est transparent quant au processus de tarification, les assureurs ont alors la capacité de comprendre comment est calculé chaque tarif, ce qui facilite ainsi la communication entre les parties prenantes. Il est possible de faire des ajustements tarifaires en temps réel ce qui permet aux assureurs de rester compétitifs et réactifs face aux changements potentiels du marché.

Les modèles d'IA de cette plateforme permettent ainsi de fixer des tarifs plus précis en tenant compte d'une multitude de variables. Différents modules sont proposés notamment :

- La gestion des données : Téléchargement de jeux de données, modification et création de variables (Module « Data »)
- La modélisation des risques : Construction de modèles de risques, exploitation de la puissance du Machine Learning pour calculer vos primes pures (Module « Risk »)
- Création et analyse des scénarios de taux : établir la meilleure stratégie et prendre des décisions de tarification (Module « Rate »)

- Construction de modèles comportementaux de conversion et de rétention, mesure de l'élasticité des prix (Module « Demand »)
- Configuration des espaces de travail et des utilisateurs

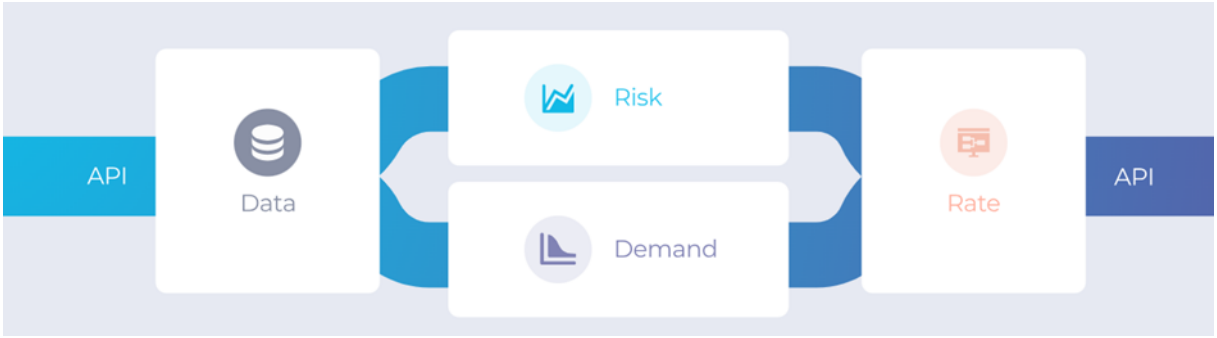


FIGURE 7.1 – Modules AKUR8

7.2 Tarification : résultats et comparaison des modèles obtenus

Après avoir effectué toutes les analyses préliminaires à la modélisation, il est maintenant possible de calculer la prime pure de la garantie RCC, la prime pure forfaitaire de la garantie RCM et la prime pure hors forfaits de la garantie RCM. La prime pure de la responsabilité civile est la somme de la prime pure des garanties RCM et RCC.

$$\text{Prime pure}_{\text{Responsabilité Civile}} = \begin{cases} \text{Coût moyen}_{\text{RCM}} \times \text{Fréquence forfaitaire}_{\text{RCM}} & + \\ \text{Coût moyen}_{\text{RCM}} \times \text{Fréquence non forfaitaire}_{\text{RCM}} & + \\ \text{Coût moyen}_{\text{RCC}} \times \text{Fréquence}_{\text{RCC}} \end{cases}$$

Pour ce mémoire, il a été décidé de réaliser trois types de modélisation et de choisir par la suite quel modèle appliquer pour chacune des garanties. Ces trois types de modélisation sont les suivants :

- Un modèle utilisant les variables provenant de la base SRA
- Un modèle utilisant les variables provenant de la base SIV et de la base SRA
- Un modèle en reprenant le tarif 2022 effectué par l'équipe APAT, c'est-à-dire avec les mêmes coefficients pour chaque variable tarifaire, et il y sera ajouté les variables provenant de la base SRA

A noter : pour chacune des modélisations, les variables contrats / variables véhicules sont également présentes, ainsi que la variable *Tranche_ dist_ moy* sur la distance moyenne des sinistres par rapport à l'adresse des souscripteurs.

Les premiers modèles linéaires généralisés qui ont été réalisés sont les modèles utilisant les variables provenant de la base SIV et de la base SRA ainsi que les modèles n'utilisant que les variables SRA. Cependant, après étude, ces modèles n'amélioraient pas les performances par rapport au tarif existant. C'est dans ce cadre qu'il a été décidé d'ajouter un type de modélisation en partant du tarif existant et d'ajouter les variables SRA afin de tenter d'améliorer le modèle et ses performances.

Les indicateurs de comparaisons des modèles entre eux sont nombreux. Parmi eux, on peut parler de l'indice de Gini ou encore de la Lift curve qui sont des indices présents dans le logiciel AKUR8.

L'indice de Gini est une mesure statistique utilisée pour évaluer la dispersion d'une distribution dans une population. Souvent utilisé pour quantifier l'inégalité économique d'un pays ou d'une région, il a été développé par le statisticien Corrado Gini et est exprimé sous forme d'un coefficient compris entre 0 et 1.

L'indice de Gini est calculé en traçant la courbe de Lorenz, qui représente graphiquement la répartition cumulative de la population en fonction de la répartition cumulative du revenu. La courbe de Lorenz est obtenue en ordonnant les observations par ordre croissant en fonction des prédictions du modèle et en calculant les valeurs cumulées observées jusqu'à l'observation i. Elle illustre comment la proportion de la valeur cumulée observée (sur l'axe des ordonnées) varie en fonction de la proportion du portefeuille requise pour atteindre cette valeur (sur l'axe des abscisses). La diagonale de la courbe de Lorenz représente l'égalité parfaite.

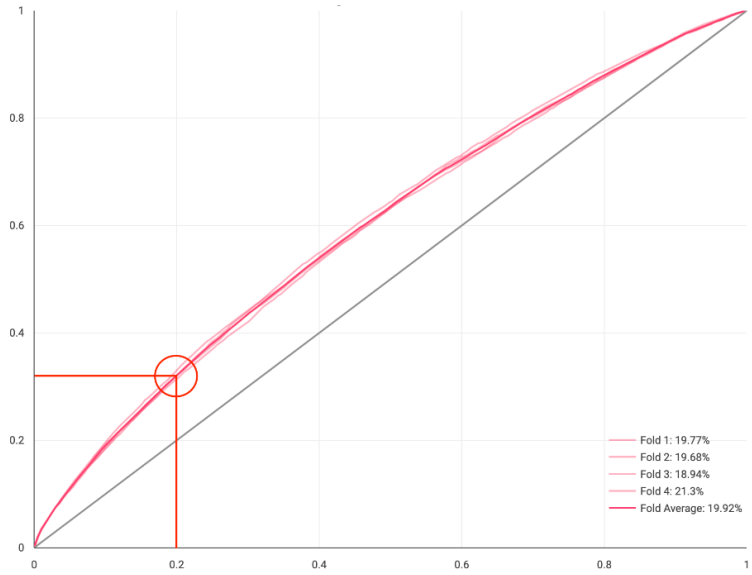


FIGURE 7.2 – Représentation graphique de la courbe de Lorenz (source AKUR8)

Dans l'exemple ci-dessus, les conclusions suivantes peuvent être émises :

— Les 20% cumulés des primes les plus élevées représentent 32% des pertes

— ou autrement dit, les 20% cumulés des risques prédits les plus élevés représentent en réalité 32% du risque observé

— Dans cet exemple, c'est une segmentation favorable du portefeuille par le modèle.

Le calcul de l'indice de Gini est basé sur l'aide entre la courbe de Lorenz et la diagonale :

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

Avec A qui représente l'aire entre la courbe de Lorenz et la diagonale et B l'aire au-dessus de la courbe de Lorenz.
Lorsque le modèle tend vers l'observé, l'indice de Gini tend vers 1. Cela indique que les prédictions du modèles sont proches de la réalité. Les modèles ayant un indice de Gini proche de 1 sont ainsi privilégiés car ils sont considérés comme plus performant en termes de prédictions.

La lift curve, courbe de gain en français, est construite en triant les prédictions de plus bas au plus élevé et en les regroupant en 20 catégories qui représentent 5% des prédictions chacune.

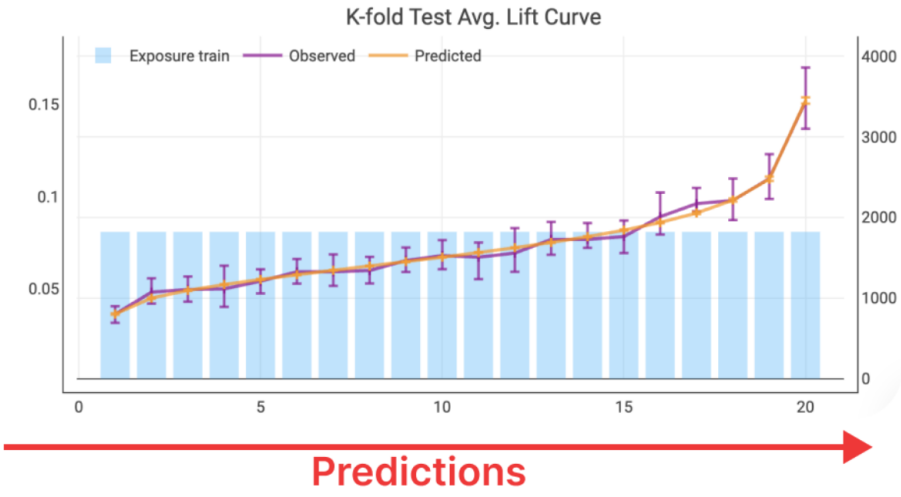


FIGURE 7.3 – Représentation de la Lift curve (source AKUR8)

La lift curve est utilisée afin d'évaluer la performance d'un modèle par rapport à une référence de base. Elle évalue l'efficacité d'un modèle. Pour chaque groupe, chaque quantile, le taux d'observation réel par rapport au taux d'observation attendu est calculé. L'interprétation se fait en observant la distance entre la courbe des prédits et la courbe des observés. Les courbes sont construites à partir des données de test de la Validation croisée, de manière à refléter les performances hors échantillon du modèle. Dans le cas

de l'exemple ci-dessus, les moyennes observées par quantiles sont représentées en violet (Observed) et les modélisés sont en jaune (Predicted). Le modèle s'ajuste correctement aux données dans cet exemple.

Le spread est également important a regarder pour chaque variable. Le spread d'une variable est calculé à partir de ses coefficients :

— Le Spread 100/0% prend l'ensemble des données et correspond à la différence entre les coefficient maximum et minimum :

$$\text{Spread} = \frac{\max(\text{Coefficients})}{\min(\text{Coefficients})} - 1$$

Si une variable n'est pas sélectionnée dans le modèle, tous ses coefficients sont égaux et le spread est égal à 0

— Le Spread 95/0% correspond à la même mesure après suppression des coefficients correspondant aux 5% les plus hauts et les plus bas du jeu de données. Il représente une vision plus robuste de l'impact de la variable.

Ces différents spread sont des éléments permettant de voir quel modèle est le meilleur parmi les trois types décrits précédemment. Par exemple, pour la modélisation de la fréquence des sinistres hors forfaits de la garantie RCM, il a fallu sélectionner le nombre de variables pour la modélisation des différents types et ensuite choisir quelle modélisation convient le mieux :

Fréquence hors forfait RCM - Modélisation SIV et SRA

Sur AKUR8, il est possible de construire une grid search qui permet de créer plusieurs modèles avec plusieurs variables discriminantes. Une fois les différents modèles créés, le graphique de la grid search permet de visualiser le nombre de variables dans le modèle sur l'axe horizontal et les performances hors échantillon du modèle (estimées en fonction du k-bloc sur le jeux de données de modélisation) sur l'axe vertical. Ce graphique offre une visualisation claire du compromis entre la complexité du modèle et ses performances. En effet, un modèle avec plus de variables sera plus difficile à interpréter qu'un modèle avec moins de variables. A performances égales, il est plus judicieux de prendre le modèle comportant moins de variables.

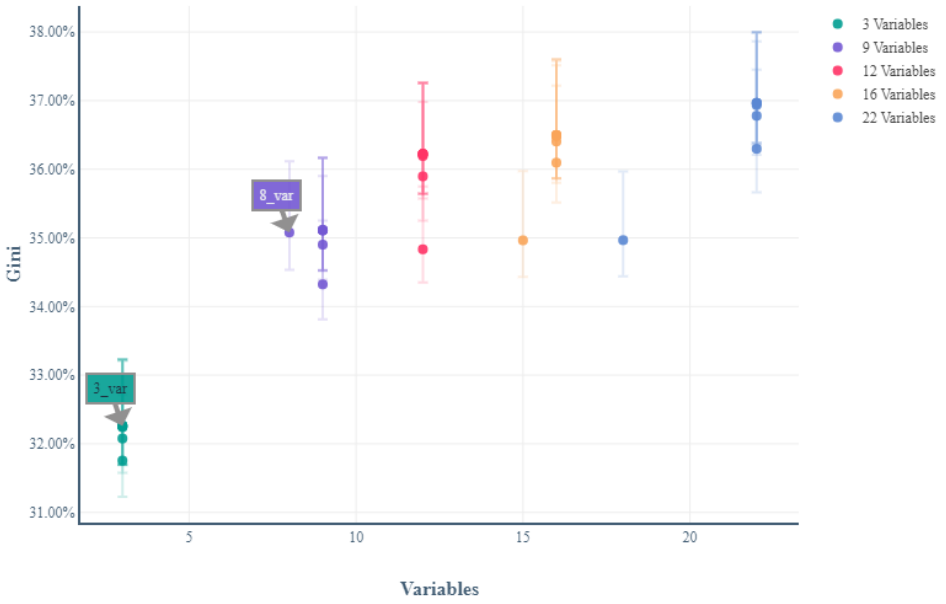


FIGURE 7.4 – Grid search des modèles de fréquence non forfaitaires utilisant les variables SRA et SIV

Sur le graphique ci-dessus, les modèles qui pourraient être intéressants sont le modèle à 3 variables et le modèle à 8 variables. Il est à remarquer que le modèle à 12 variables n'ajoute que 0,011 point au Gini comparé au modèle à 8 variables, ce qui n'est pas significatif au vu du nombre de variables à ajouter. Le modèle à 12 variables serait trop complexe dans l'interprétation et n'apporte pas beaucoup plus de performances. Après étude, les deux modèles comportait la variable année (*year*) qui ne peut pas être utilisée dans la modélisation. Il a donc été nécessaire de forcer les coefficients à 1 afin de pouvoir annuler l'impact de la variable et simplifier le modèle. Il y a ainsi deux modèles : un avec deux variables et l'autre avec 7 variables. Le modèle avec 7 variables a ainsi été choisi :

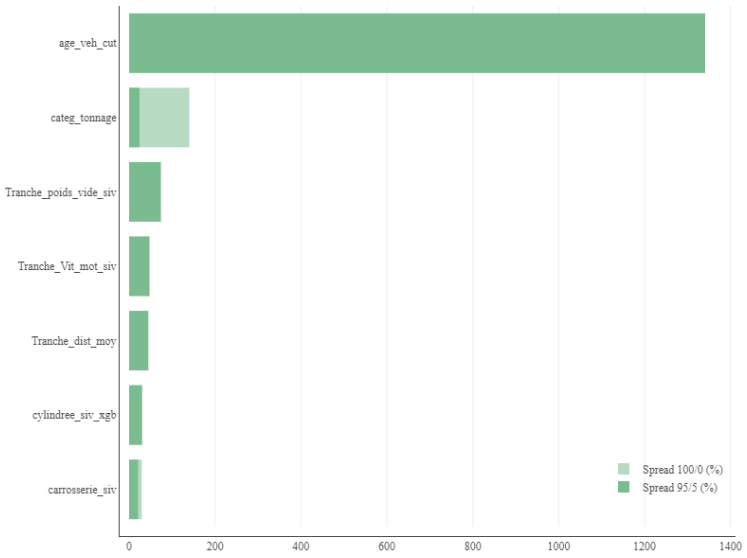


FIGURE 7.5 – Variables pour la prédiction de la fréquence des sinistres non forfaitaires - Modélisation SIV+SRA

Il est à noter que la variable qui possède la plus grande influence sur le modèle est la variable qui renseigne sur l'âge du véhicule. Les autres variables n'ont qu'une influence minime. Après avoir lissé certains coefficients pour certaines variables, on obtient les informations suivantes :

Indicateur	**Base train complète**	**Base train K-FOLD***	**Base test K-FOLD***
Gini	32,63%	32,89%	31,84%
Pseudo-R^2	5,49%	5,58%	5,24%
RMSE	0.2669	0.2683	0.2629
Deviance	190 600	143 000	47 660

TABLE 7.1 – Indicateurs du modèle SIV + SRA de fréquence des sinistres non forfaitaires RCM

Le **Pseudo -** R^2 utilisé est le Pseudo R^2 de McFadden. Il varie de 0 à 1 et est calculé comme suit :

$$\text{Pseudo-}R^2 = 1 - \frac{\text{ModelDeviance}}{\text{NullModelDeviance}}$$

Avec ModelDeviance qui représente la déviance du modèle ajusté, qui mesure l'ajustement du modèle aux données, et, NullModelDeviance qui représente la déviance du modèle nul, qui est un modèle de référence sans aucune variable indépendante. Plus il est proche de 1, meilleur est l'ajustement du modèle par rapport au modèle nul.

La validation croisée k-fold* est une méthode qui consiste à découper la base d'apprentissage en plusieurs échantillons, plus précisément en k échantillons (k plis) de façon aléatoire et pour chaque pli unique, le considérer comme la base test et le restant des plus comme l'ensemble de la base d'entraînement (d'apprentissage, base train). Cette technique permet d'évaluer la performance d'un modèle prédictif et d'estimer son aptitude à généraliser les nouvelles données qui n'ont pas été utilisées lors de son apprentissage. Elle aide à éviter le surapprentissage et offre une évaluation plus réaliste des performances d'un modèle.

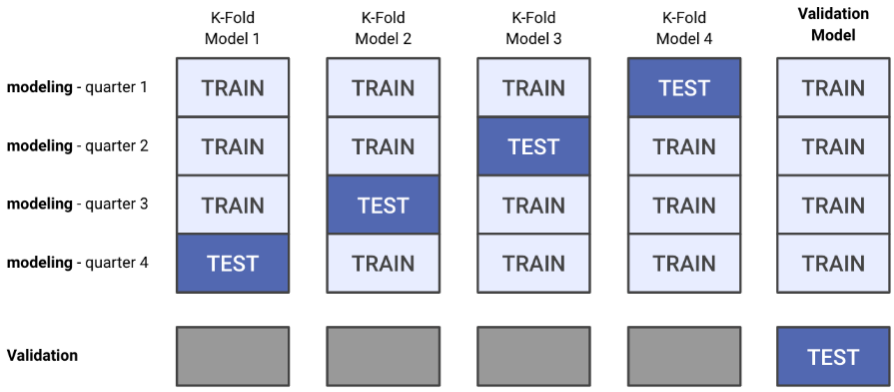


FIGURE 7.6 – Schéma de la validation croisée

Cette technique fournit une estimation plus robuste de la performance du modèle comparée à une seule division des données en base train et base test.

Fréquence hors forfait RCM - Modélisation SRA

Pour cette modélisation, les variables contrats, véhicules et les variables provenant de la base SRA ont été utilisées. Le même processus que pour la modélisation utilisant SIV et SRA a été réalisé. Pour la sélection de variables :

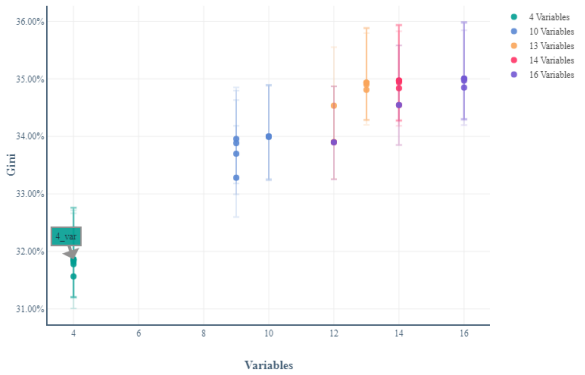


FIGURE 7.7 – Grid search des modèles de fréquence non forfaitaires utilisant les variables SRA uniquement

Pour cette modélisation, c'est le modèle à 4 variables qui a été retenu. Le modèle à 10 variables est trop complexe et n'apporte pas une performance significativement importante (+0,021). Après avoir également forcés les coefficients de la variable année à 1, le spread des 3 variables restantes est le suivant :

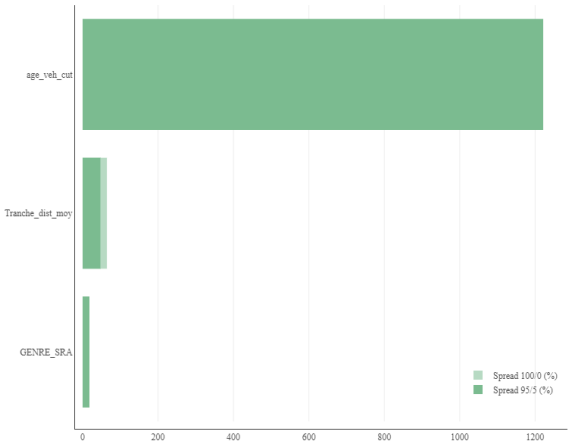


FIGURE 7.8 – Variables pour la prédiction de la fréquence des sinistres non forfaitaires - Modélisation SRA

Les indicateurs de cette modélisation sont les suivants :

Indicateur	**Base train complète**	**Base train K-FOLD**	**Base test K-FOLD**
Gini	29,82%	29,81%	29,81%
Pseudo-R^2	4,7%	4,72%	4,64%
RMSE	0.2672	0.2672	0.2672
Deviance	192 200	144 200	48 060

TABLE 7.2 – Indicateurs du modèle SRA de fréquence des sinistres non forfaitaires RCM

Pour cette modélisation, les indicateurs indiquent qu'il est moins performant que celui utilisant les variables SIV et SRA. La variable sur l'âge du véhicule est également une variable qui est a la plus grande importance.

Fréquence hors forfait RCM - Modélisation en repartant du tarif 2022 avec ajout des variables SRA

Enfin, la modélisation en repartant de la tarification effectuée en 2022 par l'équipe APAT avec ajout des variables SRA (ainsi que la variable *Tranche_ dist_ moy*) a été réalisée avec les mêmes étapes :

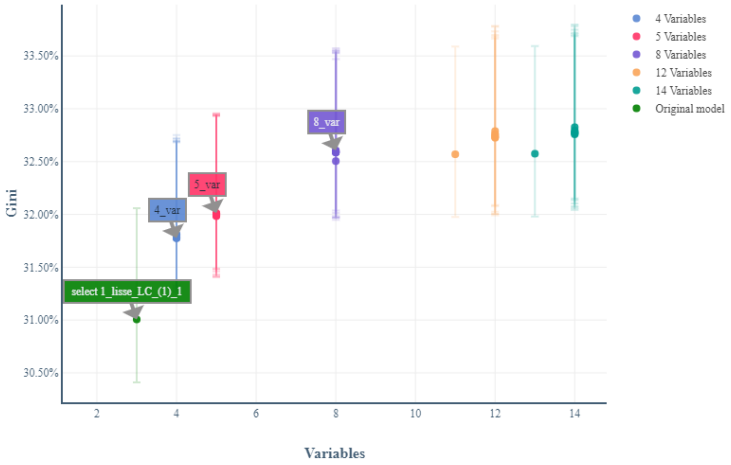


FIGURE 7.9 – Grid search des modèles de fréquence non forfaitaires utilisant la modélisation 2022 avec ajout des variables SRA

Ici, "select 1_lisse_LC_(1)_1" représente la modélisation effectuée en 2022 qui comporte 3 variables. Il a ainsi été intéressant de regarder les modèles à 4 variables (rajoutant ainsi une variable au modèle existant), 5 variables et 8 variables. Pour le modèle à 8 variables, il a été décidé que le nombre de variable était trop important et complexifierait ainsi trop le modèle. De plus, sur les variables ajoutées, certaines ne comportaient que des coefficients nuls exceptés pour une modalité ce qui rend la variable inintéressante dans la modélisation. Le choix entre le modèle à 4 ou 5 variables a été réalisé en regardant pour chaque nouvelle variable tarifaire quel modèle s'approche le plus de l'observé. Ainsi le choix définitif a été celui du modèle à 3 variables :

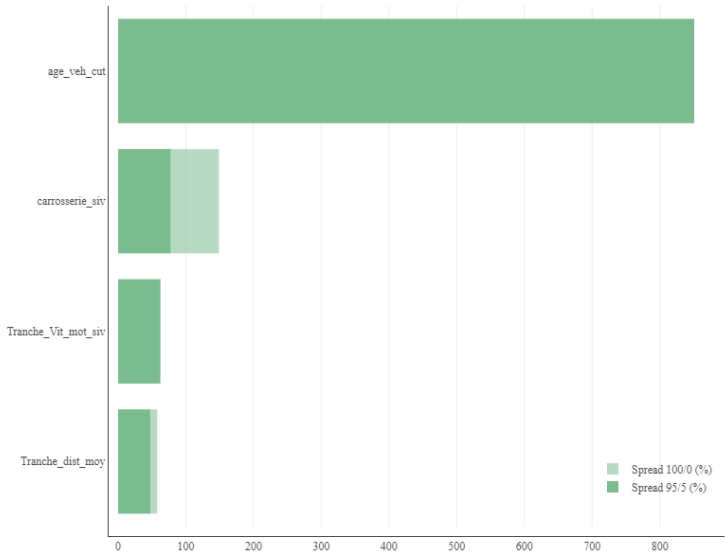


FIGURE 7.10 – Variables pour la prédiction de la fréquence des sinistres non forfaitaires - Modélisation tarif 2022 + SRA

Les indicateurs de cette modélisation sont les suivants :

Indicateur	**Base train complète**	**Base train K-FOLD**	**Base test K-FOLD**
Gini	31,4%	31,4%	31,4%
Pseudo-R^2	4,86%	4,86%	4,86%
RMSE	0.2672	0.2672	0.2672
Deviance	191 900	143 900	47 980

TABLE 7.3 – Indicateurs du modèle Tarif 2022 + SRA de fréquence des sinistres non forfaitaires RCM

Choix final du modèle pour la fréquence des sinistres non forfaitaires de la garantie RCM :

Le choix final du modèle pour la fréquence des sinistres non forfaitaires, garantie RCM, se fait ainsi entre la modélisation utilisant les variables SIV et SRA et la modélisation utilisant le tarif 2022 avec ajout des variables SRA. En effet, ce sont les deux modèles candidats qui offrent de bonnes performances. Le choix entre les deux se fait alors avec le nombre de variables le plus faible. En effet, le nombre de variables est beaucoup plus important pour la modélisation SIV/SRA que pour l'autre. Ce choix évite ainsi d'avoir un modèle trop complexe, un modèle surajusté qui pourrait entraîner une mauvaise capacité de généralisation à de nouvelles données, et une interprétation trop difficile à savoir quelles sont les variables qui ont réellement impacter les prédictions du modèle.

Résultats pour les autres garanties :

La même méthodologie a été réalisée pour chaque garantie par coût moyen / fréquence afin de déterminer quel type de modélisation choisir. Les résultats sont les suivant :

Garantie	RCM			RCC	
Modèle	Cout moyen	Fréquence sinistres forfaitaires	Fréquence hors sinistres forfaitaires	Cout moyen	Fréquence
Choix	SRA + SIV	Tarif 2022 + SRA	Tarif 2022 + SRA	Aucun modèle -> Cout moyen observé segmenté par la variable catégorie du véhicule x tonnage	Tarif 2022 + SRA
Variables retenues	poids_vide_siv_xgb carrosserie_siv GENRE_SRA	age_veh_cut categ_tonnage carrosserie_siv Tranche_poids_vide_siv CDGENRN Tranche_dist_moy	age_veh_cut carrosserie_siv Tranche_Vit_mot_siv Tranche_dist_moy		age_veh_cut carrosserie_siv categ_tonnage Tranche_dist_moy

FIGURE 7.11 – Résultats des modèles

A noter concernant le coût moyen de la garantie RCC, aucun modèle n'arrive à capter le coût moyen pour cette garantie. En effet, pour le tarif 2022, l'équipe APAT n'avait pas non plus réussi à trouver un modèle qui permet d'obtenir le coût moyen en RCC, probablement dû à la complexité des coûts lors des sinistres qui impliquent des dommages corporels plus ou moins important (blessures ou décès). En effet, le coût d'un sinistre varie considérablement d'un sinistre à un autre en fonction de plusieurs facteurs comme la gravité des blessures, le nombre de personnes impliquées, les frais médicaux, les pertes de revenus en cas de blessures entraînant une incapacité (temporaire ou permanente) de travailler, les frais juridiques ou encore les indemnités et compensations. Le choix qui a alors été fait pour le coût moyen RCC est de calculer pour chaque type de véhicule, le coût moyen sur les années 2019 à 2021. Pour chaque modalité de la variable catégorie $\times$ tonnage (categ_tonnage), on attribue le coût moyen observé ($= \frac{\text{charge}}{\text{nombre de sinistres}}$).
Il est également observable que finalement, les variables SRA ne ressortent pas des modèles exceptée la variable concernant le genre du véhicule.

Après la sélection des modèles, il devient possible de calculer plusieurs primes pures : la prime pure RCM forfaitaire, la prime pure RCM hors forfait, la prime pure RCM, la prime pure RCC et enfin la prime pure RC. Ensuite, la nouvelle méthodologie de calcul des primes pures peut être appliquée aux données de la base FND pour l'année 2022, permettant ainsi de comparer les primes pures issues de cette nouvelle méthode avec celles du tarif de l'année 2022 ainsi qu'avec les primes pures observées. Il est constaté que même si l'indice de Gini est plus bas par rapport au tarif de 2022, une analyse détaillée des différentes variables tarifaires révèle que la nouvelle méthode de calcul des primes pures se rapproche considérablement des primes pures observées. De plus, les tendances entre les primes pures observées et celles calculées par la nouvelle méthode sont très similaires.

L'application de cette nouvelle modélisation sur la base FND de 2022 peut ainsi être comparée au tarif 2022 ainsi qu'à la prime pure réellement observée. L'indice de Gini de la nouvelle modélisation atteint un niveau 31,94% contre 36,01% pour la tarification de 2022. Ceci indique une modélisation avec de moins bonne performance que le modèle existant. Cependant, en regardant les montants des primes pures moyenne, la prime pure de la nouvelle modélisation est plus proche de l'observé que de le tarif 2022 :

KPI (Key Performance Indicator)	Montant
Prime pure moyenne proposée	147, 30
Prime pure moyenne tarif 2022	169,18
Prime pure moyenne observée	146, 48

TABLE 7.4 – KPI des primes pures moyennes

Un autre indicateur important à regarder est la lift curve. Afin de comparer le tarif proposé, le tarif actuel 2022 existant (current) ainsi que l'observé, la double lift curve peut être utile :

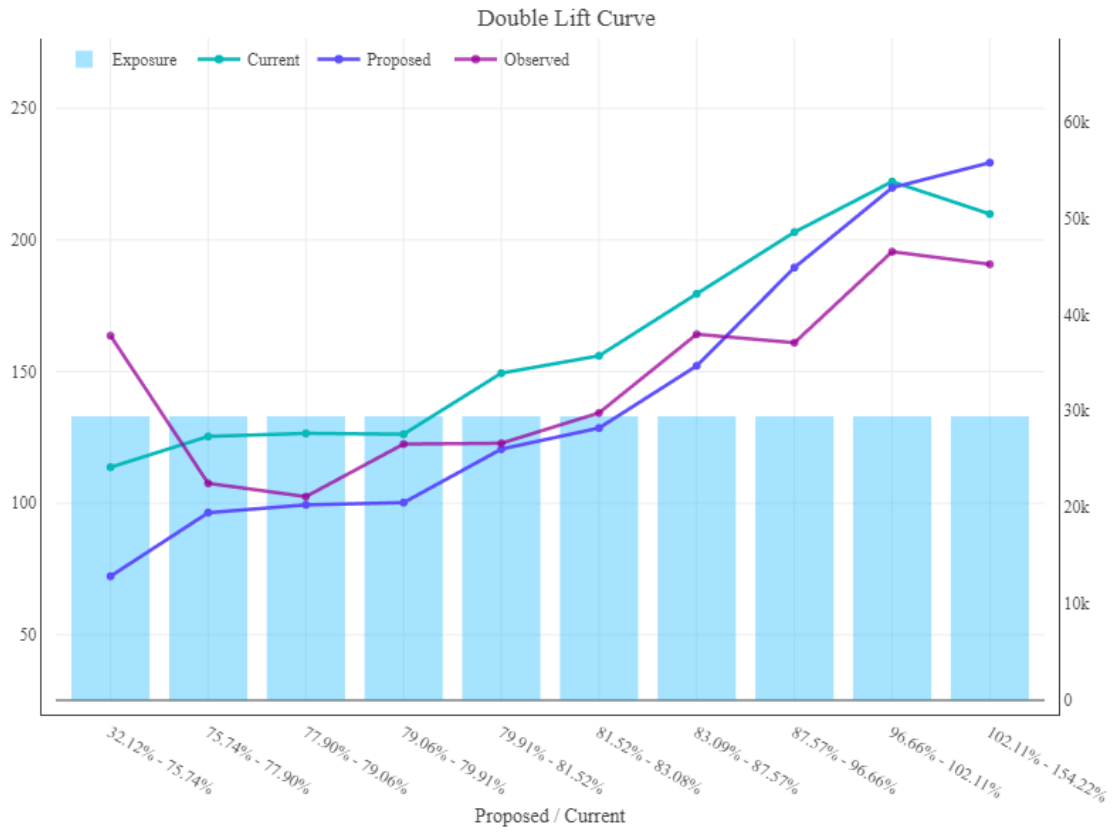


FIGURE 7.12 – Double lift curve - Comparaison nouvelle modélisation, tarif 2022 et observé

La courbe de la prime pure proposée semble plus proche de la prime pure réellement observée que la prime pure actuelle excepté pour les prédictions les plus basses (les plus à gauche sur le graphique). Ici, les prédictions sont regroupées en 10 catégories qui représentent 10% des prédictions chacune. Il serait également intéressant de regarder pour les variables tarifaires, si la prime pure de la nouvelle modélisation est également plus proche de la prime pure observée que de la prime pure du tarif 2022.

Pour des raisons de confidentialité, il a été décidé en accord avec les managers de ne pas donner le nom des variables tarifaires. Les variables seront ainsi nommées « Variable 1 », « Variable 2 » et « Variable 3 ». La **Variable 1** est ainsi une variable tarifaire regardée :

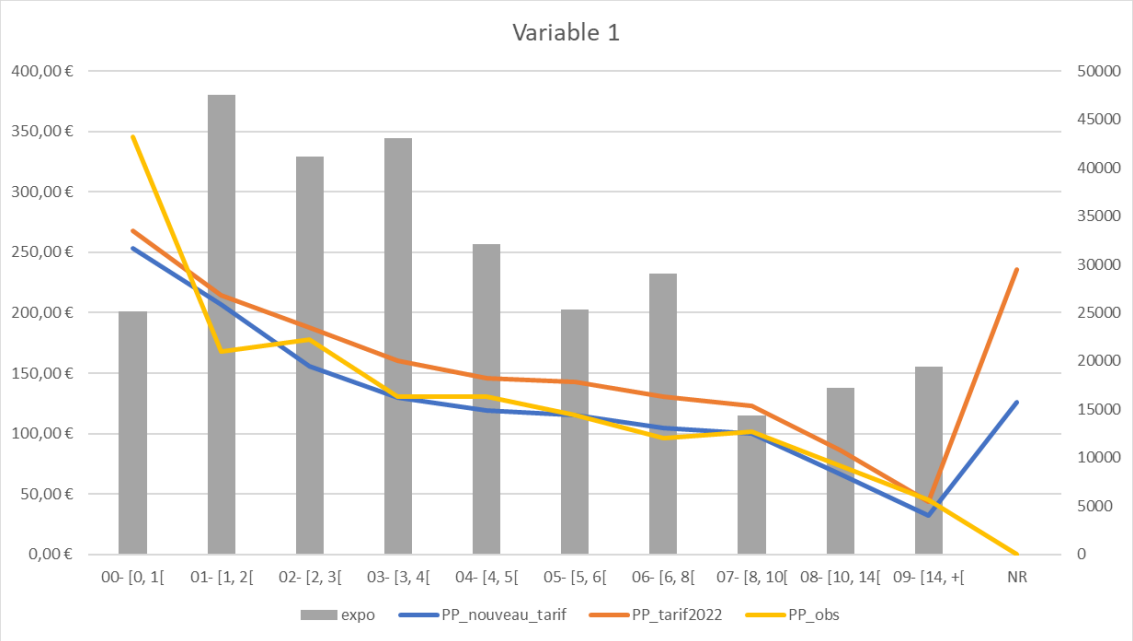


FIGURE 7.13 – Variable 1

Sur cette variable, il est facile de voir que la prime pure proposée (en bleu) est plus basse que la prime pure du tarif 2022 (en orange). La prime pure du nouveau tarif est globalement plus proche de la prime pure observée ce qui peut prouver que malgré un Gini moins bon que celui du tarif 2022, cette nouvelle modélisation suit correctement la prime pure observée.

Concernant **la variable 2**, un dépit d'un léger pic au niveau de la troisième tranche, cette variable est plus proche de la prime pure observée et comporte la même tendance que la prime pure :

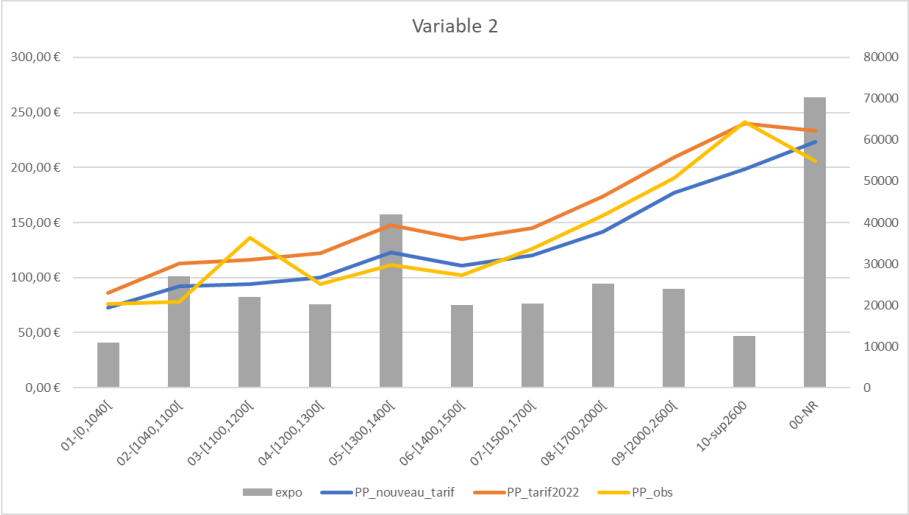


FIGURE 7.14 – Variable 2

Pour la **variables 3** :

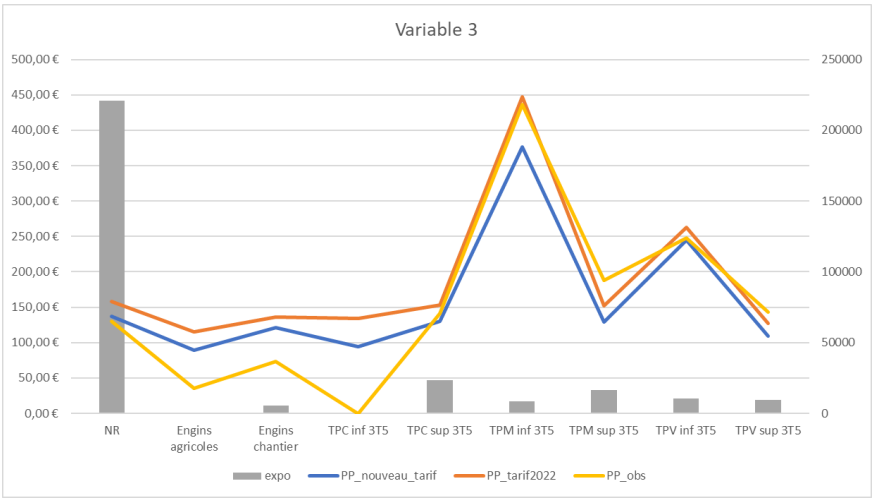


FIGURE 7.15 – Variable 3

On remarque que l'exposition des valeurs manquantes, "NR", est très forte et peu apporter un biais dans les résultats. En effet, cela peut apporter une perte d'informations qui aurait pu contribuer à une meilleure compréhension et une meilleure modélisation. Cependant, il est encore notable que la prime pure proposée suit la même tendance que la prime pure observée. Il est également possible d'observer pour certaines modalités que la prime pure proposée est inférieur à la prime pure observée ce qui peut indiquer une sous-estimation du risque.

Une dernière variable qui peut être à observer est la variable sur la **distance moyenne des sinistres depuis l'adresse des souscripteurs**, variable créée notamment pour tenter d'améliorer le modèle :

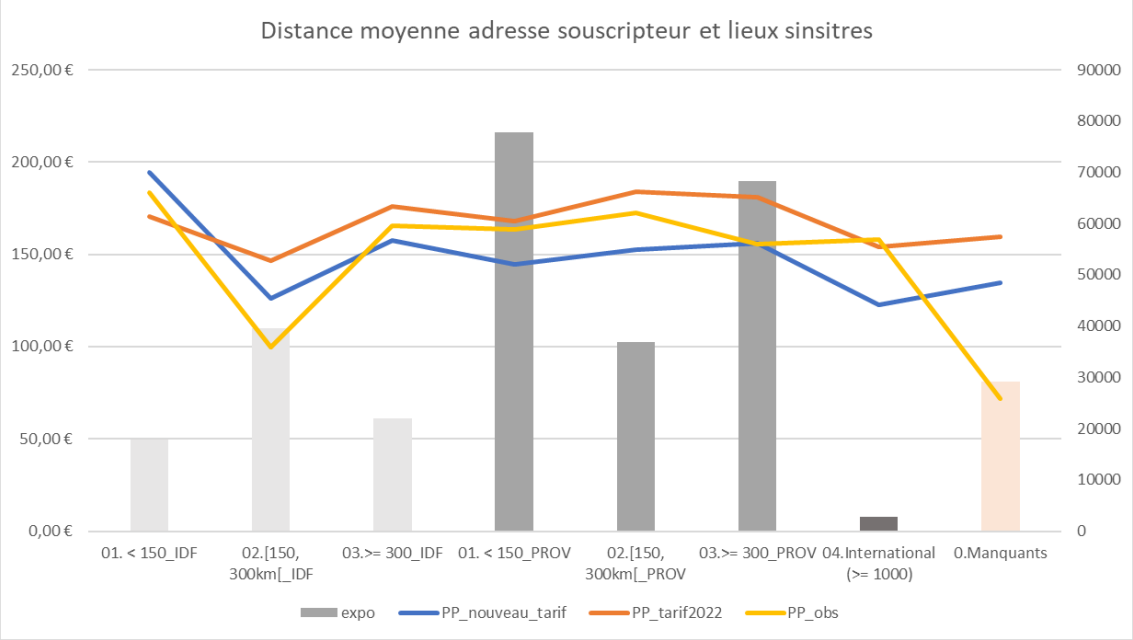


FIGURE 7.16 – Distance moyenne des sinistres depuis l'adresse du souscripteur, par tranche

Bien qu'elle n'était pas présente dans la modélisation du tarif 2022, la prime pure 2022 se rapproche tout de même de la prime pure proposée. La prime pure observée reste plus proche de la prime pure du nouveau tarif.

Il y a également eu une application de cette nouvelle modélisation sur la base PARC de l'année 2021 afin de comparer la prime pure appliquée sur cette base à la prime pure du tarif du Parc, du nom de Tarif E7, ainsi que la prime pure observée.

La même méthodologie a été réalisée pour la comparaison des primes pures. De manière générale, la prime pure issue de la modélisation FND est plus faible que la prime pure du tarif E7. L'écart entre les deux primes pures a été calculé de la manière suivante :

$$\text{ecart} = \frac{\text{PP_FND}}{\text{PP_PARC}} - 1$$

La densité des écarts est alors la suivante :

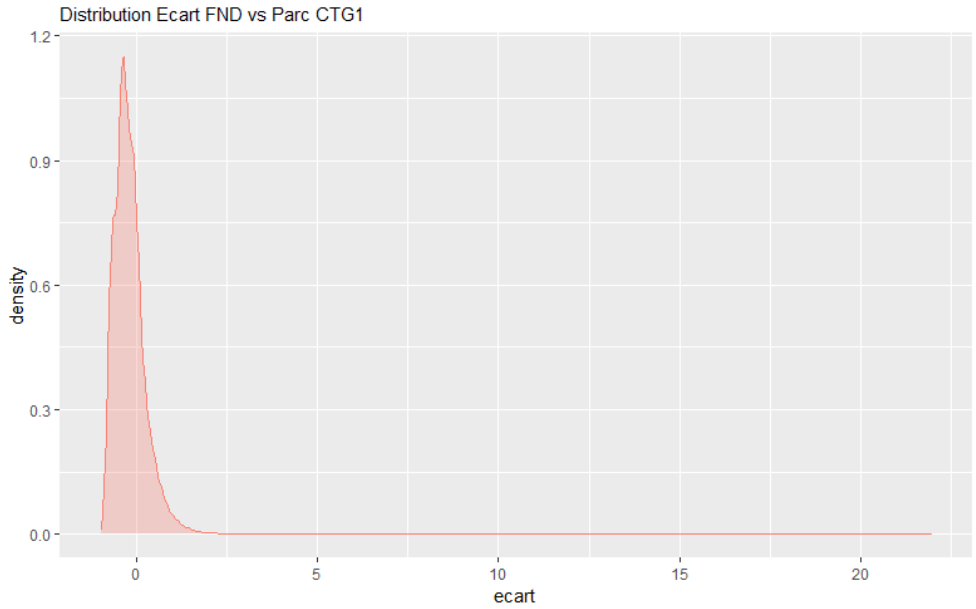


FIGURE 7.17 – Graphique de la densité des écarts

La densité est plus forte pour les écarts inférieur à 0, ce qui indique que les écarts sont plutôt négatifs :

$$\text{ecart} < 0 \Leftrightarrow \frac{\text{PP_FND}}{\text{PP_PARC}} < 1 \Leftrightarrow \text{PP_FND} < \text{PP_PARC}$$

En observant la distribution jusqu'au quantile à 96%, cela confirme que la prime pure du tarif E7 est plus élevée que celle proposée :

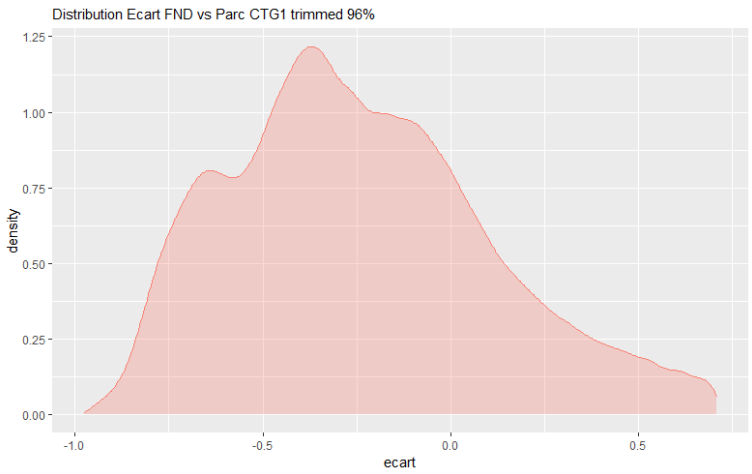


FIGURE 7.18 – Densité des écarts coupés au quantile 96%

Également, un nuage de points, aussi appelé scatter plot a été réalisé afin d'observer la relation entre les variables, ainsi que la dispersion des points par rapport à une ligne de tendance potentielle :

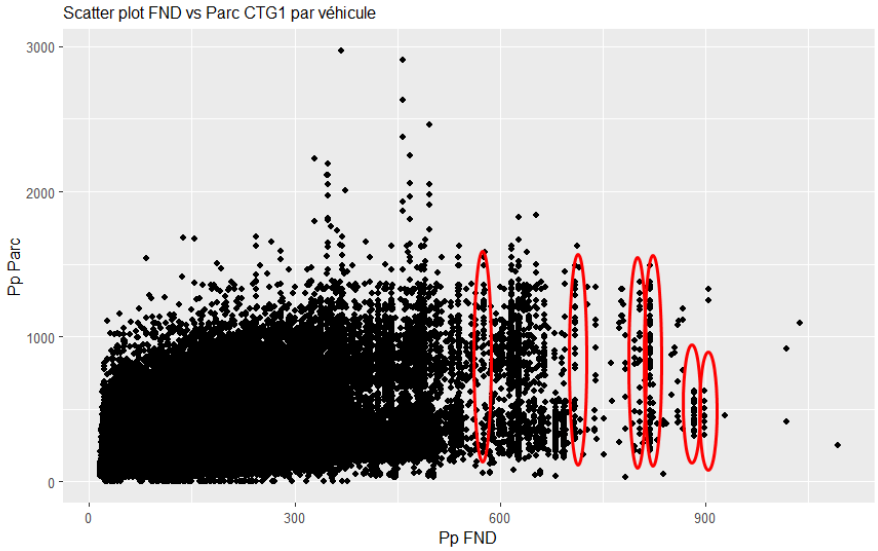


FIGURE 7.19 – Nuage de points de la prime pure Parc et de la prime pure FND

Les points verticaux (tarif constant FND contre tarif segmenté Parc) reflètent le caractère plus "fin" de la prime pure Parc par rapport à cella de la prime pure proposée FND. En effet, le modèle du Parc prend en compte plus de variables que les modèles FND car les flottes Parc comportent des informations qui ne sont pas disponibles pour les flottes non dénommées.

Il est possible également de regarder pour les variables tarifaires la prime pure moyenne du tarif E7, la prime pure moyenne FND proposée et la prime pure pure moyenne observée :

La variable 1 :

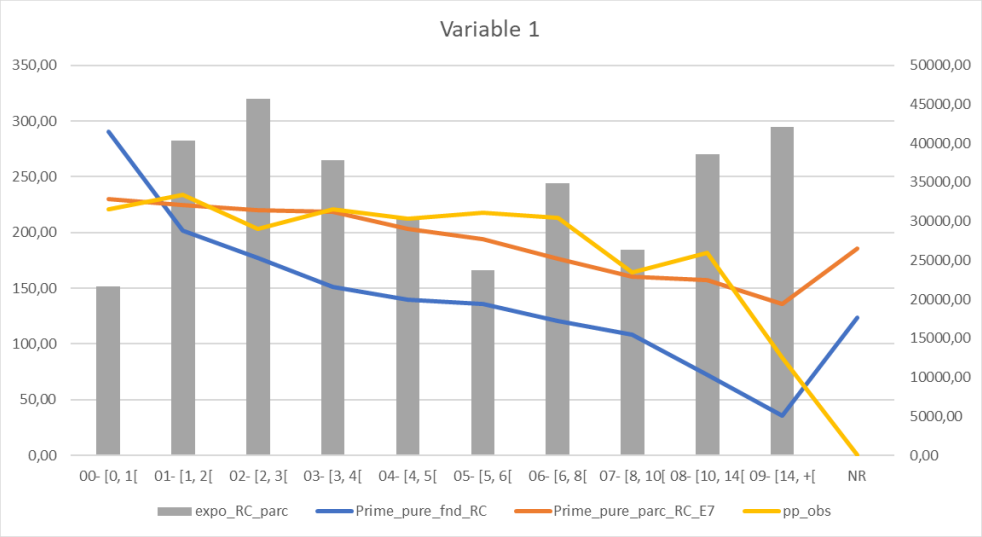


FIGURE 7.20 – Primes pures moyennes de la variable 1

Pour cette variable, la prime pure proposée sous-estime largement la prime pure observée ainsi que la prime pure E7. Cependant, il est possible d'observer la même tendance entre la prime pure FND et la prime pure Parc E7.

La variable 2 : La prime pure établie avec le nouveau tarif se rapproche plus de la prime pure observée que le tarif E7. Elle est légèrement en dessous de la prime pure observée mais reste plus proche que la prime pure du tarif E7.

Cette variable est intéressante car cela montre qu'avec cette tarification, il est possible de se rapprocher de la tarification du parc. C'est vers ce type de tarification que le tarif FND essaie de se rapprocher afin d'avoir de meilleures performances et un tarif plus fin.

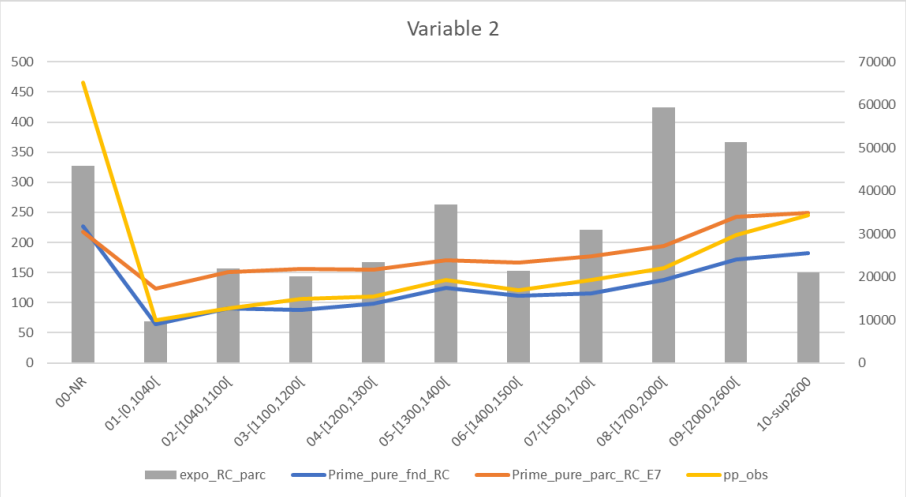


FIGURE 7.21 – Primes pures moyennes de la variable 2

La variable sur la distance moyenne des sinistres par rapport à l'adresse du souscripteur :

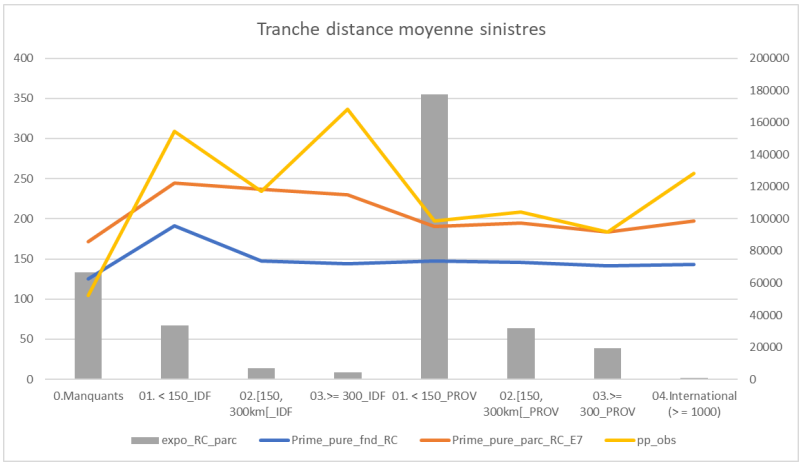


FIGURE 7.22 – Primes pures moyennes de la variable tranche de distance moyenne des sinistres

Cette variable ne semble pas significative avec la modélisation provenant du produit FND, car il n'y a pas de variations sur les différentes tranches pour la prime pure proposée. Cependant, la prime pure observée semble plus intéressante avec des variations sur les différentes tranches ce qui pourrait indiquer que cette variable pourrait être intéressante dans le tarif du Parc.

Enfin, la **variable 3** est également pertinente car il est possible de voir que la prime pure FND a la même tendance que la prime pure observée, excepté pour les engins agricoles où la prime pure observée est plus élevée que ce qui a été prédit avec les modèles FND et E7. La prime pure proposée est globalement plus faible que l'observée et que la prime pure E7. De plus, sur les mêmes modalités vues précédemment, il est possible d'observer une sous-estimation du risque avec la courbe de la prime pure proposée largement plus faible que la prime pure observée.

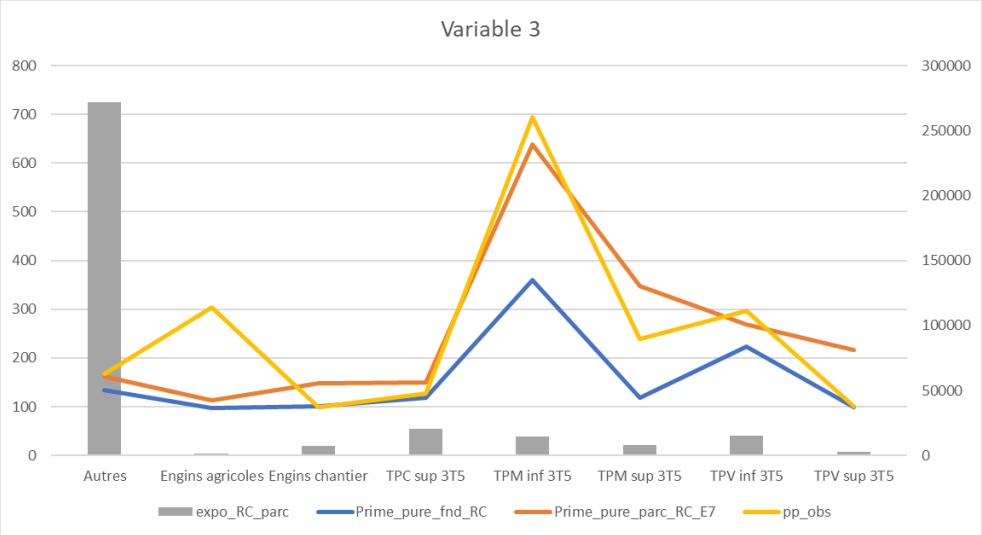


FIGURE 7.23 – Primes pures moyenne en fonction de la variable 3

En Annexe (Annexe E 7.2), d'autres graphiques et indicateurs sont présents.

Conclusion

L'objectif principal de ce mémoire était d'ajouter de nouvelles variables, notamment des variables SRA, afin de proposer une nouvelle modélisation de la prime pure de la responsabilité civile. Les données provenant de la base Sécurité et Réparation Automobiles n'étaient pas utilisées dans les dernières modélisations du tarif de la prime pure pour les flottes ouvertes. Le but a ainsi été d'améliorer les performances du modèle et de tester si l'ajout de ces variables est pertinent pour la modélisation de la prime pure RC ou bien si les variables sont plutôt redondantes avec les variables déjà présentes, notamment les variables provenant du Système d'Immatriculation des Véhicules. Les données provenant de la base SRA n'étant que des véhicules légers, inférieurs à 3,5 tonnes, ainsi que les deux roues et trois roues et quad motorisés, l'étude de ce mémoire ne s'est ainsi concentrée que sur la modélisation des véhicules de catégorie 1 (VL et VU de M3T5). Ce mémoire s'inscrit donc dans la stratégie d'AXA, "Driving Process 2023", dans le cadre des actions "Renforcer la performance de souscription" ainsi que "Accroître les cash-flow dans l'ensemble du Groupe", avec l'innovation d'ajouter des variables SRA, variables qui devraient être des variables robustes et importantes dans les modèles d'après l'organisme, dans le but d'améliorer les performances du modèle de prime pure RC et de proposer un tarif plus segmenté, plus précis et ainsi plus attractif pour les clients et plus compétitif.

L'obligation pour les compagnies d'assurances de déclarer les véhicules qu'ils assurent, avec la création du Fichier des Véhicules Assurés, a également été l'occasion pour les flottes non dénommées d'accroître la connaissance des véhicules grâce à l'acquisition de l'immatriculation des véhicules. La récupération de celle-ci se fait grâce au FVA qui est alimenté par les distributeurs : les agents et les courtiers. Les agents font partie d'AXA et peuvent rentrer les véhicules directement sur des outils AXA, où il est possible de récupérer les informations pour les transmettre à l'AGIRA. Ces informations peuvent être directement utilisées et accessibles. Cependant, pour les courtiers, certains sont des courtiers AXA, utilisant les outils propre à l'entreprise, et d'autres sont des courtiers dit déclarants, qui déclarent eux-mêmes les véhicules à l'AGIRA. Des études pour récupérer les informations des véhicules, et des études de contrôle de qualité des données sont encore en cours et en évolution constante. C'est dans ce cadre que la tarification par véhicule a été possible pour les flottes ouvertes.

En plus de l'ajout de nouvelles variables, une étude sur la géolocalisation des sinistres a pu être réalisée dans le cadre de ce mémoire afin de faire une première approche sur l'adresse des sinistres. Dépendant des résultats, la variable pourrait s'avérer intéressante dans le futur, avec potentiellement une évolution vers de nouvelles tranches plus fine sur l'activité des véhicules de l'entreprise comme une circulation locale, départementale, régionale, nationale, internationale, ou encore la mise en place d'un zonier.

Finalement, les trois types de modélisations qui ont été mis en place ne sont pas tous concluants. En effet, le modèle n'utilisant que les variables SRA n'a pas été le modèle le plus performant que ce soit pour le coût moyen ou la fréquence (forfaitaire ou non) des garanties RCM et RCC. Ce modèle ne présentait pas des performances complètement inefficaces mais restait largement plus faible que les autres types de modélisation. Pour le modèle utilisant les variables SIV et SRA, celui-ci était plus performant que le précédent mais finalement n'a été retenu qu'une fois pour la modélisation du coût moyen des sinistres de la garantie RCM, car en comparaison du modèle existant, les performances restaient de moindre qualité ce qui n'était pas intéressant dans le cadre d'une proposition d'une nouvelle modélisation pour améliorer le tarif existant. Finalement, au vu des résultats précédents des deux types de modélisation, un troisième type de modélisation a été proposé : partir du modèle existant, pour améliorer ses performances avec les variables SRA, ainsi que la variable de distance moyenne des sinistres. C'est finalement ce type de modélisation qui a été retenu pour la modélisation du coût moyen et de la fréquence des sinistres de la garantie RCC, ainsi que pour la modélisation de la fréquence forfaitaire et non forfaitaire de la garantie RCM. Une fois les modèles construits pour chaque garantie, il a été possible d'avoir la prime pure de la garantie Responsabilité Civile (par somme des primes pures des garanties RCM et RCC).

Afin de pouvoir tester la robustesse et les performances du modèle de prime pure RC, il a été réalisé une étape dite de backtesting. Cela consiste à évaluer la performance de la nouvelle modélisation en utilisant de nouvelles données, par exemple, des données FND 2022, ou encore des données Parc 2021, en comparant avec des résultats de prime pure déjà connus : la prime pure du tarif 2022 ou encore la prime pure du tarif Parc E7.

La comparaison entre la prime pure de la modélisation 2022 et la nouvelle modélisation proposée appliquée sur une nouvelle base FND de 2022 a pu montrer qu'en défaut d'un indice de Gini plus faible (-4 points) pour la nouvelle modélisation par rapport à l'existant, la prime pure proposée se rapproche globalement plus de la prime pure réellement observé. En effet, sur les variables explicatives ainsi que sur la double lift curve, il a été observé que les données observées se rapprochent plus des données prédites que des données actuelles. Finalement, il y a une légère amélioration du modèle et la nouvelle variable sur la distance moyenne des sinistres est une variable qui ressort des modèles ce qui montre son importance.

La comparaison avec le tarif E7 du Parc a pu démontrer que, globalement, la prime pure RC FND est plus basse que celle du tarif E7, mais la tendance reste proche de l'observé ainsi que de la prime pure E7. Le fait que la prime pure FND soit plus basse que

celle du Parc peut être expliqué du fait que le mix portefeuille de la composition du parc est différent de celui des flottes non dénommées, mais également du fait que sur le parc, un coefficient Technique Personnalité (CTP) a été appliqué. Ce coefficient correspond à un équivalent de bonus/malus existant sur le marche Particuliers & Professionnels afin de prendre en compte la fréquence sinistre d'un assuré.

Les pistes d'amélioration pour le futur sont les suivantes :

— Réduction du taux de valeurs manquantes, "NR" : construction d'un modèle qui prédit les manquants à partir des autres variables

— Amélioration de la variance distance moyenne des sinistres : Méthode de Text Mining pour récupérer des informations sur les rapports de sinistres ; Proposition d'un autre topage que l'Île-de-France ; Affiner la variable avec de nouvelles tranches

— Ajout d'autres variables externes, opendata : depuis des sites de l'État comme data.gouv par exemple

— Continuer les études de jointure entre les bases FVA, SRA, et même SIV, afin d'avoir une meilleure qualité de donnée

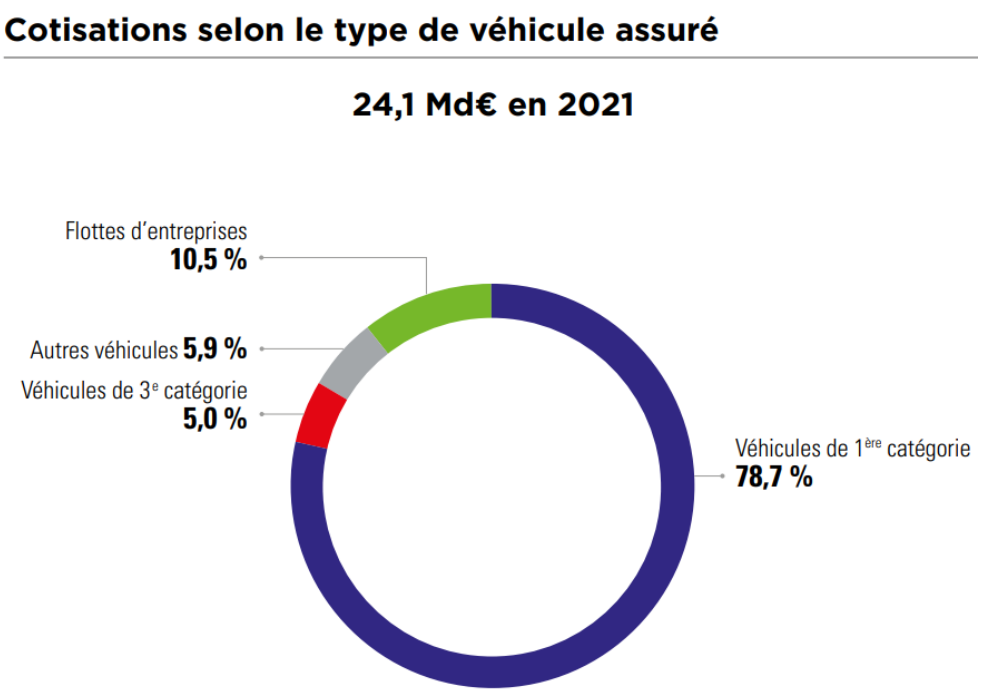
Annexes

Annexe A : Cotisations selon le type de véhicule assuré en 2021

Cotisations selon le type de véhicule assuré

M€	2017	2018	2019	2020	2021	21/20
Cotisations	**21 387**	**22 124**	**22 807**	**23 496**	**24 084**	**+ 2,5 %**
Véhicules de 1ère catégorie (1)	17 109	17 562	18 072	18 523	18 949	+ 2,3 %
Véhicules de 3e catégorie (2)	936	1 066	1 108	1 147	1 199	+ 4,5 %
Autres véhicules (3)	1 262	1 310	1 325	1 414	1 411	- 0,2 %
Flottes d'entreprises	2 080	2 186	2 302	2 412	2 525	+ 4,7 %

(1) Véhicules 4 roues de moins de 3,5 t hors flottes.
(2) Principalement des 2 roues motorisés hors flottes.
(3) Véhicules agricoles, véhicules de deuxième catégorie, missions, remorques, caravanes, engins spéciaux, polices garages, etc…



Source : Assurance française : Données clés 2021 - France Assureurs

Annexe B : L'estimateur de Pickands

$$\hat{\xi}^P_{(k(n),n)} = \frac{1}{\log(2)} \log\left(\frac{X_{(n-k(n+1,n)} - X_{(n-2k(n)+1,n)}}{X_{(n-2k(n)+1,n)} - X_{(n-4k(n)+1,n)}} \right)$$

Avec $k(n)$ est un entier tel que $1 \leq k(n) \leq n/4$.
On peut établir les résultats de convergence suivants lorsque F appartient au domaine d'attraction de H_ξ, $\xi \in \mathbb{R}$.

- Si $\lim_n k(n) = +\infty$ et $\lim_n k(n)/n = 0$ alors $\hat{\xi}^P_{(k(n),n)}$ converge en probabilité vers ξ lorsque n tend vers l'infini/
- Si de plus $\lim_{n\to\infty} k(n)/\log(\log(n)) = \infty$, alors la convergence a lieu presque sûrement.
- Sous des hypothèses supplémentaires sur $k(n)$ et F, on a la convergence en loi

$$\sqrt{k(n)} \left(\hat{\xi}^P_{(k(n),n)} - \xi \right) \Longrightarrow \mathcal{N}\left(0, \frac{\xi^2 \left(2^{2\xi+1} + 1\right)}{\left[2\left(2^\xi - 1\right)\log(2)\right]^2} \right)$$

Idée de la construction de l'estimateur : Soit $U(t) = F^{-1}(1-1/t)$, le quantile d'ordre $1-1/t$ de F, avec $t>0$. Nous avons établi pour $x, y > 0$ et $y \neq 1$

$$\lim_{s\to\infty} \frac{U(sx)-U(s)}{U(sy)-U(s)} = \begin{cases} \frac{x^\xi-1}{y^\xi-1}, & \text{si } \xi \neq 0, \\ \frac{\log(x)}{\log(y)}, & \text{si } \xi = 0. \end{cases}$$

En choisissant $x = 2, y = 1/2$ et $s = t/2$, on obtient

$$\lim_{t\to\infty} \frac{U(t)-U(t/2)}{U(t/2)-U(t/4)} = 2^\xi, \quad \forall \xi \in \mathbb{R},$$

et plus généralement si c_1 et c_2 sont deux fonctions telles que $c_1(t) \to 1/2$ et $c_2(t) \to 1/4$ lorsque $t \to \infty$ on a

$$\lim_{t\to\infty} \frac{U(t)-U\left(tc_1(t)\right)}{U\left(tc_1(t)\right)-U\left(tc_2(t)\right)} = 2^\xi, \quad \forall \xi \in \mathbb{R}.$$

L'idée de la construction de l'estimateur de Pickands est alors de trouver c_1 et c_2 en fonction des observations (X_i).

Soit $(V_i)_{1\leq i\leq n}$ un échantillon de loi de Pareto de paramètre 1 notée F_V, $F_V(x) = 1-1/x$, pour $x \geq 1$, et $k(n)$ entier tel que $1 \leq k(n) \leq n/4$, $\lim_n k(n) = +\infty$ et $\lim_n k(n)/n = 0$. On admet les convergences suivantes :

$$\lim_{n\to\infty} V_{(n-k(n)+1,n)} = \infty, \quad \lim_{n\to\infty} \frac{V_{(n-2k(n)+1,n)}}{V_{(n-k(n)+1,n)}} = \frac{1}{2}, \quad \lim_{n\to\infty} \frac{V_{(n-4k(n)+1,n)}}{V_{(n-k(n)+1,n)}} = \frac{1}{4}$$

On en déduit

$$\lim_{n\to\infty} \frac{U\left(V_{(n-k(n)+1,n)}\right)-U\left(V_{(n-2k(n)+1,n)}\right)}{U\left(V_{(n-2k(n)+1,n)}\right)-U\left(V_{(n-4k(n)+1,n)}\right)} = 2^\xi.$$

Mais par définition de U et F_V, on a $U(V) = F^{-1}\left(F_V(V)\right)$ et donc

$$\left(U\left(V_{(1,n)}\right), \ldots, U\left(V_{(n,n)}\right)\right) = \left(F^{-1}\left(F_V\left(V_{(1,n)}\right)\right), \ldots, F^{-1}\left(F_V\left(V_{(n,n)}\right)\right)\right)$$

Mais on sait que si X a pour loi $F, F(X)$ suit une loi uniforme sur $[0,1], \mathcal{U}$, et que $F^{-1}(\mathcal{U})$ a même loi que X. Ceci permet de conclure que le vecteur $\left(U\left(V_{(1,n)}\right), \ldots, U\left(V_{(n,n)}\right)\right)$ a même loi que $\left(X_{(1,n)}, \ldots, X_{(n,n)}\right)$, on en déduit l'égalité en loi

$$\frac{U\left(V_{(n-k(n)+1,n)}\right)-U\left(V_{(n-2k(n)+1,n)}\right)}{U\left(V_{(n-2k(n)+1,n)}\right)-U\left(V_{(n-4k(n)+1,n)}\right)} = 2^{\hat{\xi}^P_{(k(n),n)}}$$

(égalité en loi). On peut donc conclure que $\hat{\xi}^P_{(k(n),n)}$ converge en loi vers ξ.

Annexe C : L'estimateur de Hill

$$\hat{\xi}^H_{(k(n),n)} = \frac{1}{k(n)} \sum_{i=n-k(n)+1}^{n} (\log(X_{(i,n)}) - \log(X_{(n-k(n)+1,n)}, 1 \leq \wr(n) \leq n$$

On peut établir les résultats de convergence suivants lorsque F appartient au domaine d'attraction de $H_\xi, \xi > 0$. - Si $\lim_n k(n) = +\infty$ et $\lim_n k(n)/n = 0$ alors $\hat{\xi}^H_{(k(n),n)}$ converge en probabilité vers ξ lorsque n tend vers l'infini. - Si de plus $\lim_n k(n)/\log(\log(n)) = \infty$, alors la convergence a lieu presque sûrement. - Sous des hypothèses supplémentaires sur $k(n)$ et F, on a la convergence en loi

$$\sqrt{k(n)} \left(\hat{\xi}^H_{(k(n),n)} - \xi \right) \Longrightarrow \mathcal{N}\left(0, \xi^2\right)$$

Idée de la construction de l'estimateur : Comme $\xi > 0, F$ appartient au domaine d'attraction d'une loi de Fréchet (avec $\alpha = 1/\xi$). On peut montrer dans ce cas que (en notant $\bar{F} = 1 - F$)

$$\lim_{t\to\infty} \frac{\bar{F}(tx)}{\bar{F}(t)} = \frac{1}{x^\alpha}.$$

On peut donc écrire $\bar{F}(x) = \frac{1}{x^\alpha} L(x)$ où L est à variation lente et le Théorème de Karamata donne

$$\int_x^\infty \frac{L(t)}{t^{\eta+1}} dt \sim \infty \frac{1}{\eta x^\eta} L(x), \quad \forall \eta > 0.$$

Pour X variable aléatoire de loi F, calculons $\mathbb{E}(\log(X) - \log(t) \mid X > t)$, en supposant pour simplifier que F admet pour densité f. On a

$$\begin{aligned}
\mathbb{E}(\log(X) - \log(t) \mid X > t) &= \frac{1}{\bar{F}(t)} \mathbb{E}\left((\log(X) - \log(t)) 1_{\{X>t\}} \right) \\
&= \frac{1}{\bar{F}(t)} \int_t^\infty \frac{\bar{F}(u)}{u} du
\end{aligned}$$

On en déduit

$$\mathbb{E}(\log(X) - \log(t) \mid X > t) = \frac{\int_t^\infty \frac{L(u)}{u^{\alpha+1}} du}{\frac{L(t)}{t^\alpha}} \sim_\infty \frac{1}{\alpha}$$

On peut approcher $\mathbb{E}(\log(X) - \log(t) \mid X > t)$ par la loi des grands nombres :

$$\lim_{n\to\infty} \frac{\frac{1}{n} \sum_{i=1}^n \left(\log\left(X_i\right) - \log(t)\right) 1_{\{X_i>t\}}}{\frac{1}{n} \sum_{i=1}^n 1_{\{X_i>t\}}} = \mathbb{E}(\log(X) - \log(t) \mid X > t)$$

On remplace ensuite t par $X_{(n-k(n)+1,n)}$. D'après la section précédente $X_{(n-k(n)+1,n)}$ a même loi que $U\left(V_{(n-k(n)+1,n)}\right)$ et $\lim_n V_{(n-k(n)+1,n)} = \infty$. On en déduit que $\lim_n X_{(n-k(n)+1,n)} = \infty$. On obtient

$$\begin{aligned}
\hat{\xi}^H_{(k(n),n)} &= \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\log\left(X_i\right) - \log\left(X_{(n-k(n)+1,n)}\right)1_{\left\{X_i > X_{(n-k(n)+1,n)}\right\}}}{\frac{1}{n}\sum_{i=1}^{n}1_{\left\{X_i > X_{(n-k(n)+1,n)}\right\}}} \\
&= \frac{\frac{1}{n}\sum_{i=n-k(n)+1}^{n}\left(\log\left(X_{(i,n)}\right) - \log\left(X_{(n-k(n)+1,n)}\right)\right)}{\frac{k(n)-1}{n}} \\
&= \frac{1}{(k(n)-1)}\sum_{i=n-k(n)+1}^{n}\left(\log\left(X_{(i,n)}\right) - \log\left(X_{(n-k(n)+1,n)}\right)\right).
\end{aligned}$$

On conclut alors

$$\lim_{n\to\infty}\hat{\xi}^H_{(k(n),n)} = \frac{1}{\alpha} = \xi.$$

Annexe D : La famille exponentielle

Dans cette annexe, il va être démontré que les lois suivantes appartiennent bien à la famille exponentielle :

La Loi Normale :

Soit X $\sim N\left(\mu, \sigma^2\right)$ alors :

$$\begin{aligned}
f_X(y) &= \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2}\cdot\frac{(y-\mu)^2}{\sigma^2}} \\
&= \exp\left[\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}\cdot\left(y^2 - 2y\mu - \mu^2\right)\right] \\
&= \exp\left[\frac{y\cdot\mu - \mu^2/2}{\sigma^2} - \frac{1}{2}\cdot\left(\frac{y^2}{\sigma^2} + \ln\left(2\pi\sigma^2\right)\right)\right]
\end{aligned}$$

Avec $\mu = \theta$, $\sigma^2 = a(\phi) = \phi$, $\mu^2/2 = b(\theta) = \theta^2/2$ et la fonction $c(y, \theta)$ est représentée par la partie de droite de l'équation.

La Loi de Poisson :

Soit Y $\sim P(\lambda)$ alors :

$$f_Y(y) = \frac{\lambda^y}{y!} \cdot e^{-\lambda} = \exp[-\lambda + y \cdot \ln(\lambda) - \ln(y!)]$$
$$= \exp\left[\frac{y \cdot \ln(\lambda) - \lambda}{1} - \ln(y!)\right]$$

Avec $\ln(\lambda) = \theta$, $1 = a(\phi) = \phi$, $\lambda = b(\theta) = e^\theta$ et la fonction $c(y, \theta)$ est représentée par la partie de droite de l'équation, $c(y, \theta) = ln(y!)$.

La Loi de Gamma :

Soit $Y \sim \Gamma(\alpha, \lambda)$ alors :

$$f_Y(y) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \cdot e^{-\lambda y} \cdot y^{\alpha-1} \cdot \mathbb{1}_{\mathbb{R}^+}(y)$$

En posant $\beta = \frac{\alpha}{\lambda}$ et $\alpha = \frac{1}{\phi}$ on peut montrer que la loi gamma appartient bien à la famille exponentielle avec :

$$\theta = -\frac{1}{\beta} = -\frac{\alpha}{\lambda}, \quad a(\phi) = \phi = \frac{1}{\alpha},$$
$$b(\theta) = -\ln(-\theta) \text{ et } c(y, \phi) = \frac{\ln\left(\frac{y}{\phi}\right)}{\phi} - \ln(y) - \ln\left(\Gamma\left(\frac{1}{\phi}\right)\right)$$

Annexe E : Les autres résultats des modèles

Indicateurs des statistiques de modèles retenus

Indicateur	Base train complète	Base train K-FOLD*	Base test K-FOLD*
Gini	34,53%	34,53%	34,56%
Pseudo-R^2	5,35%	5,36%	5,32%
RMSE	0.2007	0.201	0.1997
Deviance	134 800	101 100	33 710

TABLE 7.5 – Indicateurs du modèle Tarif 2022 + SRA de fréquence des sinistres forfaitaires RCM

Indicateur	Base train complète	Base train K-FOLD	Base test K-FOLD
Gini	31,4%	31,4%	31,4%
Pseudo-R^2	4,86%	4,86%	4,86%
RMSE	0.2672	0.2672	0.2672
Deviance	191 900	143 900	47 980

TABLE 7.6 – Indicateurs du modèle Tarif 2022 + SRA de fréquence des sinistres non forfaitaires RCM

Indicateur	Base train complète	Base train K-FOLD	Base test K-FOLD
Gini	37,97%	37,98%	38,04%
Pseudo-R^2	4,34%	4,34%	4,35%
RMSE	0.07524	0.07524	0.07523
Deviance	28 460	21 350	7 116

TABLE 7.7 – Indicateurs du modèle Tarif 2022 + SRA de fréquence des sinistres RCC

Indicateur	Base train complète	Base train K-FOLD	Base test K-FOLD
Gini	13,48%	13,53%	13,54%
Pseudo-R^2	4,14%	4,14%	4,12%
RMSE	2 145	2 145	2 145
Deviance	20 610	15 460	5 153

TABLE 7.8 – Indicateurs du modèle SIV + SRA du coût moyen des sinistres RCM

Courbe de Lorenz pour la comparaison de la nouvelle modélisation comparée à l'ancienne

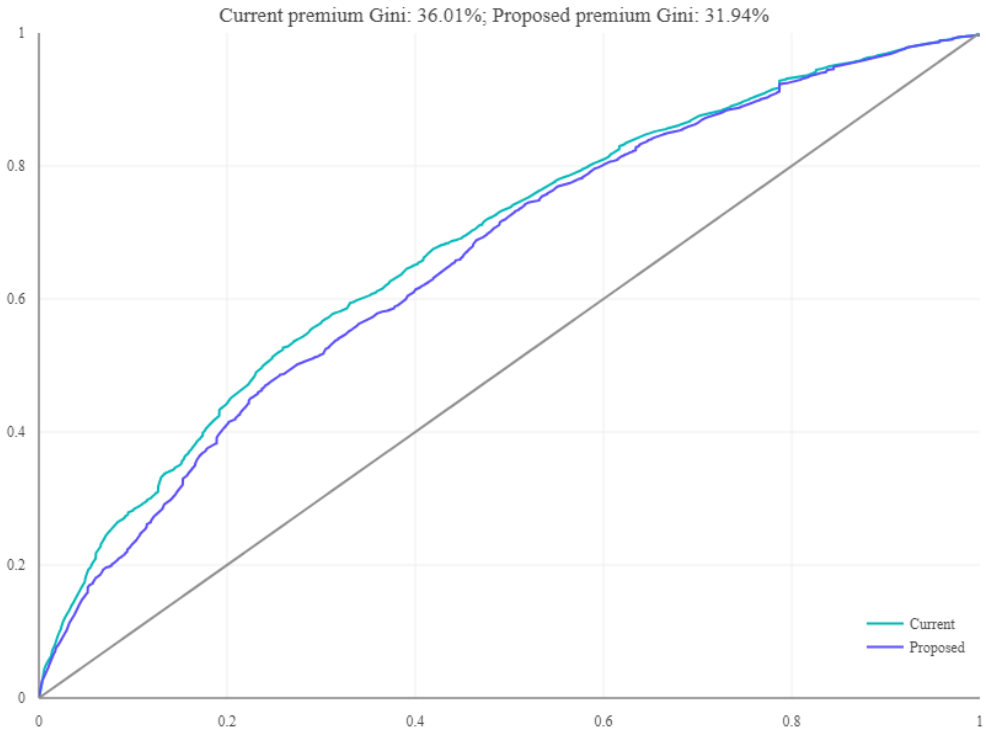


FIGURE 7.24 – Courbe de Lorenz avec en bleu foncé la nouvelle proposition de modélisation (Proposed)

Bibliographie

[Berson, 2021] Berson, E. (2021). Refonte de la garantie responsabilité civile automobile du produit garages.

[Boyer Chammard, 2008] Boyer Chammard, R. (2008). Processus de surveillance et de majoration des contrats flottes d'entreprise d'axa france.

[Charpentier et Denuit, 2005] Charpentier, A. et Denuit, M. (2005). *Mathématiques de l'assurance non-vie - tome 2*. Economica.

[Chatelain, 2020] Chatelain, P. (2020). Integrating data quality into GLM for insurance pricing. working paper or preprint.

[Cours EURIA Gruchet Xavier, 2023] Cours EURIA Gruchet Xavier (2023). Mathématiques des assurances iard et santé.

[Cours EURIA Laily Romain, 2023] Cours EURIA Laily Romain (2022-2023). Applications sur r en tarification non-vie.

[Cours Université Gustave Eiffel Clément Emmanuelle, 2022] Cours Université Gustave Eiffel Clément Emmanuelle (2022). Estimation empirique – valeurs extrêmes.

[De Jong et Heller, 2008] De Jong, P. et Heller, G. J. (2008). *Generalized linear models for insurance data*. International Serires on Actuairal Science, Cambridge University Press.

[Kim et Delecroix, 2009] Kim, E. Y. et Delecroix, M. (2009). Modèles de régression à directions révélatrices Applications en assurance non-vie. *In 41èmes Journées de Statistique, SFdS, Bordeaux*, Bordeaux, France, France.

[McCullagh et Nelder, 1989] McCullagh, P. et Nelder, N. (1989). *Generalized linear models, second ed*. Chapman and Hall.

[Nakache et Guegen, 2005] Nakache, J. et Guegen, A. (2005). Analyse multidimensionnelle de données incomplètes. *Rev. Statistique Appliquée*.

[Paglia et Phélippé-Guinvarc'h, 2011] Paglia, A. et Phélippé-Guinvarc'h, M. V. (2011). Tarification des risques en assurance non-vie, une approche par modèle d'apprentissage statistique. *Bulletin français d'actuariat*, (Juillet-Décembre).

[Schorgen, 2012] Schorgen, A. (2012). *Valeurs extrêmes : covariables et cadre bivarié*. Theses, Université de Strasbourg.

[Tomas, 2014] Tomas, J. (2014). Modèles linéaires glms : Analyse logit régression de poisson - analyse d'un portefeuille d'assurance - algorithme irwls avec r.

Table des figures

1.1 Répartition des cotisations dans l'assurance de biens et de responsabilité. Source : Données clés 2021 FFA 3
1.2 Part du Chiffre d'affaire AXA IARD Entreprise en 2021 5
1.3 Tableau de composition des flottes ouvertes 5

2.1 Types de véhicules impliqués 8

3.1 Schéma synthèse formation base véhicule 18
3.2 Charge moyenne des sinistres par garanties par année 20
3.3 Fréquence des sinistres par garanties par année 21

4.1 Choix de la méthode de discrétisation pour chaque variable quantitative . 25
4.2 Poids à vide du véhicule avant discrétisation 26
4.3 Poids à vide du véhicule après discrétisation utilisant la méthode XGBoost 26
4.4 Poids à vide du véhicule après discrétisation utilisant la méthode des quantiles 27
4.5 Discrétisation finale du poids à vide SRA utilisant la méthode des quantiles 28
4.6 Variables disponible dans les bases de modélisation 29
4.7 Densité GEV : Weibull ($\xi = -1$, darkblue), Gumbel (red) et Fréchet ($\xi = 1$, blue) 31
4.8 Estimateur de Pickands, construit manuellement 34
4.9 Estimateur de Hill, construit grâce à la fonction *hillplot* 34
4.10 Mean Excess Plot, construit manuellement 35
4.11 Seuils déterminé avec la méthode des quantiles 35
4.12 Loi du coût moyen de la garantie RCM 37
4.13 Coût moyen des sinistres de la garantie RCM 38
4.14 Fréquence des sinistres forfaitaires et non forfaitaires de la garantie RCM . 38
4.15 Coût moyen des sinistres de la variable concernant le modèle des véhicules SRA 39

5.1 Taux de manquants par contrat sur la localisation des sinistres 41
5.2 Coût moyen des sinistres sur la garantie RCM en fonction de la variable Tranche_dist_moy 43
5.3 Coût moyen des sinistres garantie RCC en fonction de la variable Tranche_dist_moy 43

5.4 Fréquence des sinistres forfaitaires (f) et hors forfait (nf) - garantie RCC en fonction de la variable Tranche_dist_moy 44
5.5 Fréquence des sinistres - garantie RCC en fonction de la variable Tranche_dist_moy 44

6.1 Corrélogramme des variables catégorielles 55
6.2 Corrélogramme des variables quantitatives 56

7.1 Modules AKUR8 62
7.2 Représentation graphique de la courbe de Lorenz (source AKUR8) 63
7.3 Représentation de la Lift curve (source AKUR8) 64
7.4 Grid search des modèles de fréquence non forfaitaires utilisant les variables SRA et SIV 66
7.5 Variables pour la prédiction de la fréquence des sinistres non forfaitaires - Modélisation SIV+SRA 67
7.6 Schéma de la validation croisée 68
7.7 Grid search des modèles de fréquence non forfaitaires utilisant les variables SRA uniquement 69
7.8 Variables pour la prédiction de la fréquence des sinistres non forfaitaires - Modélisation SRA 69
7.9 Grid search des modèles de fréquence non forfaitaires utilisant la modélisation 2022 avec ajout des variables SRA 71
7.10 Variables pour la prédiction de la fréquence des sinistres non forfaitaires - Modélisation tarif 2022 + SRA 72
7.11 Résultats des modèles 73
7.12 Double lift curve - Comparaison nouvelle modélisation, tarif 2022 et observé 74
7.13 Variable 1 75
7.14 Variable 2 76
7.15 Variable 3 76
7.16 Distance moyenne des sinistres depuis l'adresse du souscripteur, par tranche 77
7.17 Graphique de la densité des écarts 78
7.18 Densité des écarts coupés au quantile 96% 79
7.19 Nuage de points de la prime pure Parc et de la prime pure FND 79
7.20 Primes pures moyennes de la variable 1 80
7.21 Primes pures moyennes de la variable 2 81
7.22 Primes pures moyennes de la variable tranche de distance moyenne des sinistres 81
7.23 Primes pures moyenne en fonction de la variable 3 82
7.24 Courbe de Lorenz avec en bleu foncé la nouvelle proposition de modélisation (Proposed) 93

Liste des tableaux

2.1 Montant à payer selon le délai et le mode de paiement 10

3.1 Listes des variables SIV 16
3.2 Listes des variables SRA 17
3.3 Sinistres par garanties sur les années 2019 à 2021 19

4.1 Définition de la variable *SEGMENT_SRA* 39

5.1 Récapitulatif des choix de prise en compte de la variable Tranche_dist_moy 45

6.1 Famille exponentielle 50
6.2 Fonctions de lien 51
6.3 Expression des modèles en fonction des lois 51
6.4 Sélection des variables pour la base de modélisation du coût moyen des sinitres de la garantie RCM 58

7.1 Indicateurs du modèle SIV + SRA de fréquence des sinistres non forfaitaires RCM 67
7.2 Indicateurs du modèle SRA de fréquence des sinistres non forfaitaires RCM 70
7.3 Indicateurs du modèle Tarif 2022 + SRA de fréquence des sinistres non forfaitaires RCM 72
7.4 KPI des primes pures moyennes 74
7.5 Indicateurs du modèle Tarif 2022 + SRA de fréquence des sinistres forfaitaires RCM 91
7.6 Indicateurs du modèle Tarif 2022 + SRA de fréquence des sinistres non forfaitaires RCM 92
7.7 Indicateurs du modèle Tarif 2022 + SRA de fréquence des sinistres RCC . 92
7.8 Indicateurs du modèle SIV + SRA du coût moyen des sinistres RCM . . . 92