Mémoire présenté devant l'Université de Paris-Dauphine
pour l'obtention du Certificat d'Actuaire de Paris-Dauphine
et l'admission à l'Institut des Actuaires
le

Par : Oksana ALLAIRE

Titre : Comparaison de différentes méthodes pour la modélisation de la prime pure d'un produit risque aggravé en assurance automobile

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

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

Membres présents du jury de l'Institut des Actuaires

Entreprise :
Nom : Generali France – L'ÉQUITÉ
Signature :

Membres présents du jury du Master Actuariat de Paris Dauphine

Directeur de mémoire en entreprise :
Nom : Amin TOUSSI
Signature :

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

Secrétariat :

Signature du responsable entreprise

Bibliothèque :

Signature du candidat

Remerciements

Je tiens tout d'abord à remercier Monsieur Amin TOUSSI, mon tuteur de stage et manager de l'équipe Solutions d'assurance auto à la Direction des Partenariats de Generali.

J'adresse également mes remerciements à l'ensemble de l'équipe auto pour leur accueil et le partage de leurs connaissances.

Je remercie mon tuteur académique Monsieur André GRONDIN pour son suivi et ses conseils.

Je souhaite remercier le corps enseignant de l'Université Paris-Dauphine et de l'INSA Rouen pour leur formation d'excellente qualité ainsi que Monsieur Nicolas FORCADEL qui m'a donné la possibilité de suivre un double cursus.

Enfin, j'exprime ma reconnaissance à toutes les personnes qui m'ont soutenue durant ma scolarité et la rédaction de ce mémoire.

Résumé

Depuis 1958, la garantie « Responsabilité civile » de l'assurance automobile est obligatoire en France. Ce marché représente un secteur très compétitif. En effet, de nombreux acteurs (agents généraux d'assurance, courtiers, mutuelles, bancassureurs, vente directe, etc…) proposent des tarifs de plus en plus compétitifs et attractifs pour les clients.

Dans le but d'améliorer continuellement les tarifs proposés, tout en restant solvables, les assureurs doivent affiner leur segmentation. Celle-ci consiste à créer des classes de risques homogènes qui regroupent des individus pour lesquels la sinistralité est similaire, et qui paieront donc la même prime.

Afin de réduire le montant des sinistres payés, les assureurs cherchent à attirer les profils les moins risqués. Par conséquent, certains conducteurs peuvent se voir refuser la souscription d'une assurance, ou se faire résilier leur assurance en cours en cas de risque trop important (coefficient de réduction-majoration (CRM) trop élevé, non-paiement de prime, sinistres trop fréquents, fausse déclaration, graves infractions au Code de la route (conduite en état d'ébriété, conduite sous l'emprise de stupéfiants, délit de fuite, excès de vitesse)). Cependant, ces conducteurs sont tout de même tenus à l'obligation d'assurance exigée par le code de la Route. C'est la raison pour laquelle il est intéressant de savoir tarifer correctement ces risques dits « aggravés ».

L'objectif de ce mémoire est d'analyser et de comparer plusieurs méthodes de tarification sur le marché du risque aggravé.

Dans un premier temps, le marché de l'assurance automobile en France, ses spécificités et ses principaux acteurs sont exposés. Puis, après avoir analysé la base de données à disposition, la modélisation de la prime pure est développée pour les bris de glace et dommages tous accidents. Le modèle linéaire généralisé, qui est le modèle le plus utilisé par les assureurs pour tarifer leurs produits est présenté et appliqué aux données. Ensuite, certaines méthodes d'apprentissage statistique sont abordées : les arbres de classification et de régression CART et les forêts aléatoires. Enfin, la dernière partie de ce mémoire est consacrée à la comparaison des différents résultats obtenus.

<u>Mots clés</u> : Tarification, assurance automobile, risque aggravé, modèle linéaire généralisé (GLM), apprentissage statistique, arbres de classification et régression CART, forêts aléatoires.

Abstract

Since 1958 in France, the "Third Party Liability" coverage of car insurance policies is compulsory. This market is a very competitive sector. Indeed, many parties (general insurance agents, insurance brokers, mutual insurance companies, bank-insurers, direct sales, etc. ...) offer increasingly competitive and attractive rates for customers.

In order to continuously improve the rates offered, while remaining solvent, insurers must refine their segmentation. This consists in creating homogeneous risk classes that bring together people for whom the loss experience is similar, and who will therefore pay the same premium.

In order to reduce the amounts paid for the claims, insurers seek to attract least risky profiles. Therefore, some drivers may be refused by insurances, or their current contract may be terminated if the risk is too high (excessively high bonus-malus, non-payment of premium, too frequent claims, fraudulent misrepresentation or serious traffic offenses like drunk-driving, drug-driving, hit-and-run and speeding). However, even these drivers are required, under the French Highway Code, to subscribe to an insurance policy. This is the reason why it is interesting to know how to correctly price these so-called "aggravated" risks.

The purpose of this writings is to analyze and compare several pricing methods in the aggravated risk market.

First, the car insurance market in France, its specificities and its main parties are going to be described. Then, after analyzing the database, pure premium modelling is developed for glass breakage and all-accident damages. The generalized linear model, which is the model most used by insurers to price their products, is presented and applied to the data. Then, some Machine Learning methods are introduced: classification and regression trees and random forests. Finally, the aim of the last part of this writings is the comparison of the different obtained results.

Keywords: Pricing, car insurance, aggravated risk, generalized linear model (GLM), Machine Learning, classification and regression trees (CART), random forests.

Note de synthèse

Contexte :

Afin de faire face à la forte concurrence en assurance automobile, les assureurs doivent à la fois proposer des tarifs attractifs, tout en veillant à leur solvabilité. La tarification représente ainsi une opération importante lors du lancement d'un produit et le suivi de sa rentabilité.

L'enjeu majeur de la tarification consiste à déterminer la prime pure, c'est-à-dire le montant probable des sinistres. Ce montant est réparti pour l'ensemble des assurés. Il s'agit de la mutualisation des risques. L'approche tarifaire communément utilisée par les actuaires IARD consiste à décomposer la charge en produit de la fréquence et du coût. Bien que l'assurance repose sur le principe de mutualisation, il est nécessaire de proposer une prime adaptée au risque associé à chacun des assurés afin d'éviter tout phénomène d'antisélection. De ce fait, les assureurs cherchent constamment à affiner leur segmentation.

Les assureurs cherchent à proposer des tarifs compétitifs, mais aussi à réduire le risque de leur portefeuille. Par conséquent, les compagnies d'assurance cherchent à attirer les profils les moins risqués afin de réduire le montant des sinistres payés. Ainsi, il leur est possible de refuser lors de la souscription certains individus qu'ils jugent trop exposés au risque ou, sous certaines conditions, de résilier un contrat d'assurance notamment lorsque l'assuré n'a pas payé sa prime, à la suite d'une fausse déclaration volontaire de la part de l'assuré, pour cause d'aggravation du risque, après un accident en état d'ivresse, sous l'emprise de stupéfiant ou ayant entraîné une suspension ou une annulation de permis, ou bien à l'échéance du contrat. Toutefois, tout véhicule est soumis à l'obligation d'assurance imposé par la réglementation française. Pour cette raison, il est intéressant de savoir tarifer correctement les risques dits « aggravés » afin d'investir cette niche

Ainsi, les produits d'assurance dits « risques aggravés » permettent aux conducteurs novices ainsi qu'aux conducteurs ayant un CRM élevé, des antécédents d'alcoolémie, une suspension ou annulation de permis, des interruptions d'assurance, une résiliation de contrat par un assureur précédent ou des antécédents de sinistralité très importants, de pouvoir souscrire une assurance automobile.

L'objectif du mémoire est de comparer différentes méthodes de modélisation de la prime pure, afin d'estimer si, face à un risque non standard, les modèles classiquement utilisés par les assureurs présentent de bonnes performances, ou si, au contraire, ils peuvent être améliorés grâce à des méthodes plus innovantes.

Quel que soit la méthode de modélisation implémentée, il est impératif de fixer un compromis entre le biais et la variance. Le biais correspond à l'erreur de modélisation de l'algorithme (sous-apprentissage, mauvaises hypothèses, etc…) et la variance est due aux variations de l'échantillon d'apprentissage (sur-apprentissage, stabilité du modèle). Ainsi, un modèle doit un bon prédicteur mais il ne doit pas pour autant être trop complexe afin d'éviter le sur-apprentissage.

Données utilisées :

Avant toute modélisation, il est nécessaire de construire une base de données regroupant les caractéristiques des assurés et de leur véhicule, ainsi que leur cause d'aggravation. Ces différents risques sont ensuite fusionnés avec la base de sinistralité afin de modéliser le coût et la fréquence des sinistres de deux garanties : le bris de glace et les dommages au véhicule.

Les données étudiées sont issues du portefeuille des produits d'assurance automobile aggravés de cinq courtiers travaillant avec la Direction des Partenariats de Generali. Les produits d'assurance étant conçus

sur-mesure pour les courtiers, les conditions de souscription et les primes actuellement proposés varient légèrement. Cependant, afin du justifier la pertinence de ce regroupement une étude est menée afin d'analyser la composition du portefeuille et la sinistralité. Les objectifs de ce regroupement sont multiples. Il permet de rassembler plus de données afin de garantir la robustesse des modèles et améliorer les tarifs actuellement proposés pour qu'ils correspondent de manière la plus juste possible au risque réel des conducteurs et de leur véhicule. Il permet également d'avoir un tarif commun à l'ensemble des courtiers proposant des contrats « risque aggravé » afin d'élargir les conditions de souscription des polices déjà sur le marché, mais aussi de proposer une prime pour le lancement de nouveaux partenariats avec des courtiers n'ayant pour le moment aucune donnée.

L'étude menée sur la base de données permet de montrer que comme les conditions de souscription diffèrent selon les courtiers, ceux-ci ne sont pas apporteurs des mêmes types de conducteurs. Ainsi, le distributeur A est principalement apporteur de conducteurs ayant eu une suspension de permis. Le courtier B possède essentiellement des jeunes conducteurs et ayant peu d'expérience de conduite. Le courtier C est apporteur d'assurés ayant des antécédents importants de sinistralité (sinistres corporels et matériels). Une grosse proportion des assurés du courtier D sont en situation de malus. Enfin, de nombreux assurés du courtier E ont été résilié par leur précédent assureur à la suite d'un non-paiement de prime. Cependant, les caractéristiques des véhicules assurés et les zones géographiques du risque sont les mêmes pour l'ensemble des courtiers apporteurs. Cela permet de conclure qu'il est pertinent de concaténer les bases des cinq courtiers puisque hormis les conditions de souscription qui diffèrent, les bases présentent de nombreux points communs.

La deuxième partie de l'étude de la base de données consiste à prouver que la sinistralité est due aux caractéristiques intrinsèques des assurés, et non au courtier apporteur. Si tel n'était pas le cas, le regroupement des cinq bases ne serait pas cohérent. Cette analyse permet également de constater que la sinistralité du portefeuille « aggravé » est plus dégradée que la sinistralité nationale. Parmi ces risques élevés, il est possible d'évaluer les caractéristiques du risque les plus aggravantes. Ainsi, les assurés vivant dans des zones les plus risquées d'après le zonier ont une fréquence et un coût de sinistralité plus élevés. De manière générale, les caractéristiques du véhicule ont un impact sur le coût des sinistres tandis que les caractéristiques des conducteurs influent sur la fréquence. Ainsi, plus la classe et le groupe SRA sont grands, plus la fréquence et le coût moyen des sinistres le sont également. En revanche, le coût moyen des vieux véhicules est inférieur à celui des véhicules neufs. Par ailleurs, les profils d'assurés ayant les plus hauts taux de sinistralité sont les artisans, les commerçants et les professions libérales, les jeunes conducteurs et ceux ayant obtenu leur permis de conduire récemment. De plus, les antécédents de sinistralité sont des éléments importants puisque plus le CRM est élevé, plus la fréquence et le coût moyen sont importants. Il en est de même pour les assurés ayant eu un grand nombre de sinistres au cours des trois dernières années et ceux ayant fait l'objet d'une annulation de permis.

Modèle linéaire généralisé :

Le modèle linéaire généralisé est la méthode de modélisation communément utilisée par les actuaires pour tarifer les produits d'assurance IARD pour les particuliers. Dans ce mémoire, quatre modèles linéaires généralisés sont implémentés, pour la fréquence et le coût des deux garanties étudiées : le bris de glace et les dommages au véhicule.

Pour implémenter ces modèles, une loi est sélectionnée grâce aux critères AIC et de déviance. Pour les deux modèles de fréquence, la loi binomiale négative est privilégiée et pour les deux modèles de coût, la loi Gamma est choisie. Seules les variables tarifaires significatives sont gardées dans le modèle. Cette sélection est effectuée grâce à un test de Wald.

Afin de faciliter l'implémentation pour les courtiers, le lien logarithmique est privilégié, permettant ainsi d'avoir un modèle multiplicatif.

La cohérence des résultats est vérifiée en comparant les tendances des coefficients obtenus grâce aux modèles aux tendances de sinistralités remarquées dans l'analyse descriptive de la sinistralité.

Afin de valider les différents modèles, plusieurs critères sont contrôlés. Tout d'abord, la déviance est calculée, puis, deux types de résidus sont examinés, les résidus de Pearson et les résidus quantiles normalisés.

Arbre de régression et de classification :

Un inconvénient des modèles linéaires généralisé vient du fait qu'ils sont paramétriques. Il est parfois complexe de trouver une loi s'adaptant aux données et le choix des variables significatives peut également s'avérer difficile. Face à cette limite, les arbres de régression peuvent être implémentés. Il s'agit d'une méthode de Machine Learning visant à partitionner récursivement les variables explicatives.

La construction d'un arbre se décompose en deux étapes. Tout d'abord, un arbre saturé est construit, puis, un arbre optimal, élagué est obtenu, grâce à une validation croisée.

Quatre arbres optimaux sont obtenus : pour la fréquence et le coût des garanties bris de glace et dommages.

Forêt aléatoire :

Le principal défaut des arbres est leur manque de robustesse. Pour limiter ce problème, il est possible d'utiliser des méthodes d'agrégation. L'idée principale consiste à construire un grand nombre d'arbres puis à les réunir afin d'obtenir une meilleure estimation. Cette méthode permet de réduire la variance de l'estimation et ainsi d'obtenir des résultats plus robustes.

Le principe des forêts aléatoires repose sur l'agrégation d'une combinaison d'arbres, apportant une plus grande efficacité. Les arbres sont construits de manière à être le plus décorrélés possible, de sorte qu'ils soient indépendants entre eux, ce qui permet de réduire la variance du modèle. Pour cela, l'algorithme procède à un double échantillonnage : à la fois sur les observations (*tree bagging*, les données sont rééchantillonnées), mais aussi sur les variables (*feature sampling*, le choix de la division optimale s'effectue sur un sous-ensemble des variables explicatives*).*

Pour chacune des quatre forêts modélisées, il est possible de déterminer l'importance des variables.

Comparaison des modèles :

Dans un premier temps, les variables les plus significatives pour les modèles sont comparées. Leur importance est déterminée grâce à la valeur de la p-value pour le modèle linéaire généralisé, grâce à leur position dans l'arbre pour l'algorithme CART et grâce à l'erreur Out-of-Bag pour la forêt aléatoire.

Afin de vérifier que la tarification issue des différents modèles n'est ni surévaluée ni sous-évaluée, la somme des primes demandées aux assurés est simulée et comparée à la charge réellement payée.

Puis, pour vérifier que la segmentation est suffisamment affinée, les courbes de Lorenz sont tracées pour les douze modèles. De manière générale, le modèle linéaire généralisé et la forêt aléatoire permettent la même forme de segmentation. Les arbres CART segmentent moins bien les risques, ce qui peut être expliqué par le peu de feuilles conservées après la phase d'élagage.

Afin de tester la qualité des modèles et dans le but de vérifier qu'ils se généralisent correctement, l'erreur quadratique moyenne est calculée sur une base de test, indépendante de la base de modélisation.

La comparaison des différents modèles permet de conclure que les modèles linéaires généralisés, les arbres CART et les forêts aléatoires modélisent correctement la fréquence et le coût des sinistres Bris de glace et Dommages au véhicule. Cependant, les modèles linéaires généralisés et les forêts aléatoires présentent les meilleurs résultats.

Conclusion :

Le mémoire réalisé permet de modéliser la prime pure de deux garanties en utilisant plusieurs méthodes de tarification : les modèles linéaires généralisés et deux méthodes de Machine Learning, les arbres de régression CART et les forêts aléatoires. Ces modélisations permettent à la Direction des Partenariats de proposer des primes segmentées et adaptées aux risques des assurés. Les forêts aléatoires proposent des résultats qui se généralisent mieux aux données de la base test. Cependant, lorsqu'une loi s'adapte bien à la distribution de la variable réponse, les modèles linéaires généralisés proposent également une bonne modélisation des risques. Ils ont de plus l'avantage d'être facilement interprétables et de pouvoir être implémentés informatiquement facilement car ils fournissent une prime de référence et des coefficients multiplicatifs. Par ailleurs, le temps de calcul des modèles linéaires généralisés est inférieur à celui nécessaire aux méthodes de Machine Learning.

Synthesis note

Context:

In order to face strong competition in automobile insurance, insurers must both offer attractive rates, while ensuring their solvency. Pricing represents an important operation when launching a product and monitoring its profitability.

The major issue of pricing is to determine the pure premium, that is to say the probable amount of claims. This amount is distributed for all insured persons. It is the pooling of risks. The pricing approach commonly used by property and casualty actuaries is to break down the charge into a product of frequency and cost. Although insurance is based on the principle of pooling, it is necessary to offer a premium adapted to the risk associated with each policyholder to avoid any adverse selection phenomenon. As a result, insurers are constantly seeking to refine their segmentation.

Insurers seek to offer competitive rates, but also to reduce the risk of their portfolio. As a result, insurance companies seek to attract the least risky profiles in order to reduce the amount of claims paid. Thus, it is possible for them to refuse at the time of the subscription certain individuals that they consider too exposed to the risk or, under certain conditions, to terminate an insurance contract in particular when the insured did not pay his premium, to the following a deliberate misrepresentation by the insured, because of increased risk, after a drunken accident, under the influence of a narcotic or resulting in a suspension or cancellation of a license, or well at the end of the contract. However, any vehicle is subject to the insurance obligation imposed by French regulations. For this reason, it is interesting to know how to correctly price the so-called "aggravated" risks in order to invest this niche.

For example, "aggravated risk" insurance products allow novice drivers as well as drivers with high CRM, a history of alcohol abuse, suspension or cancellation of licenses, insurance interruptions, termination of contract by a previous insurer or a very significant loss history, to be able to take out automobile insurance.

The objective of the thesis is to compare different methods of pure premium modeling, in order to estimate if, compared to a non-standard risk, the models conventionally used by insurers show good performances, or if, on the contrary, they can be improved through more innovative methods.

Regardless of the modeling method implemented, it is imperative to set a trade-off between bias and variance. The bias corresponds to the algorithm's modeling error (under-learning, bad assumptions, etc.) and the variance is due to variations in the learning sample (over-learning, model stability). Thus, a model must be a good predictor, but it must not be too complex to avoid over-learning.

Data:

Before modeling, it is necessary to build a database of the characteristics of the insured and their vehicle, as well as their cause of aggravation. These various risks are then merged with the loss base in order to model the cost and the frequency of claims for two warranties: broken glass and damage to the vehicle.

The data studied comes from the portfolio of car insurance products aggravated by five brokers working with Generali's Partnerships Directorate. Because insurance products are custom-designed for brokers, the underwriting conditions and premiums currently offered vary slightly. However, to justify the relevance of this grouping, a study is conducted to analyze the composition of the portfolio and the loss experience. The objectives of this grouping are multiple. It allows more data to be collected in order to guarantee the robustness of the models and improve the rates currently proposed so that they correspond as accurately as possible to the real risk of drivers and their vehicles. It also allows for a common tariff for all brokers

offering "aggravated risk" contracts in order to broaden the underwriting conditions for policies already on the market, but also to offer a bonus for the launch of new partnerships with Brokers have no data for the moment.

The study conducted on the database shows that since the subscription conditions differ depending on the broker, they are not contributors of the same types of drivers. Thus, the distributor A is primarily a driver of drivers who had a license suspension. Broker B has mostly young drivers with little driving experience. Broker C is a contributor of policyholders with a significant history of claims (personal and material losses). A large proportion of D broker's policyholders are in a bad situation. Finally, many of the insureds of Broker E were terminated by their previous insurer as a result of a non-payment of premium. However, the characteristics of insured vehicles and geographical areas of risk are the same for all brokers. This makes it possible to conclude that it is relevant to concatenate the five brokers' bases since, apart from the different subscription conditions, the bases have many points in common.

The second part of the database study consists in proving that the loss experience is due to the intrinsic characteristics of the policyholders, and not to the contributing broker. If this were not the case, the grouping of the five bases would not be coherent. This analysis also shows that the loss experience of the "aggravated" portfolio is worse than the national loss experience. Among these high risks, it is possible to assess the most aggravating risk characteristics. Thus, policyholders living in riskier areas according to the zone have a higher frequency and cost of loss. In general, the characteristics of the vehicle have an impact on the cost of claims while the characteristics of the drivers affect the frequency. Thus, the larger the SRA class and group, the higher the frequency and average cost of claims. On the other hand, the average cost of old vehicles is lower than that of new vehicles. In addition, the profiles of policyholders with the highest loss ratios are craftsmen, tradesmen and the liberal professions, young drivers and those who have recently obtained their driving license. In addition, historical loss experience is important because the higher the CRM, the higher the frequency and the average cost. The same applies to policyholders who have had many claims in the last three years and those who have been the subject of a license cancellation.

Generalized linear model:

The generalized linear model is the modeling method commonly used by actuaries to price property and casualty insurance products for individuals. In this thesis, four generalized linear models are implemented, for the frequency and the cost of the two guarantees studied: the breaking of ice and the damage to the vehicle.

To implement these models, a law is selected using AIC and deviance criteria. For both frequency models, the negative binomial distribution is preferred and for both cost models, the Gamma distribution is chosen. Only significant tariff variables are kept in the model. This selection is made through a Wald test.

To facilitate the implementation for brokers, the logarithmic link is privileged, allowing to have a multiplicative model.

The consistency of the results is verified by comparing the trends of the coefficients obtained by the models with the loss trends observed in the descriptive analysis of the loss experience.

In order to validate the different models, several criteria are checked. First, the deviance is calculated, then two types of residues are examined, the Pearson residuals and the standardized quantile residues.

Regression and classification tree:

A disadvantage of generalized linear models is that they are parametric. It is sometimes complex to find a law that adapts to the data and the choice of significant variables can also be difficult. Faced with this

limitation, regression trees can be implemented. This is a Machine Learning method that attempts to recursively partition explanatory variables.

The construction of a tree breaks down in two stages. First, a saturated tree is built, and then, an optimal tree, pruned is obtained, through cross-validation.

Four optimal trees are obtained: for the frequency and the cost of the broken guarantees of ice and damages.

Random forest:

The main defect of the trees is their lack of robustness. To limit this problem, it is possible to use aggregation methods. The main idea is to build many trees and then bring them together to get a better estimate. This method makes it possible to reduce the variance of the estimate and thus to obtain more robust results.

The principle of random forests is based on the aggregation of a combination of trees, bringing greater efficiency. Trees are constructed to be as decorrelated as possible, so that they are independent of each other, reducing the variance of the model. For that, the algorithm carries out a double sampling: on both the observations (tree bagging, the data are resampled), but also on the variables (feature sampling, the choice of the optimal division is carried out on a sub-sampling set of explanatory variables).

For each of the four modeled forests, it is possible to determine the importance of the variables.

Comparison of models:

In a first step, the most significant variables for the models are compared. Their importance is determined by the value of the p-value for the generalized linear model, thanks to their position in the tree for the CART algorithm and thanks to the out-of-bag error for the random forest.

In order to verify that the pricing resulting from the different models is neither overvalued nor undervalued, the sum of the premiums requested from the insureds is simulated and compared to the charge actually paid.

Then, to verify that the segmentation is sufficiently refined, the Lorenz curves are plotted for the twelve models. In general, the generalized linear model and the random forest allow the same form of segmentation. CART trees have a lower risk profile, which can be explained by the few leaves left after the pruning phase.

In order to test the quality of the models and in order to verify that they are generalizing correctly, the mean squared error is calculated on a test basis, independent of the modeling basis.

A comparison of the different models leads to the conclusion that generalized linear models, CART trees, and random forests correctly model the frequency and cost of Ice Breakdown and Vehicle Damage claims. However, generalized linear models and random forests show the best results.

Conclusion:

The writing allows to get the pure premium of two guarantees by using several methods of pricing: the generalized linear models and two methods of Machine Learning, the regression trees CART and the random forests. These models allow the Partnerships Department to offer premiums segmented and adapted to the risks of policyholders. Random forests provide results that generalize better to the data in

the test database. However, when a law adapts well to the distribution of the response variable, generalized linear models also offer good risk modeling. Moreover, they have the advantage of being easily interpretable and of being able to be implemented easily because they provide a reference premium and multiplicative coefficients. In addition, the calculation time of generalized linear models is less than that required for Machine Learning methods.

Table des matières

Remerciements 3
Résumé 4
Abstract 5
Note de synthèse 6
Synthesis note 10
Introduction 16
Chapitre 1 : Contexte global 17
1.1 L'assurance automobile en France 17
1.1.1 Les assureurs français 17
1.1.2 La réglementation 18
1.1.3 Les garanties proposées 21
1.1.4 Le marché français 22
1.1.5 Le risque aggravé 24
1.2 Modélisation de la prime pure 25
1.2.1 Principes de tarification en assurance automobile 25
1.2.2 Problématiques liées à la modélisation 27
Chapitre 2 : Présentation et analyse des données 30
2.1 Construction de la base de données 30
2.1.1 Bases utilisées 30
2.1.2 Harmonisation et fusion des bases 31
2.1.3 Description des variables 32
2.1.4 Construction des bases d'étude 33
2.2 Analyse de la base de données 34
2.2.1 Analyses descriptives et retraitement éventuel des données 35
2.2.2 Dépendance des variables 52
2.2.3 Seuil d'écrêtement des sinistres graves 55
Chapitre 3 : Modèle linéaire généralisé 58
3.1 Théorie sur le modèle linéaire généralisé 58
3.1.1 Notations 58
3.1.2 Modèle linéaire classique 58
3.1.3 Les trois composantes d'un modèle linéaire généralisé 59
3.1.4 Calibration d'un modèle linéaire généralisé 61
3.1.5 Validation du modèle 63

3.2 Modélisation des garanties 65
3.2.1 Modélisation pour la garantie bris de glace 65
3.2.2 Modélisation pour la garantie dommages tous accidents 71
Chapitre 4 : Modèles de Machine Learning 76
4.1 Théorie 76
4.1.1 Arbre de régression CART 76
4.1.2 Forêt aléatoire 82
4.2 Modélisation 85
4.2.1 Modélisation pour la garantie bris de glace 85
4.2.2 Modélisation pour la garantie dommages tous accidents 90
Chapitre 5 : Analyse des résultats 96
5.1 Variables explicatives utilisées 96
5.2 Comparaison de la prime pure modélisée 98
5.3 Évaluation du caractère prédictif des modèles 103
Conclusion 106
Bibliographie 107
Annexes 108

Introduction

Face aux constantes évolutions de la réglementation (obligations des assureurs envers les assurés et durcissement des normes de solvabilité) et au développement des comparateurs d'assurance, le marché de l'assurance automobile est devenu très concurrentiel. Ce phénomène est accentué par le fait que la garantie « Responsabilité civile » est obligatoire en France depuis 1958, ce qui pousse de nombreux acteurs (agents généraux d'assurance, courtiers, mutuelles, bancassureurs, vente directe, etc…) à proposer des tarifs toujours plus compétitifs.

Proposer des primes défiant les offres concurrentes tout en restant solvables est devenu le défi des assureurs automobile. La tarification est une opération importante relative au lancement et au suivi d'un produit d'assurance. Les primes proposées aux assurés doivent répondre au principe de mutualisation des risques, tout en étant adaptées au risque associé à chacun des assurés. Ainsi, les assureurs doivent affiner leur segmentation afin de créer des classes de risques homogènes regroupant des individus dont la sinistralité est similaire.

Dans le but de réduire le montant des sinistres payés, les assureurs cherchent à attirer les profils les moins risqués. De ce fait, ils peuvent refuser lors de la souscription certains conducteurs qu'ils jugent trop risqués, notamment lorsque ces derniers ont des antécédents de sinistralité importants ou lorsqu'ils ont fait l'objet de procédures judiciaires (suspension ou annulation de permis, conduite sous l'emprise de l'alcool ou de stupéfiants, délit de fuite, etc…). Ils peuvent également résilier le contrat d'un assuré sous certaines conditions (fausse déclaration, non-paiement de prime, aggravation du risque, etc…). Malgré le fait que les assureurs restent souverains en matière d'appréciation des risques et de politique de souscription, tout véhicule mis en circulation doit être assuré. C'est la raison pour laquelle il est intéressant de savoir tarifer correctement les risques dits « aggravés » afin d'investir cette niche.

Les données étudiées dans le cadre de ce mémoire sont issues des produits d'assurance « aggravés » de cinq courtiers associés à la Direction des Partenariats de Generali. L'objectif est de proposer un tarif adapté aux profils risqués pour deux garanties : le bris de glace et les dommages au véhicule.

Dans un premier temps, le marché de l'assurance automobile en France, ses spécificités et ses principaux acteurs sont présentés. Puis, les grands principes de tarification sont rappelés. Le deuxième chapitre de ce mémoire est dédié à la construction et à l'analyse de la base de données. Cette étape est indispensable à la compréhension du portefeuille afin de connaitre les principales tendances de sinistralité. Elle permet également d'étudier la corrélation entre les variables et de fixer le seuil à partir duquel les sinistres sont considérés comme « graves ».

Après avoir exposé le contexte de l'étude, les deux chapitres suivants sont consacrés à la modélisation de la prime pure. Tout d'abord, les modèles linéaires généralisés, traditionnellement mis en œuvre par les assureurs en raison de leur performance et de leur robustesse sont présentés et appliqués aux données. Face au marché fortement concurrentiel et à la spécificité des produits d'assurance « aggravés », le but de ce mémoire est de tester si d'autres modèles peuvent être employés. Deux méthodes de Machine Learning sont alors implémentées : les arbres de régression CART et les forêts aléatoires.

Enfin, la dernière partie de ce mémoire est dédiée à la comparaison des différents modèles et à l'analyse des résultats obtenus.

Chapitre 1 : Contexte global

1.1 L'assurance automobile en France

1.1.1 Les assureurs français

L'assurance automobile représente le marché le plus important de l'assurance de biens. En effet, d'après le rapport annuel de la Fédération Française de l'Assurance (FFA), en 2018, l'ensemble de l'assurance française a généré 219 milliards d'euros de chiffre d'affaires, dont 56 milliards pour l'assurance de biens et de responsabilité. Avec 22 milliards d'euros de chiffre d'affaires, soit 39% de celui de l'assurance de biens et de responsabilité, le secteur de l'assurance automobile représente un marché extrêmement développé.

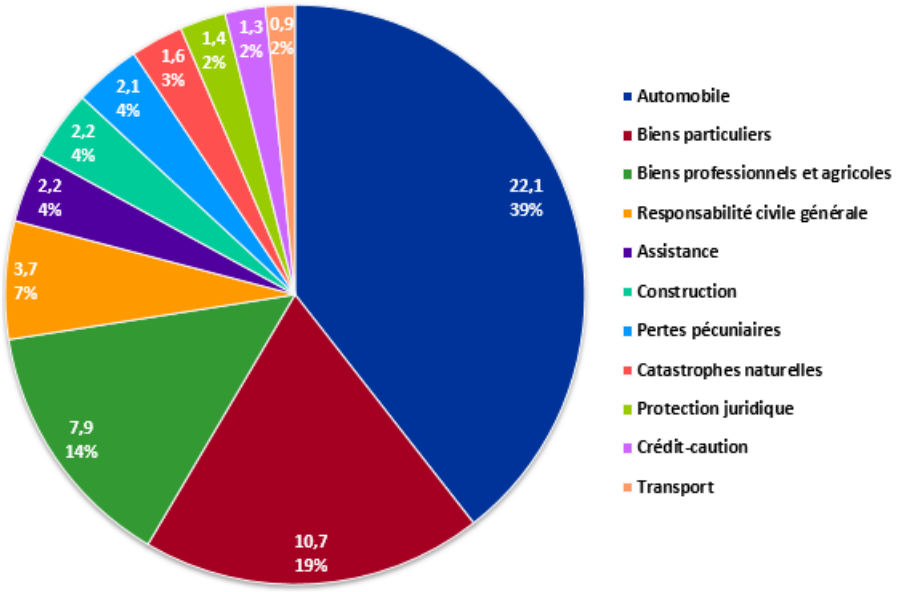

Figure 1 : Répartition des primes perçues en 2018 en milliards d'euros en assurance de biens et de responsabilité

En conséquence de la forte demande d'assurance automobile, de nombreux acteurs, tels que les agents généraux d'assurance, les courtiers, les mutuelles, les bancassureurs et les assureurs utilisant le canal de la vente directe, proposent des produits sur le marché de l'assurance automobile. De ce fait, ce secteur subit une concurrence très forte. En effet, une centaine d'acteurs opèrent sur ce marché. Cependant, les dix premiers groupes comptabilisent à eux seuls près de 80% du marché français de l'assurance automobile.

Le tableau page suivante énumère les 10 meilleurs assureurs en automobile selon leur chiffre d'affaires.

Rang	Assureur	Chiffre d'affaires 2018 (en M€)	Nombre de contrats en 2018	Part du segment auto dans le CA global en 2018
1	Covéa	3 866	9 863 001	26%
2	Axa	2 514	4 208 037	10%
3	Groupe Macif	1 921	5 184 961	61%
4	Groupama	1 844	4 060 859	Non communiqué
5	Allianz	1 602	2 845 310	Non communiqué
6	Groupe Maif	1 449	3 496 634	51%
7	Crédit agricole Assurances	1 164	2 983 457	Non communiqué
8	Generali	1 011	2 413 000	37%
9	Groupe des Assurances du Crédit mutuel	995	2 665 176	9%
10	Matmut	990	2 735 315	Non communiqué

Tableau 1 Top 10 des assureurs auto (chiffre d'affaires hors taxe)

1.1.2 La réglementation

1.1.2.1 Obligation d'assurance et fichier des véhicules assurés

Selon la loi du 27 février 1958, reprise par le Code des assurances et le Code de la route, la souscription d'une assurance responsabilité civile (RC) est obligatoire pour les véhicules terrestres à moteur. *« Toute personne physique ou toute personne morale autre que l'Etat, dont la responsabilité civile peut être engagée en raison de dommages subis par des tiers résultant d'atteintes aux personnes ou aux biens dans la réalisation desquels un véhicule est impliqué, doit, pour faire circuler celui-ci, être couverte par une assurance garantissant cette responsabilité, dans les conditions fixées par décret en Conseil d'Etat. Pour l'application du présent article, on entend par "véhicule" tout véhicule terrestre à moteur, c'est-à-dire tout véhicule automoteur destiné à circuler sur le sol et qui peut être actionné par une force mécanique sans être lié à une voie ferrée, ainsi que toute remorque, même non attelée. »* (Article L211-1 du Code des assurances).

Cette obligation, également nommée « garantie au tiers », a été instaurée dans le but de protéger les victimes afin de couvrir les frais liés aux dégâts matériels et corporels. De plus, elle profite également à l'assuré puisqu'elle protège son patrimoine personnel en cas d'indemnisation d'un tiers.

Cependant, malgré cette obligation d'assurance, le Fonds de Garantie estime qu'environ 750 000 véhicules circulent sans être assurés. Le Fonds de Garantie des Assurances Obligatoires de dommages (FGAO), créé en 1951 et financé par les assureurs et les assurés, indemnise les victimes d'accidents de la circulation provoqués par des personnes non assurées ou non identifiées. En 2017, le FGAO a réglé 85 millions d'euros aux victimes d'accidents impliquant des conducteurs non assurés, ce qui représente 27 970 dossiers.

Pour faire face à l'augmentation significative ces dernières années du nombre de véhicules non assurés, la création d'un fichier national recensant tous les véhicules assurés a été adoptée dans la loi de modernisation de la justice le 18 novembre 2016. Le fichier des véhicules assurés (FVA), actif depuis le 1er janvier 2019, doit être complété par les assureurs et leurs délégataires. Croisé avec le système d'immatriculation des véhicules (SIV), ce fichier permet de lister les véhicules sans assurance et de faciliter les contrôles des forces de l'ordre.

1.1.2.2 Indemnisation des sinistres

Les conventions d'indemnisation automobile ont été mises en place par les sociétés d'assurance afin de faciliter la gestion des sinistres d'un faible montant impliquant deux assureurs différents. La plupart des sociétés d'assurance adhèrent à ces conventions.

Sinistres matériels :

La Convention d'indemnisation directe de l'assuré et de recours entre sociétés d'assurance automobile (IRSA) date de 1968 et concerne uniquement les sinistres matériels.

Cette convention définit les conditions d'application du forfait IDA (Indemnisation directe de l'assuré). Elle concerne les sinistres matériels impliquant deux véhicules et dont le montant du dommage est inférieur à 6 500 €. Dans ce cas, l'assureur indemnise directement son assuré à hauteur du préjudice subi, puis, il émet un recours auprès de l'assureur adverse. Le montant remboursé par la compagnie d'assurance opposée est forfaitaire. En 2019, il s'élève à 1 482 € en cas de responsabilité totale, et à 741 € en cas de responsabilité partagée.

Lorsque le préjudice est supérieur à 6 500 €, la convention IRSA prévoit que la demande de recours ne soit plus forfaitaire, mais corresponde au montant réel du sinistre.

Sinistres corporels :

L'AIPP (atteinte à l'intégrité physique et psychique) permet de mesurer la gravité d'un sinistre corporel. Ce taux mesure la réduction des capacités physiques ou intellectuelles d'une victime d'un accident de la route et est déterminé par un médecin expert. Par conséquent, il s'agit d'un élément essentiel permettant de déterminer le montant d'indemnisation.

La loi du 5 juillet 1985 tendant à l'amélioration de la situation des victimes d'accidents de la circulation et à l'accélération des procédures d'indemnisation, dite « loi Badinter », a pour but de faciliter et d'accélérer l'indemnisation des victimes des accidents de la route. Elle a pour objectif de protéger les victimes, en leur garantissant une indemnisation intégrale des préjudices corporels et économiques.

Suite à la loi Badinter, la Convention d'indemnisation et de recours corporel automobile (IRCA) a été signée en 2002. Elle facilite la gestion des petits sinistres corporels survenus en France, impliquant au moins deux véhicules immatriculés en France et dont les séquelles permanentes des victimes sont faibles (AIPP inférieure ou égale à 5%). Dans ce cas, l'indemnisation des victimes est effectuée directement par l'assureur des assurés, puis, cet assureur exerce postérieurement un recours forfaitaire auprès de l'assureur adverse. Dans le cas où l'une des trois conditions précédentes n'est pas vérifiée, c'est à la compagnie de l'assuré ayant souscrit une responsabilité civile à indemniser les victimes.

1.1.2.3 Coefficient de réduction-majoration

L'assureur est libre de fixer le montant des primes qu'il souhaite appliquer. Cependant, selon l'Annexe à l'article A121-1 du Code des assurances, il est tenu de multiplier ce montant par un coefficient de réduction-majoration (CRM), communément dénommé « bonus-malus ».

Le CRM initial est fixé à 1. Un assuré bénéficie d'une réduction de 5% de son CRM pour chaque année sans accident comportant une part de responsabilité de celui-ci. La réduction maximale est fixée à 50%. Ainsi, 13 ans sont nécessaires pour atteindre un bonus de 0,5. Au-delà, il ne peut plus diminuer.

Lorsque l'assuré est déclaré responsable lors d'un sinistre, son CRM est majoré de 25% par accident et de 12,5% lorsqu'il n'est que partiellement responsable. Le CRM est plafonné à 3,5.

De plus, après deux années consécutives sans sinistre, le coefficient de réduction-majoration ne peut pas être supérieur à 1. On parle alors de « descente rapide ».

D'après l'article 4 de l'annexe à l'article A121-1 du Code des assurances, certains sinistres sont explicitement exclus d'une majoration du coefficient de réduction-majoration : un accident de stationnement sans tiers identifié et les garanties vol, incendie et bris de glace. Ce même article stipule que si un assuré possède un CRM de 0,5 depuis au moins 3 années, le premier accident impliquant une part de sa responsabilité n'entraîne pas de majoration de son CRM. Cet assuré devra ensuite attendre 3 années sans sinistre responsable avant de pouvoir bénéficier à nouveau de cet avantage.

Il est à noter que le CRM se conserve. Ainsi, lors d'un changement de véhicule ou d'assureur, le CRM précédent de l'assuré est appliqué.

1.1.2.4 Autres

D'autres lois ont été promulguées dans le but de protéger les assurés.

La loi Consommation, dite Loi Hamon, autorise, depuis 2015, les assurés à résilier leur assurance à n'importe quel moment après un an d'engagement, ce qui leur permet de s'affranchir des contraintes liées à la tacite reconduction des contrats.

La directive sur la distribution d'assurances (DDA), entrée en vigueur en 2018, contraint les compagnies d'assurance à fournir des conseils et des informations objectives lors de la distribution d'un contrat d'assurance.

Par ailleurs, le règlement général sur la protection des données (RGPD) impose, depuis 2018, des procédures concernant le traitement des données personnelles des assurés.

D'autres lois à propos de l'assurance auto en France sont actuellement en cours d'examen, telles que la loi d'orientation des mobilités (LOM) et la réforme de la responsabilité civile.

1.1.3 Les garanties proposées

1.1.3.1 La responsabilité civile

Comme évoqué dans le paragraphe précédent, il s'agit de la seule garantie obligatoire. Elle couvre les dommages causés à des tiers (passager, autre conducteur, piéton) ou à un autre véhicule.

1.1.3.2 Les garanties de dommages au véhicule

- Garantie vol : couvre le véhicule assuré en cas de vol ou de tentative de vol (serrure forcée, modification des branchements électriques, etc…).

- Garantie incendie : prend en charge les dégâts du véhicule lorsque celui-ci est endommagé ou détruit par un incendie, qu'il provienne de l'intérieur ou de l'extérieur du véhicule.

- Garantie bris de glace : couvre les dommages du pare-brise. Elle peut également être étendue aux vitres latérales, à la vitre arrière, ainsi qu'aux rétroviseurs et aux optiques des phares.

- Garantie dommage tous accidents : couvre tous les dommages subis par le véhicule, quelles que soient la responsabilité du conducteur et les circonstances de l'accident.

- Garanties catastrophes naturelles et catastrophes technologiques : s'appliquent lorsque l'une des garanties dommages citées précédemment a été souscrite et couvrent les dégâts matériels causés respectivement par une catastrophe naturelle (inondation, coulée de boue, séisme, raz de marée, cyclone, avalanche) et par une catastrophe technologique (accident non nucléaire survenant dans une installation classée, un stockage souterrain ou suite au transport de matières dangereuses). Elles ne peuvent être mises en jeu qu'après publication au Journal Officiel de la République Française d'un arrêté interministériel de catastrophe naturelle ou de catastrophe technologique.

- Garantie tempête : couvre les dommages matériels causés par une tempête et par les effets du vent.

- Garantie attentats et actes terroristes : est également automatiquement appliquée lorsqu'une garantie dommages est souscrite. Elle indemnise les dégâts liés à un attentat, un acte de terrorisme, un sabotage, un mouvement populaire ou une émeute.

1.1.3.3 La garantie personnelle du conducteur

Il s'agit de la seule garantie couvrant les dommages corporels subis par le conducteur, en prenant en charge les frais médicaux, le préjudice subi et la perte de revenus liée à un arrêt de travail.

1.1.3.4 Les garanties de service

- Défense pénale et recours après un accident de la circulation (DPRSA) : Suite à un accident, un assuré peut être engagé dans une procédure judiciaire. La DPRSA lui permet d'être représenté et défendu par l'assureur, qui prend à sa charge les honoraires d'avocat, ainsi que les frais d'expertise et de procédure. Il obtient ainsi un pouvoir de défense lorsqu'il est poursuivi du fait de la détention ou de l'utilisation d'un véhicule en cas d'accident, ou un pouvoir de recours à l'égard du tiers responsable pour la réclamation des dommages subis par le véhicule assuré et les personnes transportées.

- Garantie assistance dépannage : Cette garantie facultative entre en jeu en cas d'immobilisation du véhicule suite à une panne, une crevaison ou un accident. Elle comprend le remorquage du véhicule, le rapatriement ou l'hébergement du conducteur et des passagers et le remplacement du véhicule le temps des réparations.

1.1.4 Le marché français

En France, le réseau routier est composé de plus d'un million de kilomètres de voies nationales, départementales et communales. Selon les données de la FFA, plus de 42 millions de véhicules (hors flottes) de première catégorie (c'est-à-dire de voitures particulières et de camionnettes) sont assurés en France au 31 décembre 2018, ainsi que près de 3 millions de deux roues, de tricycles, de quadricycles à moteur, de voiturettes et de voitures sans permis. D'après ces mêmes données, l'âge moyen des véhicules de première catégorie du parc français est de 10,8 ans. 62% de ces véhicules sont alimentés par du diesel et 37% par de l'essence. La proportion de véhicules hybrides et électriques est de 1%, néanmoins, cette part est en augmentation depuis plusieurs années. Par ailleurs, 38% des véhicules ont une puissance fiscale de 6 à 8 cv, ce chiffre est en baisse par rapport aux années précédentes, au profit des véhicules de moins de 6 cv, qui représentent, en 2018, 47% du parc automobile. Cela peut être expliqué par l'aide financière appelée « bonus écologique » mise en place par le gouvernement dans le but de privilégier les véhicules moins polluants, et par conséquent ayant une plus petite puissance fiscale. De plus, 71% des conducteurs ont un CRM maximal, c'est-à-dire un bonus de 50%. Seulement 0,9% des conducteurs ont un coefficient correspondant à un malus.

Bien que seule la souscription de la responsabilité civile soit obligatoire, les garanties facultatives sont également très sollicitées.

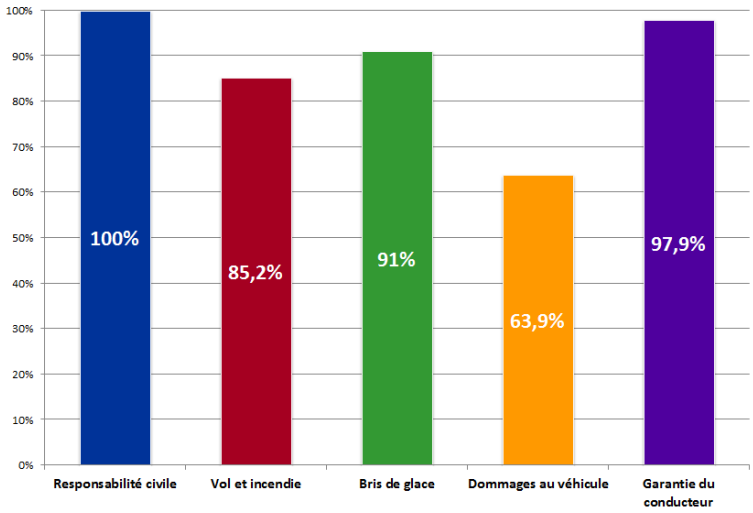

Figure 2 : Taux de souscription des différentes garanties en assurance automobile

En 2018, les primes perçues en assurance automobile s'élèvent à 22 milliards d'euros, soit une augmentation de 3,4% par rapport à 2017. En moyenne, un assuré paye sa prime 410 € hors taxes par an. Cependant, pour plus de significativité, il est nécessaire de comparer cette prime en fonction des garanties souscrites. Ainsi, un assuré ayant souscrit uniquement la garantie responsabilité civile paye sa prime en moyenne 147 €, tandis qu'un assuré ayant souscrit une assurance tous risques (hors assistance) paye sa prime annuelle en moyenne 496 € hors taxes.

La décomposition de la prime moyenne hors taxes selon les différentes garanties est détaillée ci-dessous.

Garantie	Responsabilité civile	Vol	Incendie	Bris de glace	Dommages au véhicule	Tempête, grêle, neige	Garantie du conducteur, protection juridique
Prime moyenne	147 €	41 €	7 €	46 €	195 €	8 €	52 €

Tableau 2 Décomposition de la prime moyenne hors taxe selon les garanties

Avec plus de 7,66 millions de sinistres en 2018, soit en moyenne près de 21 000 sinistres par jour, l'assurance automobile est l'assurance la plus sollicitée par les assurés. Les deux garanties les plus sollicitées sont le bris de glace et les dommages au véhicule.

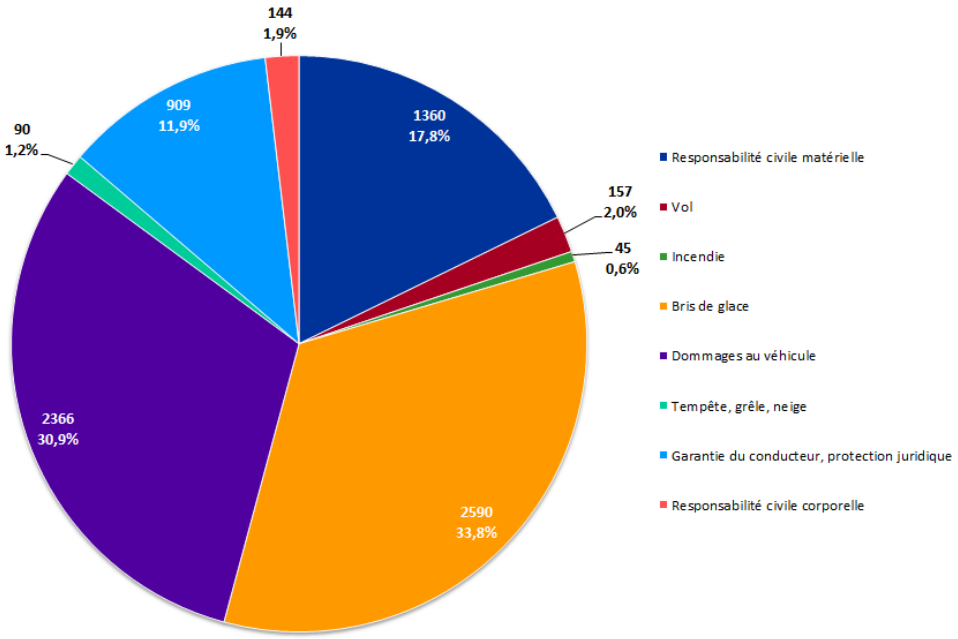

Figure 3 : Répartition du nombre de sinistres (en millier) des véhicules de première catégorie en 2018

Cependant, il est évident que la charge des sinistres n'est pas proportionnelle au nombre de sinistres. En effet, la responsabilité civile corporelle présente un coût moyen beaucoup plus élevé que d'autres garanties.

Ainsi, en 2018, 14,7 milliards d'euros ont été versés aux assurés à la suite de déclaration de sinistres automobiles. La répartition de cette charge est détaillée dans le diagramme ci-dessous.

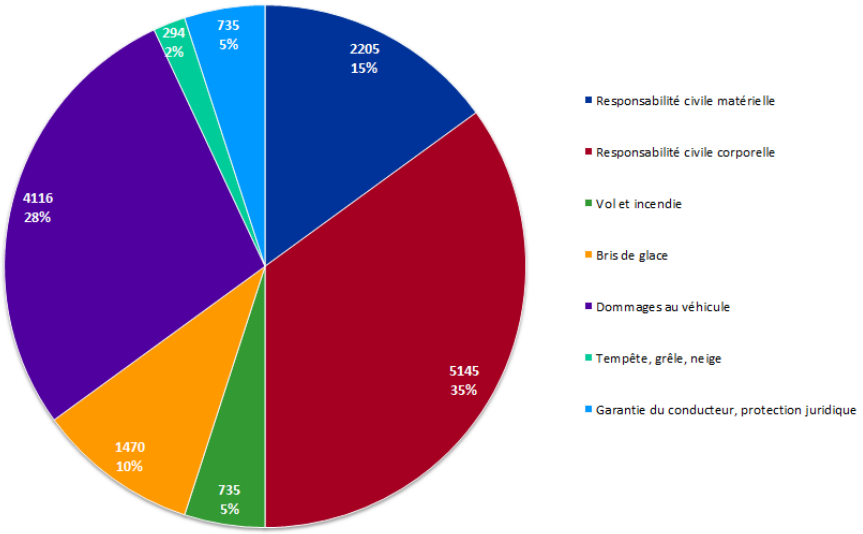

Figure 4 : Répartition de la charge (en million d'euros) des sinistres selon les garanties en 2018

Le coût moyen des sinistres matériels est en moyenne assez faible. En effet, les sinistres matériels liés à la responsabilité civile valent en moyenne 1 480 €, le vol 3 350 €, les dommages au véhicule 1 565 € et le bris de glace 475 €.

À l'inverse, les sinistres corporels représentent une grosse charge pour les assureurs. Le tableau suivant présente la part de victimes selon la gravité de l'accident et le montant de la charge des sinistres pour chaque catégorie.

	Répartition du nombre de sinistres corporels	Répartition de la charge des sinistres corporels
Blessés légers sans AIPP	68%	4%
Blessés avec AIPP	31%	91%
Dont AIPP de 1% à 5%	20%	8%
Dont AIPP de 6% à 19%	9%	22%
Dont AIPP de 20% et plus	2%	61%
Décès	1%	5%

Tableau 3 Répartition du nombre de victimes et du montant des indemnisations selon la gravité des sinistres corporels

Ainsi, on remarque que bien que les blessés ayant une AIPP supérieure à 20% ne correspondent qu'à 2% des dossiers, ils représentent à eux seuls 61% de la part des sinistres corporels. D'autre part, en 2018, 442 sinistres corporels graves ont coûté aux assureurs plus de 500 000 € chacun.

Du fait de la forte concurrence qui pousse les assureurs à réduire leurs tarifs, du coût moyen des sinistres qui est en augmentation et de la très grande volatilité des sinistres graves, l'assurance automobile n'est globalement pas rentable. En effet, le ratio combiné comptable net de réassurance s'établit à 100% en 2018. Cependant, il est à noter qu'après 12 années avec un ratio combiné comptable net de réassurance dépassant les 100%, avec un pic de 109% en 2009, ce dernier est revenu à l'équilibre en 2018.

1.1.5 Le risque aggravé

L'assurance automobile repose sur le principe de mutualisation des risques, c'est-à-dire que les primes versées par l'ensemble des assurés servent à régler des sinistres survenus seulement pour une petite partie de ces assurés. Par conséquent, les compagnies d'assurance cherchent à attirer les profils les moins risqués afin de réduire le montant des sinistres payés. Ainsi, il est possible pour les assureurs de refuser lors de la souscription certains individus qu'ils jugent trop exposés au risque ou, sous certaines conditions, de résilier un contrat d'assurance notamment lorsque l'assuré n'a pas payé sa prime, suite à une fausse déclaration volontaire de la part de l'assuré, pour cause d'aggravation du risque, après un accident en état d'ivresse, sous l'emprise de stupéfiant ou ayant entraîné une suspension ou une annulation de permis, ou bien à l'échéance du contrat.

Toutefois, comme évoqué précédemment, la souscription d'une assurance responsabilité civile est obligatoire. Par conséquent, la règlementation prévoit qu'un individu puisse être assuré contre les dommages que son véhicule pourrait causer à autrui, même s'il est considéré comme « à risque ». Ainsi, d'après l'Article L212-1 du code des Assurances, si une compagnie d'assurance a refusé la souscription d'une garantie Responsabilité civile à un individu, celui-ci peut solliciter le Bureau central de tarification (BCT) qui est chargé de fixer le montant de prime que l'assuré devra payer à la compagnie d'assurance, celle-ci devra alors proposer la garantie Responsabilité civile selon le prix et les franchises fixés par le BCT.

Pour anticiper le fait que certains conducteurs puissent avoir recours au BCT, il peut être intéressant de prévoir directement des polices d'assurance spécifiques aux personnes ayant eu des antécédents d'assurance « risqués ».

Ainsi, les produits d'assurance dits « risques aggravés » permettent aux conducteurs novices ainsi qu'aux conducteurs ayant un CRM élevé, des antécédents d'alcoolémie, une suspension ou annulation de permis, des interruptions d'assurance, une résiliation de contrat par un assureur précédent ou un nombre de sinistres dans les 36 derniers mois très important, de pouvoir souscrire une assurance automobile.

Les conducteurs en situation de risque aggravé ne représentent qu'une petite partie de la population française. Pour faire face à la concurrence entre sociétés d'assurance automobile, la Direction des Partenariat de Generali a décidé, par l'intermédiaire de courtiers, de proposer des produits « risque aggravé » afin de se démarquer et d'attirer des conducteurs généralement écartés par les assureurs traditionnels. Toutefois, cette cible d'assurés présente des antécédents risqués, d'où la nécessité de leur proposer une prime adaptée afin de garantir la rentabilité des produits proposés.

La seconde partie de ce chapitre expose donc les principes généraux permettant de modéliser la prime pure demandée aux assurés.

1.2 Modélisation de la prime pure

La spécificité de l'assurance repose sur le phénomène d'inversion du cycle de production. En effet, contrairement aux autres secteurs, les produits d'assurance sont vendus avant que le coût définitif ne soit connu. Ainsi, les primes perçues doivent être définies avant que le montant des indemnisations versées ne soit établi. C'est la raison pour laquelle il est nécessaire pour les assureurs de bien cerner le risque qu'ils portent afin de définir des tarifs adaptés. La constitution de provisions techniques et de marges de solvabilité sont les conséquences directes de l'inversion du cycle de production. Ces deux derniers aspects ne seront pas développés dans le cadre de ce mémoire.

1.2.1 Principes de tarification en assurance automobile

La prime commerciale, qui est la prime réellement payée par les assurés, se décompose en quatre éléments :

- La prime pure (π) : couvre le montant probable des sinistres, correspond à la tarification du risque.
- Les chargements de l'assureur : couvrent les coûts d'administration supportés par l'assureur.
- La commission du courtier : finance les coûts d'acquisition et de gestion supportés par le courtier.
- Le chargement de sécurité : couvre le risque de mauvaise tarification.

L'enjeu majeur de la tarification consiste à déterminer le montant probable des sinistres. Ce montant est réparti pour l'ensemble des assurés, on parle alors de mutualisation des risques. Ce principe consiste à répartir la charge totale des sinistres entre les différents assurés, de sorte que les risques se compensent. Il s'agit du modèle collectif.

Le principe de mutualisation des risques est applicable en assurance automobile à destination des particuliers et est une application directe de la Loi des Grands Nombres.

Théorème 1 Loi des Grands Nombres (démonstration en annexe 3.A)

Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires réelles indépendantes et identiquement distribuées et définies sur le même espace de probabilité. Soit $\mu = \mathbb{E}[X_i]$ et $\sigma = \mathbb{V}[X_i] < \infty$. Alors, la convergence suivante est vérifiée :

$$\frac{1}{n}\sum_{i=1}^{n} X_i \underset{n\to\infty}{\overset{\mathbb{P}}{\to}} \mu.$$

Le montant total payé par l'assureur est une variable aléatoire car il n'est pas connu à l'avance. Cependant, l'assureur doit fixer les primes perçues au moment de la souscription ou de la reconduction des contrats. La charge totale des sinistres est notée S et le montant des primes pures perçues est noté π. Puisque la prime pure correspond à la tarification du risque, il ne faut pas que la valeur de la prime pure soit surévaluée ou sous-évaluée. Ainsi, l'assureur doit chercher à minimiser l'erreur quadratique $\mathbb{E}[(S-\pi)^2]$. La solution de ce problème d'optimisation, dont la démonstration est faite en annexe 3.B, est donnée par $\pi = \mathbb{E}[S]$.

Ainsi, en assurant un grand nombre n de risques indépendants et équivalents, les compagnies d'assurance mutualisent les risques assurés (par application de la loi des grands nombres) et font payer aux assurés des primes dont le montant correspond à la tarification du risque.

Il est possible de modéliser directement la charge S. Cependant, une méthode très répandue en tarification de produits d'assurance IARD à destination des particuliers consiste à décomposer la variable aléatoire S en deux paramètres inconnus : le nombre total de sinistres et le prix unitaire de chaque sinistre.

Ainsi, S est définie de la manière suivante :

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

avec :

- S : charge totale des sinistres.
- N : nombre total de sinistres survenus
- C_i : montant du i^{ème} sinistre, valeur strictement positive.
- $(C_i)_{1\le i\le N}$ forment un échantillon de variables aléatoires indépendantes et identiquement distribuées de loi parente C.
- $(C_i)_{1\le i\le N}$ et N sont des variables aléatoires indépendantes.

Par conséquent, il est nécessaire d'estimer l'espérance de S donnant la sinistralité attendue afin de déterminer quel montant de prime pure l'assureur doit demander à ses assurés. La variance de S permet de déterminer le niveau de risque pris par l'assureur.

Ces deux éléments sont donnés par les deux formules suivantes, dont la démonstration est détaillée en annexe 3.C :

$$\mathbb{E}[S] = \mathbb{E}[N] \times \mathbb{E}[C],$$

$$\mathbb{V}[S] = \underbrace{\mathbb{V}[C] \times \mathbb{E}[N]}_{\textit{aléa sur le coût moyen des sinistres}} + \underbrace{\mathbb{V}[N] \times \mathbb{E}[C]^2}_{\textit{aléa sur le nombre de sinistres}}.$$

L'espérance de S se décompose donc en produit de deux espérances. C'est pourquoi, la connaissance de la loi de N et de la loi de C suffit à déterminer la prime pure. Cette approche est appelée modèle « coût × fréquence ».

Cependant, cette séparation de la charge en modèle « Coût moyen × fréquence » n'est possible que si le coût et la fréquence sont indépendants. Si tel n'est pas le cas, il est possible, grâce à d'autres méthodes, de modéliser directement la prime pure.

Les résultats précédents découlent de l'hypothèse que les $(C_i)_{1\leq i\leq N}$ sont indépendants et identiquement distribués, or, dans la pratique, il est évident que cette hypothèse d'homogénéité n'est pas vérifiée. En effet, les assurés ne présentent pas tous le même niveau de risque. C'est pourquoi, il est nécessaire de les regrouper par niveau de risque homogène. Cette segmentation permet donc aux assureurs de différentier les primes payées par les assurés selon plusieurs paramètres qui seront détaillés dans le deuxième chapitre.

Ce procédé est indispensable en assurance automobile car il permet de limiter le phénomène d'antisélection. En effet, si la prime payée était un montant forfaitaire identique pour tous les assurés, correspondant à la prime moyenne, les « mauvais risques » verseraient une prime sous-tarifée, tandis que les « bons risques » trouveraient la prime moyenne trop chère et iraient s'assurer auprès d'un autre assureur. Cela aurait pour conséquence de n'attirer que les assurés ayant une fréquence et un coût de sinistre supérieurs à ceux choisis dans l'hypothèse de tarification, et dégraderait ainsi les résultats techniques.

Les chargements couvrant les frais de l'assureur et le chargement de sécurité sont en général proportionnels à la prime pure. De plus, les commissions des courtiers sont fixées par les courtiers eux-mêmes et sont donc différentes pour chacun d'entre eux. Par conséquent, l'enjeu de la tarification consiste à déterminer la prime pure, c'est-à-dire à définir une loi de probabilité afin d'estimer l'espérance de la charge des sinistres.

L'enjeu de la tarification consiste donc à modéliser la fréquence de survenance d'un sinistre et le coût moyen de ce dernier. Ce mémoire se consacre à la comparaison de trois méthodes de modélisation. Cependant, avant de donner les détails techniques de conception de la prime pure, des notions communes relatives à la modélisation selon les différentes méthodes sont présentées.

1.2.2 Problématiques liées à la modélisation

La modélisation repose sur l'exploitation d'une base de données. Cette base contient l'historique des informations possédées par la compagnie d'assurance : les risques couverts (c'est-à-dire la liste de tous les contrats avec les caractéristiques des assurés et de leur véhicule) mais aussi la sinistralité passée. Le but de la modélisation est d'expliquer une variable (le coût ou la fréquence), notée Y en fonction de variables explicatives $X = (X_1; \dots; X_k)$. L'échantillon de données à disposition est noté $D = \left(x_{1,i}; \dots; x_{k,i}; y_i\right)_{i=1}^{n}$, il s'agit des réalisations des variables aléatoires X et Y pour n contrats.

Modéliser la variable aléatoire Y consiste donc à trouver une fonction de prédiction $f : \mathbb{R}^{k}$* $\rightarrow \mathcal{A}$ (où $\mathcal{A}$ désigne l'ensemble des prédictions possibles). Quel que soit le type de modèle utilisé (modèle linéaire généralisé, Machine Learning, etc…), l'objectif est de déterminer la meilleure fonction de prédiction.

La fonction de perte (*loss function* en anglais) permet de mesurer l'écart entre les valeurs observées et les valeurs prédites par le modèle. La fonction de perte la plus communément utilisée est la fonction d'erreur quadratique : $L\big(y; f(x_1; \dots; x_k)\big) = \big(y - f(x_1; \dots; x_k)\big)^2$. Celle-ci doit être minimisée afin de déterminer le meilleur modèle. L'estimateur optimal est défini par $f^* = \underset{f\in\mathcal{F}}{\operatorname{argmin}}\, \mathbb{E}\left[L\big(Y; f(X_1; \dots; X_k)\big)\right]$. Cette espérance

* La notation $\mathbb{R}^k$ est abusive. En effet, les variables peuvent aussi bien être quantitatives que qualitatives. Cependant, cette notation est utilisée dans le reste du mémoire. Elle signifie qu'il y a k variables explicatives.

définie le risque de prédiction : $R(f) = \mathbb{E}\big[L\big(Y; f(X_1; \dots; X_k)\big)\big]$. Cependant, en pratique, elle n'est pas calculable. Par conséquent, le risque empirique est utilisé : $\widehat{R_n(f)} = \frac{1}{n}\sum_{i=1}^{n} L\big(y_i; f(x_i)\big)$.

La principale problématique des modèles linéaires généralisés correspond aux choix effectués, notamment ceux de la loi et des variables sélectionnées. Si les hypothèses prises sont fausses, la modélisation risque d'être très mauvaise.

Les algorithmes CART et Forêt Aléatoire étudiés dans ce mémoire sont des modèles non paramétriques par conséquent, ils permettent de s'affranchir des contraintes liées au choix de la loi. Cependant, en cherchant à minimiser le risque empirique, un phénomène de sur-apprentissage (*overfitting* en anglais) peut survenir. Il est en effet nécessaire de minimiser l'erreur de prédiction, cependant, il ne faut pas que le modèle s'ajuste trop aux données d'apprentissage afin qu'il ne perde pas sa capacité de prédicteur sur de nouvelles données. Ce phénomène est illustré sur la figure ci-dessous.

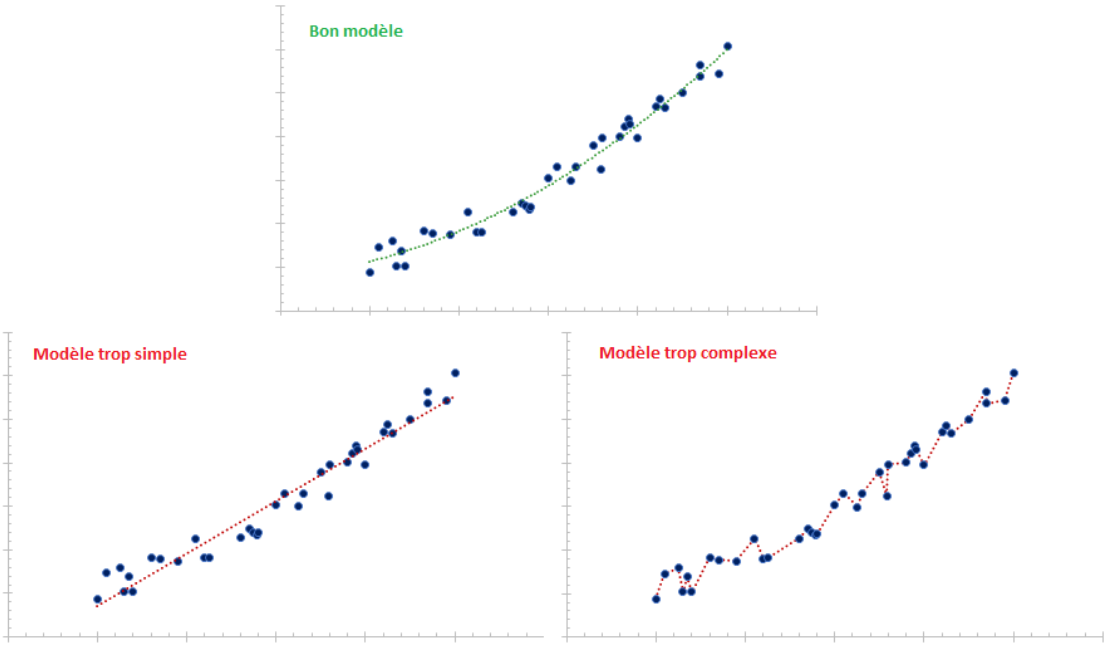

Figure 5 Exemples de modèles trop simple, correct et trop complexe

Tout l'enjeu de la modélisation consiste donc à concevoir un modèle qui soit à la fois un bon prédicteur mais qui ne soit pas pour autant trop complexe afin d'éviter le sur-apprentissage.

De manière générale, deux éléments peuvent détériorer les modèles : le biais et la variance. En considérant le risque quadratique défini précédemment, le risque de prédiction peut se décomposer de la manière suivante : $R(\hat{f}) = \mathbb{E}\left[\left(Y - \hat{f}(X)\right)^2\right] = \mathbb{E}\left[\left(Y - \mathbb{E}[Y|X] + \mathbb{E}[Y|X] - \mathbb{E}[\hat{f}(X)|X] + \mathbb{E}[\hat{f}(X)|X] - \hat{f}(X)\right)^2\right] = \underbrace{\mathbb{E}[(Y - \mathbb{E}[Y|X])^2]}_{variance\ du\ bruit} + \underbrace{\mathbb{E}\left[\left(\mathbb{E}[Y|X] - \mathbb{E}[\hat{f}(X)|X]\right)^2\right]}_{biais\ de\ \hat{f}} + \underbrace{\mathbb{E}\left[\left(\mathbb{E}[\hat{f}(X)|X] - \hat{f}(X)\right)^2\right]}_{variance\ de\ \hat{f}}$.

Le biais correspond à l'erreur de modélisation de l'algorithme (sous-apprentissage, mauvaises hypothèses, etc…) et la variance est due aux variations de l'échantillon d'apprentissage (sur-apprentissage, stabilité du modèle). Il est impossible de minimiser à la fois le biais et la variance. Par exemple, plus le nombre de variables explicatives dans le modèle est élevé, plus le biais est petit, cependant, cela aura pour conséquence d'augmenter la variance. Il est donc nécessaire de trouver un compromis entre le biais et la variance, c'est-à-dire entre la complexité du modèle et sa capacité à se généraliser.

C'est la raison pour laquelle les données utilisées par des méthodes de Machine Learning doivent être séparées : une partie sert à la modélisation et l'autre partie sert à la validation afin d'optimiser le modèle (cet aspect sera développé dans le quatrième chapitre).

Afin de tester si le modèle est un bon prédicteur pour les données, il est nécessaire de le tester sur un échantillon de données indépendantes, c'est-à-dire sur des données n'ayant pas servi à la conception du modèle. Ainsi, 20% de la base de données, choisie de manière aléatoire, sert de validation du caractère prédictif du modèle, aussi bien pour le modèle linéaire généralisé que pour les méthodes de Machine Learning.

L'objectif de ce mémoire est donc de proposer différentes méthodologies permettant de modéliser la prime pure d'un produit « risque aggravé » d'assurance automobile. Ces différentes méthodes sont ensuite comparées.

La tarification est un aspect essentiel pour le lancement ou la refonte d'un produit. En effet, bien que certaines caractéristiques telles que les franchises et les plafonds appliqués à chaque garantie, la qualité des services proposés ou encore la notoriété de la marque soient importantes, le prix du produit d'assurance reste largement le premier critère pour les assurés. Les assurés sont en général attirés par les produits les moins chers, d'autant plus qu'avec le développement des comparateurs d'assurance, il est très facile pour les assurés de comparer les prix proposés par des assureurs concurrents. Ainsi, les tarifs proposés doivent être compétitifs. Par ailleurs, la tarification technique n'est pas toujours celle réellement demandée aux assurés. En effet, selon les objectifs de la compagnie d'assurance (rentabilité, croissance, suivi de la clientèle existante), des modifications peuvent-être apportées, notamment en pratiquant des remises commerciales.

Chapitre 2 : Présentation et analyse des données

L'objectif de ce mémoire est de comparer plusieurs méthodes de modélisation de la prime pure afin de revoir la politique tarifaire des produits d'assurance automobile pour les « risques aggravés ». À titre indicatif et afin de mieux comprendre le fonctionnement de la Direction des Partenariats et des relations avec les courtiers, une présentation de Generali et de la Direction des Partenariats est disponible en annexe 1.

Au cours des années, la Direction des Partenariats a procédé, en fonction des ambitions de rentabilité et de développement commercial et en fonction des résultats techniques de chaque courtier, à des revalorisations tarifaires annuelles par rapport à la prime initialement calculée. Pour cette raison, la Direction des Partenariats souhaiterait mettre à jour, grâce à des méthodes de modélisation de la prime pure, les tarifs proposés. Les objectifs sont multiples : améliorer les tarifs actuellement proposés afin qu'ils correspondent de manière la plus juste possible au risque réel des conducteurs et de leur véhicule et avoir un tarif commun à l'ensemble des courtiers proposant des contrats « risque aggravé » afin d'élargir les conditions de souscription des polices déjà sur le marché et de proposer une prime pour le lancement de nouveaux partenariats avec des courtiers n'ayant pour le moment aucune donnée.

Avant de procéder à la modélisation de la prime pure, il est nécessaire de construire et d'étudier la base de données comportant les contrats et les sinistres des différents courtiers.

2.1 Construction de la base de données

2.1.1 Bases utilisées

La Direction des Partenariats dispose des bases contenant les informations relatives aux contrats et aux sinistres des courtiers partenaires sur les trois dernières années.

Parmi les courtiers travaillant avec l'Équité, 17 d'entre eux proposent des produits d'assurance automobile « aggravés ». Cependant, dans le cadre de l'étude menée pour ce mémoire, seules les données des cinq plus gros courtiers sont prises en compte. En effet, ceux-ci représentent 78% des contrats dit « aggravés » en portefeuille, et 71% des primes acquises. Pour des raisons de confidentialité, ils seront nommés A, B, C, D et E dans la suite de ce mémoire.

Par ailleurs, deux autres bases communes à tous les produits d'assurance automobile ont été exploitées : le zonier et le véhiculier.

La zone géographique est un critère discriminant car le risque n'est pas le même dans toutes les villes. Cependant, il n'est pas possible de concevoir un tarif spécifique à chacune des 36 681 communes françaises. Par conséquent, le zonier permet de réduire le nombre de zones géographiques tout en discriminant de façon juste les profils de risque. Le zonier utilisé est celui de la Direction des Partenariats de Generali. Il est commun à tous les partenaires et associe à chaque ville, selon son code postal, une zone numérotée de 2 (la moins risquée) à 6 (la plus risquée). Il a été conçu en fonction de la sinistralité passée de la ville (nombre de sinistres, coût moyen des sinistres, etc…), des données démographiques de celle-ci (nombre d'habitants, âge moyen des habitants, etc…) et de ses données routières (nombre de feux

tricolores, de ronds-points, densité de la circulation routière, etc…). Les coefficients de ce zonier seront supposés en conformité avec les données étudiées pour les cinq courtiers et ne seront pas améliorés dans le cadre de ce mémoire.

L'association SRA (Sécurité et réparation automobiles) est un organisme professionnel dont toutes les entreprises d'assurance automobile sont adhérentes. Elle met à disposition de celles-ci des bases de données regroupant toutes les caractéristiques techniques et commerciales des véhicules. Elle a également pour objectif de transmettre des classements selon la sécurité des véhicules et la protection contre le vol et des études sur les coûts de réparation. Le véhiculier utilisé dans le cadre de ce mémoire est celui transmis par le SRA. Il contient notamment des informations sur la marque, le modèle et la version du véhicule, le type de carrosserie, la puissance fiscale, la boîte de vitesses, la cylindrée du moteur, la vitesse maximale du véhicule, son prix d'achat ainsi que la date de début et de fin de mise en circulation. De plus, un code unique est associé à chaque modèle de véhicule afin de permettre son identification. Deux autres variables sont fixées par le SRA :

- Le groupe : variant de 20 à 50, il représente la dangerosité intrinsèque du véhicule et est définit en fonction de la puissance de celui-ci et des dispositifs de sécurité en cas d'accident.
- La classe : avec des valeurs comprises entre A et ZE (+ HC = hors classe), elle révèle la valeur du véhicule et est fixée en fonction du prix à neuf de ce dernier (hors option et hors remise commerciale) et du coût de réparation des pièces détachées.

2.1.2 Harmonisation et fusion des bases

Les données étant issues des bases de cinq courtiers différents, il est nécessaire que celles-ci soient harmonisées. En effet, tous les courtiers ne codifient pas leurs données de façon uniforme.

La première étape de construction de la base commune des contrats consiste à sélectionner les variables qui pourraient potentiellement être utilisées lors de la modélisation.

Trois types de variables se distinguent dans la base contrats :

- Les variables relatives au contrat (Numéro de contrat, nombre d'avenants, fractionnement de la prime, garanties souscrites, etc…)
- Celles concernant l'assuré (Nom, prénom, adresse, profession, date de naissance, date d'obtention du permis de conduire, antécédents d'assurance et de sinistralité, etc…)
- Celles correspondant au véhicule assuré (Marque, modèle, immatriculation, type de carburant, usage, date de mise en circulation, etc…)

Dans un premier temps, le but est de conserver un maximum de variables qui pourraient être significatives lors de l'étape de modélisation.

Cependant, certaines variables spécifiques au courtier ou au contrat ne pourront pas être exploitées par la suite. C'est le cas notamment du code de l'apporteur pour les bases des courtiers grossistes ou du nombre d'avenants depuis la date de souscription du contrat. Par ailleurs, il existe des données uniques qui ne pourront pas servir à la modélisation. Ainsi, le nom et le prénom d'un assuré ne sont nécessairement pas des variables révélatrices de son niveau de risque. Il en est de même pour le numéro d'immatriculation du véhicule. En effet, le groupe et la classe SRA étant directement donnés dans les bases, il n'est pas nécessaire de le conserver afin de croiser les informations du véhicule avec le SIV. De plus, quelques variables, telles que le type de garage, le carburant utilisé pour alimenter le véhicule, le mode d'acquisition du véhicule ou encore le nombre de mois d'antécédents d'assurance étant indisponibles pour trois, voire quatre des courtiers, il est impossible de les conserver dans le modèle. D'autre part, il aurait été intéressant de conserver la variable « sexe ». Cependant, d'après la loi n°2007-1774 du 12 décembre 2007, reprise par l'article L111-7 du Code des assurances, « *Toute discrimination directe ou indirecte fondée sur la prise en compte du sexe comme facteur dans le calcul des primes et des prestations ayant pour effet des différences en matière de primes et de prestations est interdite.* ». Il parait donc injustifié de retenir cette variable.

L'étude portant sur un portefeuille de contrats « aggravés », seul un conducteur est déclaré dans la base de données.

Les bases contrats et sinistres sont ensuite fusionnées afin de relier chaque sinistre au conducteur associé et aux caractéristiques de son véhicule.

2.1.3 Description des variables

Le tableau ci-dessous présente les variables explicatives retenues ainsi que les différentes modalités qu'elles peuvent prendre.

Variables	Description	Modalités
Courtier	Courtier apporteur	{A ; B ; C ; D ; E}
Fractionnement	Fractionnement de la prime d'assurance	{Mensuel ; Trimestriel ; Semestriel ; Annuel ; Autre}
Formule	Garanties souscrites	{RC ; RC + BDG ; RC + VI ; RC + BDG + VI ; RC + BDG + VI + DTA}
Classe SRA	Classe définie par le SRA (valeur du véhicule)	{A ; B ; … ; ZE ; HC}
Groupe SRA	Groupe défini par le SRA (dangerosité du véhicule)	{20 ; … ; 50}
Âge du véhicule	Âge en années, calculé par différence entre la date de début du contrat et la date de mise en circulation du véhicule	$\mathbb{N}$
Zone	Coefficient défini par le zonier en fonction du code postal	{2 ; 3 ; 4 ; 5 ; 6}
Usage	Trajets réalisés	{Vie privée ; Vie privée + trajets travail ; Vie privée + affaires}
Âge du conducteur	Âge en années, calculé par différence entre la date de début du contrat et la date de naissance de l'assuré	$\mathbb{N} \setminus \{0; \dots ; 17\}$
Nombre d'années de permis	Calculé par différence entre la date de début du contrat et la date d'obtention du permis de conduire	$\mathbb{N}$
Catégorie socioprofessionnelle	Classe de métier exercé	{Agriculteur ; Artisan/Commerçant ; Etudiant ; Profession libérale ; Retraité ; Salarie/Fonctionnaire ; Sans profession}

CRM	Coefficient de réduction-majoration (« bonus-malus »)	[0,5 ; 3,5]
Résiliation non-paiement	Résiliation par un assureur précédent pour non-paiement de primes	{OUI ; NON}
Résiliation fausse déclaration	Résiliation par un assureur précédent suite à une fausse déclaration	{OUI ; NON}
Résiliation nombre de sinistres	Résiliation par un assureur précédent pour cause de fréquence de sinistres trop élevée	{OUI ; NON}
Annulation de permis	Invalidation pour défaut de point ou annulation judiciaire au tribunal	{OUI ; NON}
Suspension de permis	Suspension provisoire, de 6 mois à 5 ans	{OUI ; NON}
Condamnation pour alcoolémie	Condamnation suite à taux d'alcoolémie supérieur à 0,5g/L	{OUI ; NON}
Sinistres corporels	Nombre de sinistres corporels au cours des 36 derniers mois	$\mathbb{N}$
Sinistres matériels responsables	Nombre de sinistres matériels responsables au cours des 36 derniers mois	$\mathbb{N}$
Sinistres matériels non responsables	Nombre de sinistres matériels non responsables au cours des 36 derniers mois	$\mathbb{N}$

Tableau 4 Récapitulatif des variables explicatives utilisées

2.1.4 Construction des bases d'étude

Après avoir fusionné les différentes bases, il est nécessaire de retraiter la base brute obtenue afin d'exploiter les informations concernant la fréquence et le coût des sinistres selon les conducteurs et leur véhicule.

Une base différente est élaborée pour chaque garantie. En effet, les différents types de sinistres ont des fréquences différentes, il n'est donc pas possible d'étudier simultanément plusieurs garanties.

Bases pour la fréquence :

La jointure de la base contrats et de la base sinistres peut comporter plusieurs lignes correspondant au même contrat pour la même période en portefeuille. En effet, si un assuré a eu plusieurs sinistres durant la durée de son contrat d'assurance, plusieurs lignes distinctes (correspondant au nombre de sinistres) sont inscrites dans la base recensant les sinistres. Afin d'estimer la fréquence de sinistralité de chaque garantie, il est nécessaire de regrouper ces lignes, sinon, tous les assurés seraient considérés comme ayant eu au maximum un seul accident, ce qui n'est pas nécessairement le cas.

De plus, pour calculer la fréquence, il est important de prendre en compte l'exposition. Elle correspond à la durée durant laquelle l'assuré est resté dans le portefeuille. Elle se calcule grâce à la formule suivante :

$$exposition = \frac{date\ de\ fin\ du\ contrat - date\ de\ début\ du\ contrat}{365}.$$

En effet, les contrats n'ont pas tous la même durée. Par conséquent, il est nécessaire de prendre en compte l'exposition car un assuré qui a eu un sinistre alors qu'il est resté dans le portefeuille seulement six mois est plus risqué qu'un conducteur ayant eu un sinistre durant une année complète. La fréquence de sinistralité est calculée pour chaque type de sinistre grâce à la formule :

$$fréquence = \frac{nombre\ de\ sinistres\ durant\ la\ période\ d'assurance}{exposition}.$$

Par ailleurs, il convient de garder dans la base d'un type de sinistre uniquement les assurés ayant souscrit une garantie correspondant à ce sinistre. En effet, tous les conducteurs ayant uniquement une assurance responsabilité civile ne déclarent aucun autre type de sinistres. Cependant, il serait erroné de considérer que leur fréquence de sinistralité est nulle pour les autres garanties.

Il convient de remarquer que tous les sinistres de la base de données ont été rattachés à un assuré grâce au numéro de contrat, par conséquent, aucun sinistre orphelin ne nécessite de modélisation spécifique.

Bases pour le coût moyen :

Les bases utilisées pour la modélisation du coût moyen des sinistres correspondent simplement à une extraction de la base brute issue de la fusion entre les bases contrats et sinistres.

2.2 Analyse de la base de données

La mutualisation des bases de données des cinq partenaires permet de regrouper plus d'informations. Toutefois, cela impose que les bases présentent des aspects communs. En effet, si les caractéristiques des contrats, des conducteurs, des véhicules et de la sinistralité sont totalement différentes, il ne serait pas pertinent de regrouper les données à cause de leur hétérogénéité. Cette éventualité est d'autant plus à considérer que les courtiers ne touchent pas la même cible de population du fait de leurs actuelles conditions de souscription.

L'hétérogénéité des profils cibles des courtiers apporte une meilleure couverture des profils de risque existants. En effet, élaborer un tarif à l'aide de cinq bases permet de partager les informations contenues dans celles-ci. Lorsqu'un futur client ne correspondant pas au profil des assurés souscrivant habituellement chez un certain courtier souhaite effectuer un devis auprès de ce courtier, si les données n'avaient pas été mutualisées, le risque associé à cet individu pourrait être mal déterminé. En revanche, le fait d'avoir partagé les bases de plusieurs courtiers permet de mieux calibrer le risque et ainsi, de proposer un tarif plus adapté. Cela permet aussi d'avoir un socle de tarification pour le lancement de nouveaux produits de risque aggravé par de nouveaux courtiers.

Une étape préliminaire à la modélisation de la prime pure consiste donc à analyser la base de données, afin d'en ressortir les principales tendances de chaque courtier, mais aussi d'avoir une première appréciation des profils les plus risqués.

2.2.1 Analyses descriptives et retraitement éventuel des données

Cette partie du mémoire présente l'allure des principales variables, ainsi que les tendances générales de sinistralité.

2.2.1.1 Statistiques du portefeuille

Avant de pouvoir analyser la sinistralité, il est nécessaire d'étudier les principales caractéristiques des contrats (répartition par courtier, attributs des souscripteurs et des véhicules assurés, etc…)

Certains retraitements ont été appliqués. Ils ont été mis en place lorsque des données étaient manquantes, ou présentaient des aberrances. En l'absence de correction explicitement écrite dans ce mémoire, aucune modification n'a été effectuée.

Répartition des contrats par courtier :

Les données étudiées sont issues des bases des cinq plus gros courtiers travaillant avec la Direction des Partenariats. Cependant, ces derniers n'ont pas tous la même volumétrie de contrats. Le diagramme ci-dessous représente la répartition de ces contrats « aggravés » selon les cinq courtiers.

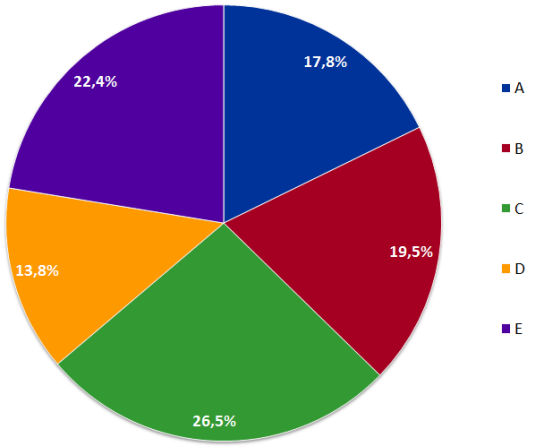

Figure 6 Répartition du nombre de contrats en fonction des courtiers

Il apparait ainsi que le courtier C a presque deux fois plus de contrats que le courtier D. Ainsi, pour permettre une meilleure analyse des résultats et une meilleure comparaison, les descriptions suivantes seront effectuées en pourcentage par rapport au nombre de contrats possédés par un courtier.

La base étudiée étant issue de contrats « aggravés », le comportement des assurés est différent de celui de l'ensemble de la population française décrit dans le premier chapitre.

Ainsi, 49% des assurés du portefeuille choisissent de ne souscrire aucune garantie sur le bris de glace, le vol, l'incendie ou les dommages au véhicule. Sur l'ensemble des assurés considérés, seulement 44% d'entre eux choisissent de souscrire une assurance en cas de vol ou d'incendie, 49% sur le bris de glace, 58% sur la garantie personnelle du conducteur et seulement 23% souscrivent une assurance sur les dommages au véhicule. Pour chacune de ces garanties, ces pourcentages représentent 40 points de moins que la moyenne nationale. Le diagramme ci-dessous exprime les spécificités de chaque courtier. Entre autres, le courtier B ne propose aucune garantie personnelle du conducteur (GPC), tandis que les assurés du distributeur E souscrivent presque tous une GPC et que très peu de ceux du partenaire C choisissent cette garantie. Cela est dû au fait que les courtiers s'adressent généralement à un type d'assurés spécifique, n'ayant pas les mêmes caractéristiques. Par ailleurs, la politique de souscription et de communication publicitaire n'est pas la même selon les distributeurs.

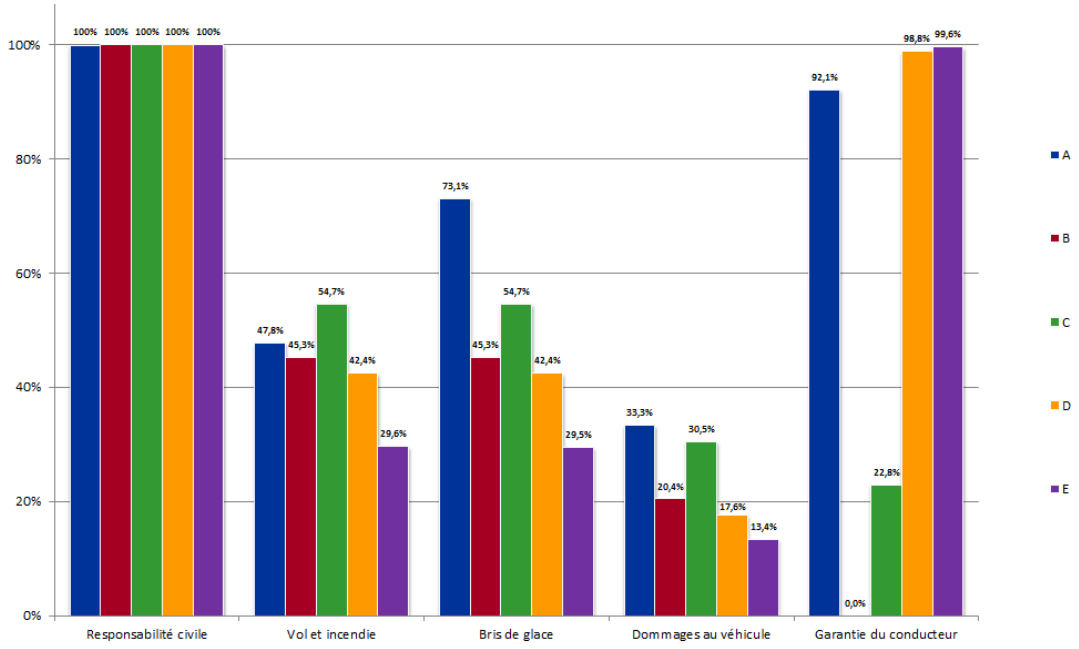

Figure 7 Taux de souscription des garanties par courtier apporteur

Comme évoqué précédemment, les conditions de souscription des polices d'assurance associées aux cinq courtiers sont différentes. Cela a un impact sur les caractéristiques des souscripteurs. De ce fait, certains courtiers visent à attirer les jeunes conducteurs et les conducteurs ayant très peu d'expérience de conduite tandis que d'autres ont dans leur portefeuille des conducteurs en situation de malus et d'autres encore ont des conducteurs ayant fait l'objet d'une résiliation par un assureur précédent.

Les différents graphiques présentés par la suite permettent d'avoir un aperçu de la répartition des assurés selon les courtiers.

Description des conducteurs :

Dans tous les contrats d'assurance automobile, l'âge des conducteurs et leur expérience de conduite ont un impact sur le tarif demandé. Le choix de l'âge minimum et de l'âge maximum des assurés fait partie des conditions de souscription et diffère selon les courtiers apporteurs. Aussi, comme en témoigne le graphique suivant, les produits d'assurance « risque aggravé » ne sont pas proposés à la même génération selon les partenaires. Alors que le courtier B adresse ses contrats à une population plutôt jeune, les courtiers A et E n'ont aucun client âgé de moins de 25 ans.

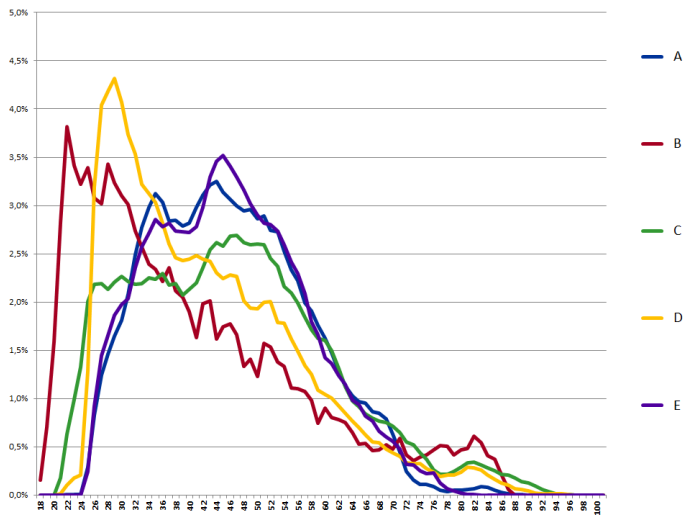

Figure 8 Distribution de l'âge des conducteurs par courtier apporteur

Il est à noter que les données ne présentent pas d'incohérence car tous les conducteurs déclarés sont majeurs. Cependant, du fait de la différence remarquable des profils de conducteurs, les valeurs manquantes ont été retraitées selon le courtier apporteur. Ainsi, les données manquantes, qui représentaient 0,12% de l'échantillon total, ont été remplacées par l'âge médian correspondant au courtier apporteur, c'est-à-dire 36 ans pour B, 40 ans pour D, et 45 ans pour A, C et E.

Du fait des spécificités liées à l'âge du conducteur déclaré, il est naturel de retrouver les mêmes différences à propos de l'ancienneté de permis du conducteur. Ainsi, un tiers des assurés chez les courtiers B et D ont obtenus leur permis moins de 5 ans avant le début de leur contrat d'assurance, tandis que ce type de conducteur ne représente que 1% du portefeuille des courtiers A et E.

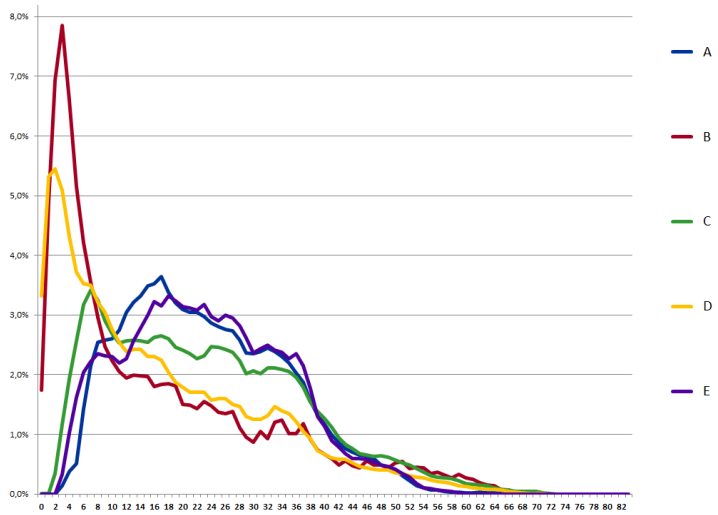

Figure 9 Distribution de l'ancienneté de permis des conducteurs par courtier apporteur

Pour les mêmes raisons que lors de la transformation des valeurs manquantes relatives à l'âge du conducteur, les valeurs manquantes de cette variable ont été remplacées par le nombre d'années de permis médian, selon le courtier rattaché, c'est-à-dire 11 ans pour le courtier B, 13 pour D, 22 pour A et C, et 23 ans pour le courtier E.

Afin de mieux comprendre les profils attirés par chaque courtier, il est nécessaire d'étudier la catégorie socio-professionnelle des assurés. En effet, cela donne des renseignements sur l'âge du conducteur (lorsque l'assuré est retraité ou étudiant), mais aussi sur l'usage de son véhicule (notamment pour les trajets professionnels). Une grande majorité des assurés en portefeuille sont salariés ou fonctionnaires : 61% pour E, 74% pour C, 75% pour A, et 79% pour les courtiers B et D. La différence assez notable pour le courtier E vient du fait qu'une grosse proportion de ses assurés sont sans profession : 19% alors que les autres courtiers ont en moyenne 4% d'assurés sans profession. Pour les autres modalités ; la répartition des assurés selon leur catégorie socio-professionnelle est globalement la même selon les courtiers : moins d'1% sont agriculteurs, 6% sont artisans ou commerçants, 3% exercent une profession libérale, 3% des assurés sont étudiants et 7% sont retraités.

Un autre critère important pour la tarification automobile est le coefficient de réduction majoration. Il est à la fois dépendant du nombre d'années depuis lesquelles le conducteur est titulaire du permis, mais aussi du nombre de sinistres responsables causés par ce conducteur. Puisque les données étudiées sont issues d'une base de contrats « aggravés », il est évident que les conducteurs du portefeuille étudié n'ont pas les mêmes caractéristiques que l'ensemble des conducteurs français. En effet, 71% des Français assurés ont un CRM de 0,5. C'est le cas de seulement 29% des assurés étudiés. À l'inverse, 13% des conducteurs étudiés ont un

coefficient correspondant à un malus, ce qui est nettement supérieur au 0,9% de la moyenne nationale. De plus, 41 conducteurs ont un malus maximal, c'est-à-dire un CRM valant 3,5.

La répartition par courtier est détaillée dans le diagramme ci-dessous. Il en ressort que très peu d'assurés du courtier D ont un bonus maximal. De plus, les courtiers B et D ont un taux d'assurés en situation de malus encore plus élevé que les autres courtiers.

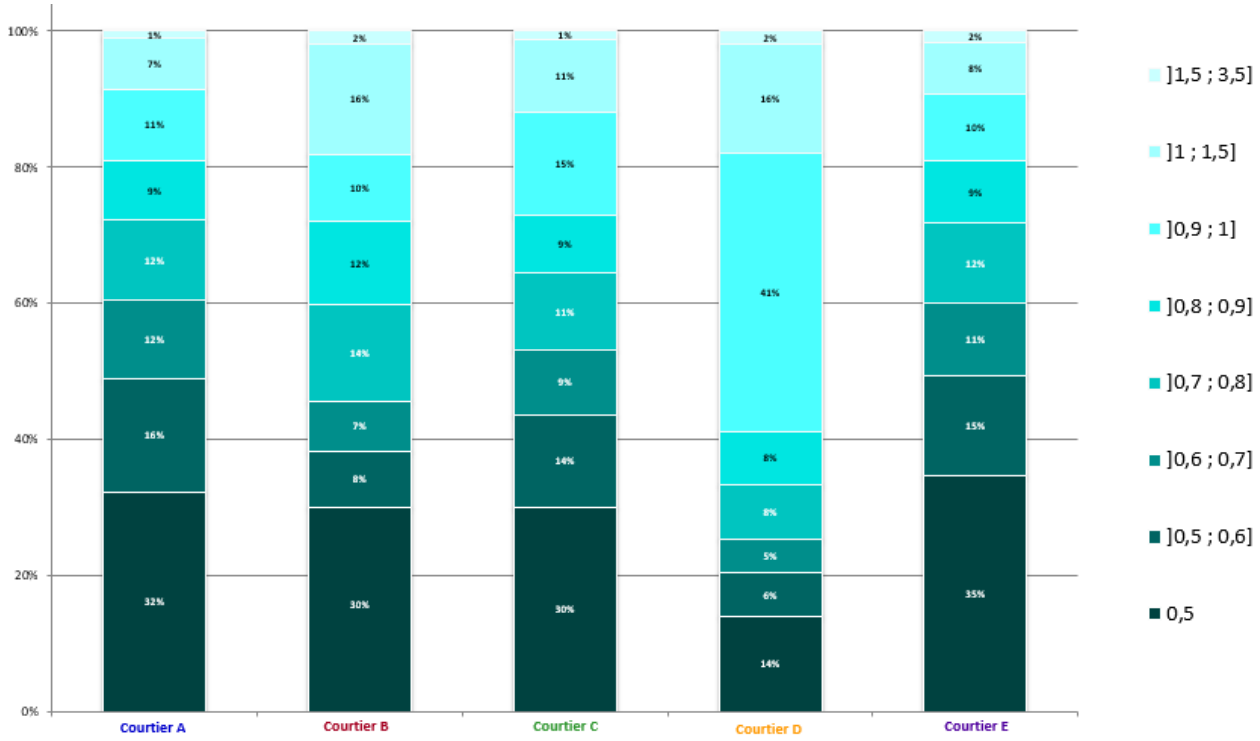

Figure 10 Distribution des coefficients de réduction-majoration selon les courtiers

Par ailleurs, aucune donnée aberrante n'a été constatée (CRM < 0,50, CRM > 3,5, CRM incompatible avec l'ancienneté de permis du conducteur, etc…)

Lorsqu'un conducteur veut souscrire une assurance auto, il doit informer l'assureur s'il a eu une résiliation par un autre assureur au cours des trois dernières années. Dans le cas de contrats d'assurance aggravés, de nombreux conducteurs ont eu une résiliation En effet, sur l'ensemble des contrats du portefeuille, seulement 47% d'entre eux n'ont fait l'objet d'aucune résiliation d'assurance de la part d'un assureur précédent ou d'aucune suspension ou annulation de permis. Ce pourcentage baisse drastiquement à 18% pour les clients du courtier E. D'après les remarques exposées dans les paragraphes précédents, le courtier D est spécialisé envers les conducteurs novices ou ayant un CRM élevé. De ce fait, seulement 5% de ses clients ont fait l'objet d'une résiliation dans les trois dernières années. Les spécificités de chaque courtier sont mises en avant dans le diagramme ci-dessous. Ainsi, il est à noter que les conditions de souscription du courtier A sont favorables aux conducteurs ayant eu une suspension de permis. Tandis que les individus ayant fait l'objet d'une résiliation pour non-paiement de prime souscrivent auprès des distributeurs C et E, ceux ayant eu un nombre très important de sinistres engendrant une résiliation sont apportés par les courtiers B et C.

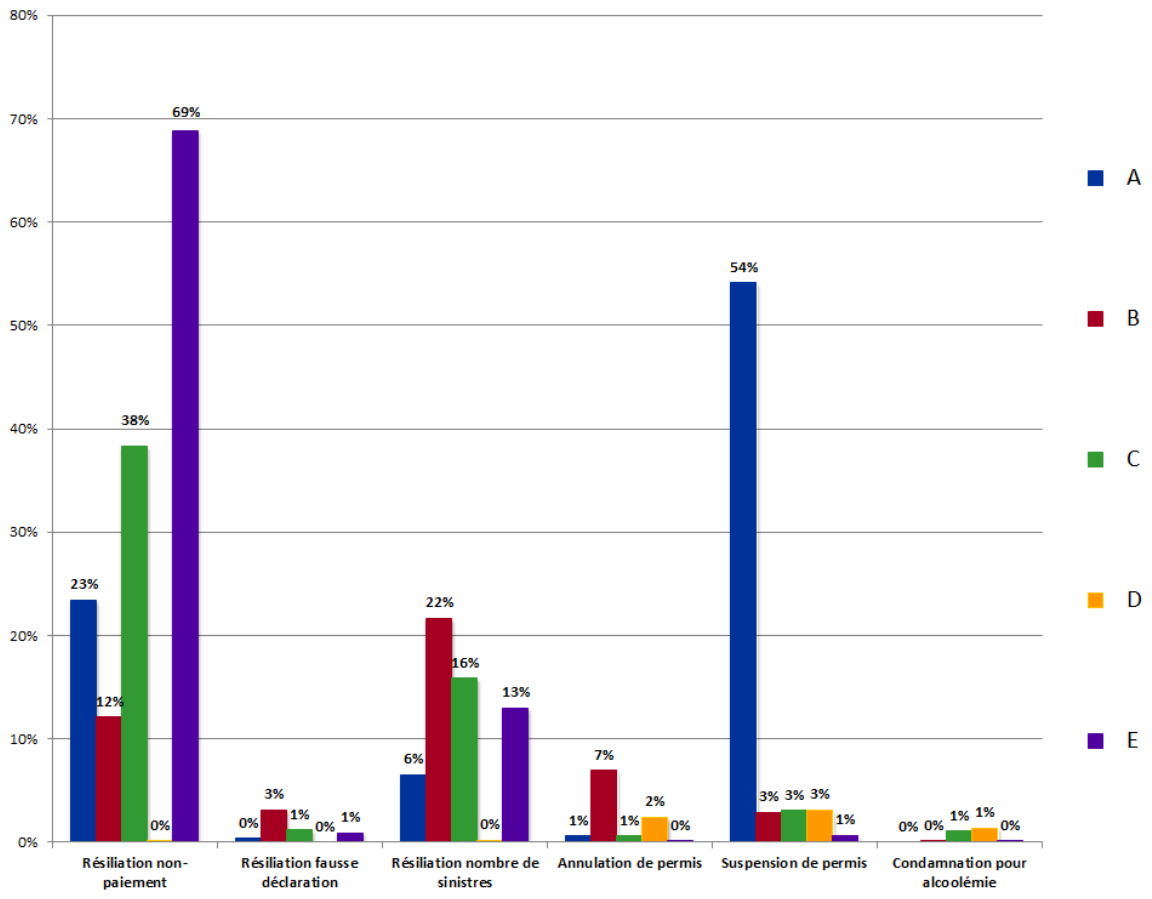

Figure 11 Distribution des causes de résiliation selon les courtiers apporteurs

Le dernier élément spécifique aux risques aggravés concerne les antécédents de sinistralité. Lors de la souscription d'un nouveau contrat d'assurance, l'assureur, par l'intermédiaire des courtiers délégataires, demande au potentiel assuré de déclarer les sinistres qu'il a eu au cours des trois dernières années. Les trois diagrammes suivants représentent la répartition du nombre d'antécédents de sinistres corporels, matériels responsables et matériels non responsables.

La plupart des contrats d'assurance automobile « classiques » excluent de leurs conditions de souscription les personnes ayant eu au moins un sinistre corporel. En effet, elles représentent un fort risque pour les assureurs, d'autant plus que les sinistres corporels sont les plus coûteux. Il est à noter que les courtiers B et C sont apporteurs de beaucoup d'assurés ayant eu un ou plusieurs sinistres corporels au cours des trois dernières années.

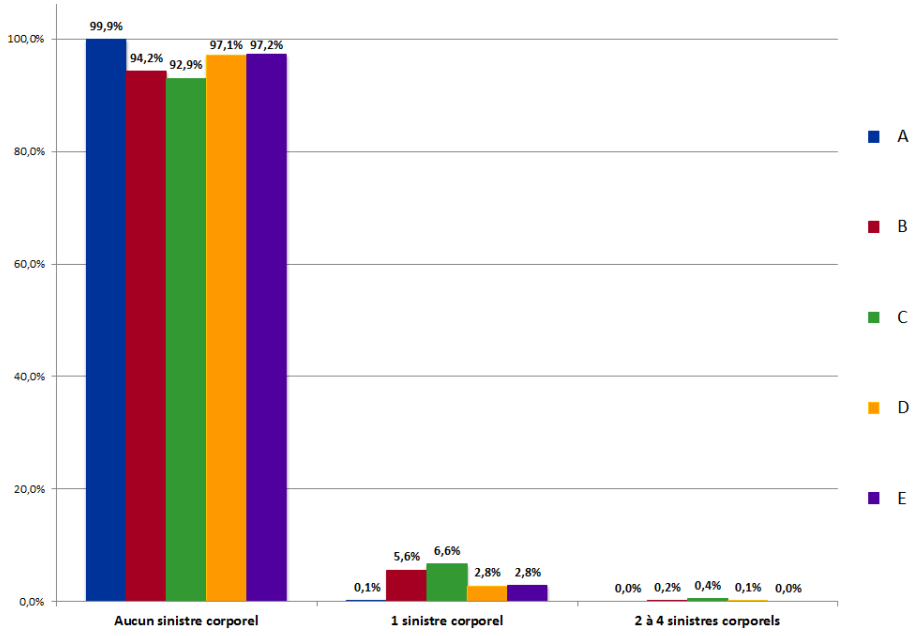

Figure 12 Répartition des antécédents de sinistralité corporelle selon les courtiers

Bien que, par définition, les sinistres non responsables ne sont pas causés par l'assuré, ils peuvent être onéreux pour l'assureur. Ainsi, en cas de vol du véhicule, l'assureur doit rembourser la valeur de ce dernier, il en est de même suite à une déclaration de bris de glace. Par ailleurs, en cas de sinistre matériel non responsable avec un tiers, du fait des conventions IRSA, si le coût du sinistre est supérieur au montant du forfait IDA, l'assureur perd de l'argent.

34,5% des assurés du courtier C et 33,4% des assurés du courtier E ont eu au moins un sinistre matériel non responsable au cours des trois dernières années.

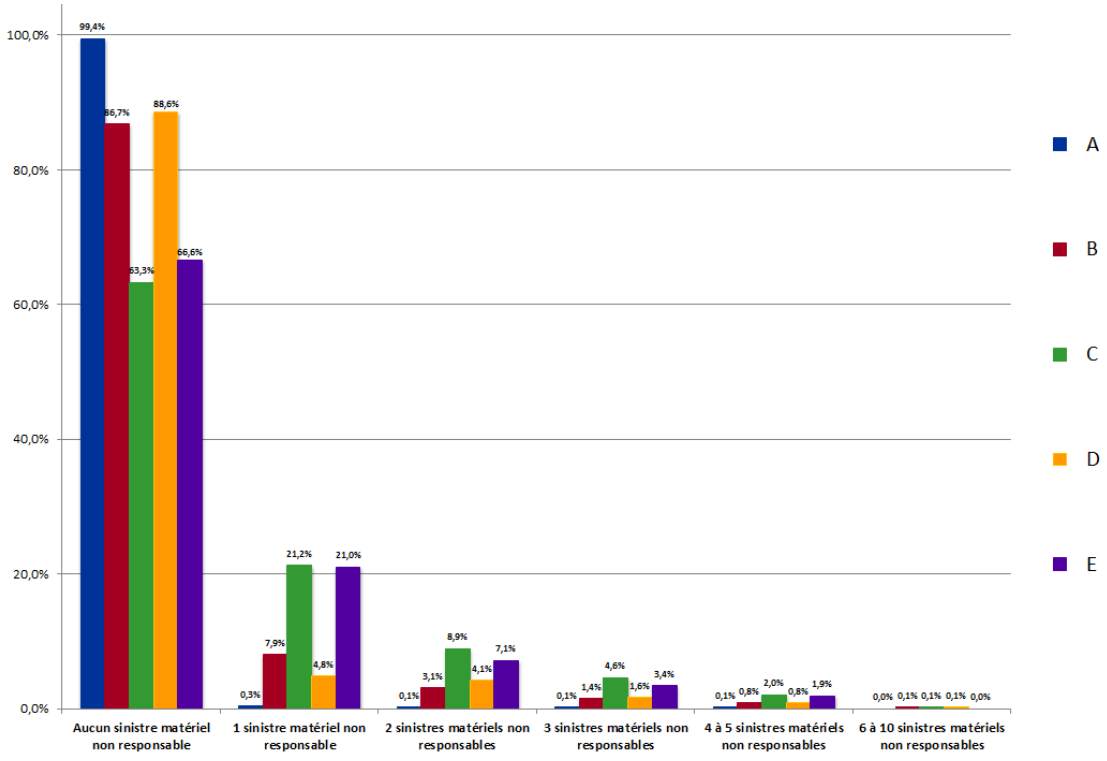

Figure 13 Répartition des antécédents de sinistralité matérielle non responsable selon les courtiers

Pour les contrats d'assurance automobile « standard », les assureurs évitent de prendre des assurés ayant eu beaucoup de sinistres responsables parce qu'ils représentent un risque pour l'assureur. En effet, ces sinistres sont directement causés par l'assuré et en cas de sinistre impliquant des tiers, l'assureur doit également les indemniser.

Les mêmes tendances que pour les sinistres matériels non responsables se dégagent, c'est-à-dire que le courtier A n'apporte que très peu d'assurés ayant eu des antécédents de sinistralité, et que le courtier C est celui qui apporte le plus d'assurés ayant eu au moins un sinistre responsable. Sur l'ensemble du portefeuille, 25% des assurés ont déclaré au moins un sinistre responsable au cours des trois dernières années précédant la date de début de leur contrat.

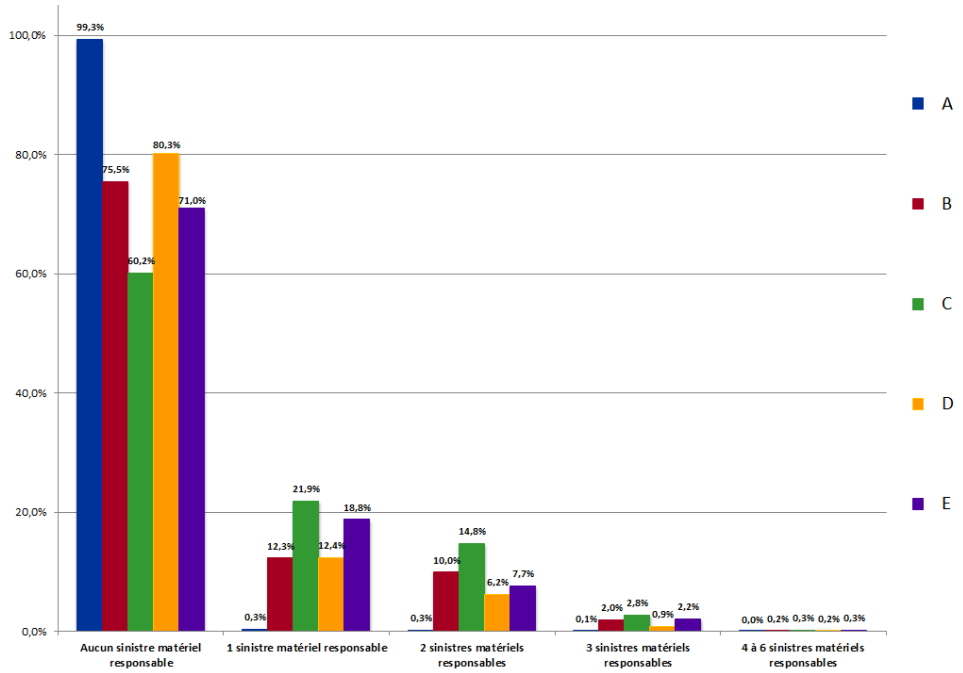

Figure 14 Répartition des antécédents de sinistralité matérielle responsable selon les courtiers

Les descriptions des conducteurs et de leurs antécédents mettent en avant des différences liées aux conditions de souscription qui varient selon les courtiers apporteurs. Pour résumer, il apparait que le courtier B est apporteur d'assurés jeunes ou ayant une faible expérience de conduite. Ces conducteurs novices n'ont eu que très peu de résiliations par un assureur précédent. À l'inverse, les assurés de l'apporteur E ont presque tous fait l'objet d'une résiliation, le plus souvent pour non-paiement de prime. Le courtier A apporte en général des assurés ayant eu une suspension de permis. Le courtier D possède un grand nombre de clients en situation de malus. Enfin, le courtier C attire les profils d'assurés ayant beaucoup d'antécédents de sinistralité et ayant entraîné une résiliation pour cause de fréquence de sinistralité trop élevée.

Malgré ces disparités liées aux conditions de souscription, il est judicieux de regrouper les données. En effet, cela permet d'avoir plus de données pour la modélisation et d'avoir une vision générale des contrats automobile aggravés dans le but de proposer de nouveaux partenariats ou d'élargir les conditions de souscription. Cette pertinence vient du fait qu'en dehors du cadre de souscription, la distribution des assurés selon les apporteurs suit la même tendance. Cette similarité est montrée dans la suite des statistiques descriptives du portefeuille.

Répartition des contrats par département :

La répartition des contrats est globalement la même selon les différents courtiers. Ainsi, la plupart des contrats ont été souscrits pour des conducteurs vivant en Île-de-France, dans le Nord-Pas-de-Calais, le long de la côte méditerranéenne, en Gironde, en Haute-Garonne, dans le Rhône ou en Isère.

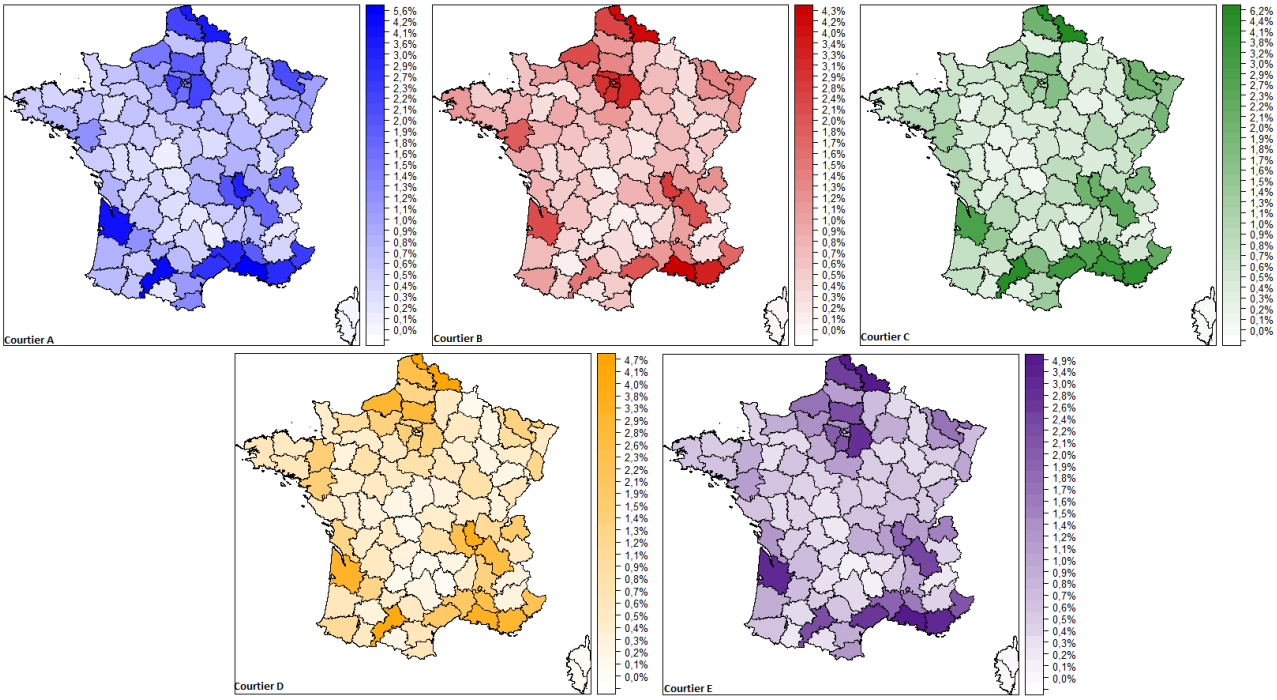

Figure 15 Répartition des contrats par département et par courtier

Description des véhicules assurés :

Les données concernant les véhicules sont homogènes pour l'ensemble des partenaires. Cette constatation peut être observée sur les trois graphiques suivants.

Les véhicules assurés ont entre 0 et 68 ans. Deux tiers d'entre eux ont été mis en circulation il y a 6 à 17 ans. L'âge moyen du parc des véhicules assurés est de 11,87 ans, soit un an de plus que la moyenne nationale.

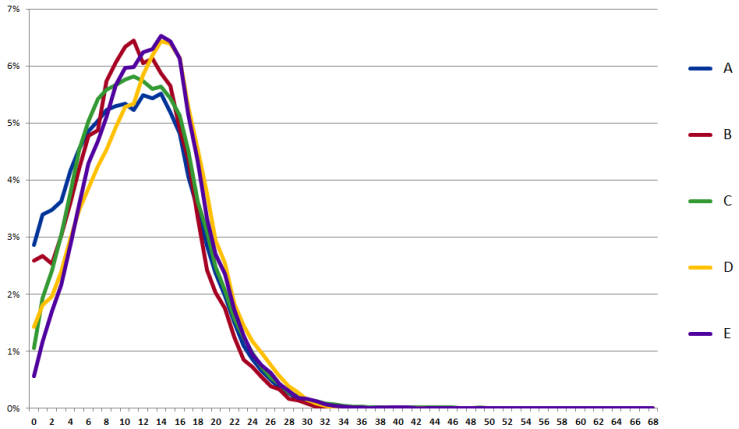

Figure 16 Distribution de l'âge des véhicules assurés par courtier apporteur

La classe est déterminée par le SRA en fonction du prix à neuf du véhicule et du prix de ses pièces détachées. Cette variable a donc un impact sur les garanties dommages matériels. Les classes observées sur l'ensemble du portefeuille varient de A à Z. 22 véhicules sont « hors classe ». Il s'agit de véhicules dont la valeur à neuf peut s'élever à plus de 300 000 €. Néanmoins, ils représentent une très faible minorité du portefeuille. En effet, 46% des véhicules appartiennent aux classes I, J, K et L. Ces véhicules ont une valeur moyenne à neuf de 22 000 €.

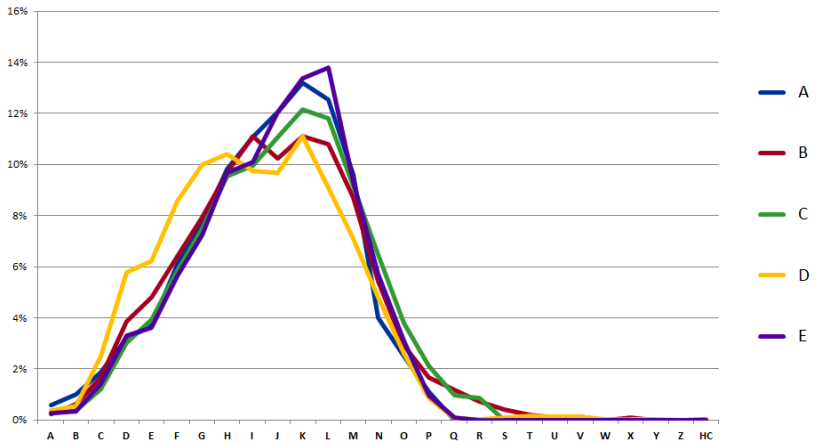

Figure 17 Distribution de la classe SRA des véhicules assurés par courtier apporteur

Le groupe est fixé par le SRA selon la dangerosité du véhicule. Il constitue donc une variable importante pour la garantie « responsabilité civile ». 80% des véhicules du portefeuille appartiennent à seulement cinq groupes, ceux variant de 28 à 32. Par ailleurs, 38% de l'ensemble des véhicules sont rattachés soit au groupe 30, soit au groupe 31. Ce type de véhicule correspond à des voitures dont la puissance fiscale moyenne est de 6 cv. Cette variable présentant certaines données incohérentes, un retraitement a été effectué. Ainsi, 0,04% des données affichaient un groupe strictement inférieur à 20, ce qui est théoriquement impossible. Celui-ci a donc été remplacé par 20. De plus, 0,03% des données étant incomplètes, elles ont été remplacées par le groupe 30, qui correspond au groupe médian.

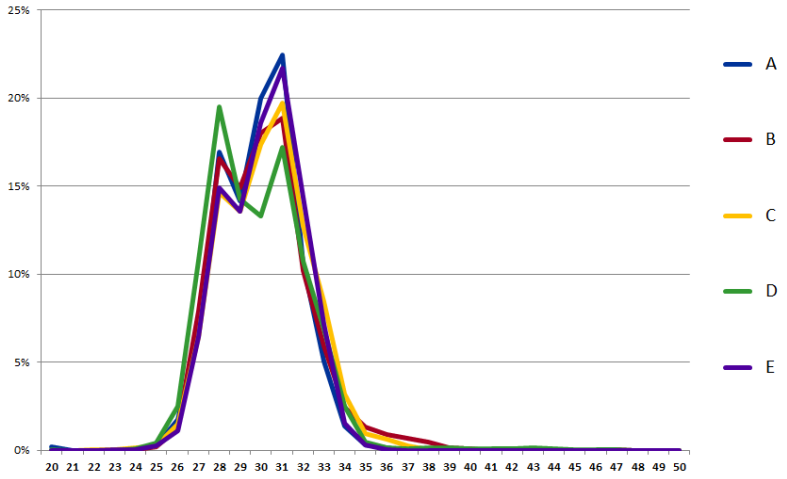

Figure 18 Distribution du groupe SRA des véhicules par courtier apporteur

Lorsqu'un assuré souscrit un contrat d'assurance automobile, il déclare l'usage qu'il effectuera de son véhicule. Trois types d'utilisation se distinguent :

- Vie privée : le véhicule n'est utilisé que pour des déplacements dans le cadre de la vie privée.

- Vie privée et trajet-travail : il s'agit des déplacements de la vie privée et des allers-retours du domicile au lieu de travail ou d'études. En cas d'accident de la route sur le chemin du travail, les lésions corporelles du conducteur sont couvertes par l'Assurance Maladie. Les dommages matériels occasionnés au véhicule assuré ainsi que tous les dommages matériels et corporels subis par les tiers sont couverts par l'assurance automobile du conducteur, si celui-ci est responsable.
- Vie privée et affaires : le véhicule est assuré pour tous les déplacements de la vie privée et professionnelle (y compris pour les trajets domicile-travail).

Sur l'ensemble du portefeuille, 24% des assurés ont souscrit une assurance « vie privée seule », 71% d'entre eux utilisent leur véhicule pour se rendre au travail et 5% effectuent des trajets privés et professionnels.

Le diagramme ci-dessous détaille ces pourcentages selon les cinq courtiers apporteurs. La répartition de l'usage déclaré du véhicule est globalement la même pour tous les courtiers. Cependant, il est à noter que le courtier E présente un taux d'assurés n'utilisant leur véhicule que dans un cadre privé plus élevé que les autres. Par ailleurs, une part importante des assurés du courtier B déclarent utiliser leur véhicule à des fins privées et pour se rendre sur leur lieu de travail.

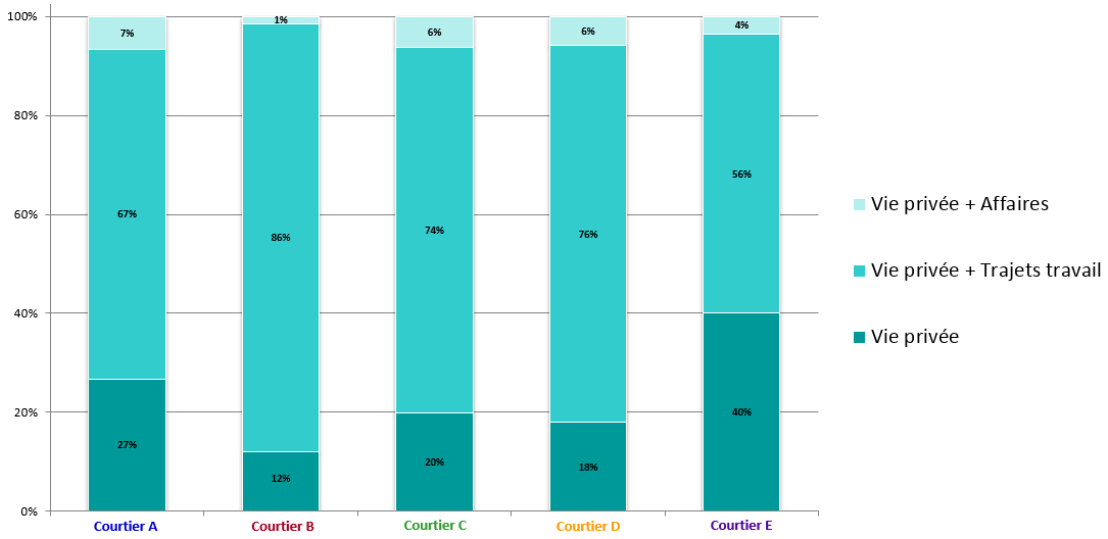

Figure 19 Distribution de l'usage du véhicule par courtier

La description des variables permet de comprendre la répartition des assurés et de leur véhicule dans le portefeuille.

Du fait des conditions de souscription qui diffèrent selon les courtiers, ceux-ci ne sont pas apporteurs des mêmes types de conducteurs. Le distributeur A est principalement apporteur de conducteurs ayant eu une suspension de permis. Le courtier B possède essentiellement des jeunes conducteurs et ayant peu d'expérience de conduite. Le courtier C est apporteur d'assurés ayant des antécédents importants de sinistralité (sinistres corporels et matériels). Une grosse proportion des assurés du courtier D sont en situation de malus. Enfin, de nombreux assurés du courtier E ont été résilié par leur précédent assureur à la suite d'un non-paiement de prime. Cependant, les caractéristiques des véhicules assurés sont les mêmes pour l'ensemble des courtiers apporteurs : l'âge moyen des automobiles est de 12 ans, elles appartiennent essentiellement aux classes I à L et aux groupes 28 à 32 et sont utilisées pour des déplacements privés et pour se rendre sur le lieu de travail. Ces véhicules circulent essentiellement en Île-de-France, dans le Nord-Pas-de-Calais et le long de la côte méditerranéenne. Ceci permet de conclure qu'il est pertinent de concaténer les bases des cinq courtiers puisque hormis les conditions de souscription qui diffèrent, les bases présentent de nombreux points communs.

Avant de chercher à modéliser la prime pure, il est nécessaire d'analyser la sinistralité du portefeuille afin d'en percevoir les différentes tendances.

2.2.1.2 Statistiques de la sinistralité

Comme évoqué dans le premier chapitre, les deux éléments les plus importants concernant la sinistralité sont le coût moyen d'un sinistre et la fréquence de survenance d'un sinistre.

Dans le but d'avoir un premier aperçu des profils les plus à risque, il est essentiel d'analyser la sinistralité selon différentes variables.

Sinistralité par garantie :

En moyenne, les sinistres coûtent à l'assureur 1 916 €. Cependant, du fait de leur nature différente, il n'est pas cohérent de les comparer directement. Pour plus de significativité, le tableau suivant présente une analyse des coûts de sinistres et de la fréquence de chaque type de sinistre.

Garantie	Coût médian	Coût moyen	Coût maximal	Répartition du nombre de sinistres	Répartition de la charge de sinistres	Fréquence de sinistralité
Bris de glace	472 €	479 €	5 228 €	29,08%	6,52%	8,37%
Catastrophe naturelle	2 000 €	3 669 €	21 344 €	0,17%	0,32%	0,07%
Dommages	1 692 €	2 345 €	74 719 €	8,69%	8,18%	4,42%
DPRSA	158 €	517 €	3 975 €	0,18%	0,05%	0,02%
Incendie	2 571 €	3 712 €	31 220 €	0,72%	1,40%	0,43%
RC corporelle	4 344 €	12 364 €	4 303 121 €	2,51%	48,47%	0,42%
RC matérielle	762 €	1 071 €	203 639 €	56,70%	31,69%	10,01%
Vol	1 478 €	3 303 €	31 266 €	1,96%	3,37%	1,10%

Tableau 5 Description des montants de sinistres et des fréquences par garantie couverte

Les trois garanties représentant la plus grosse charge sont la responsabilité civile, les dommages et le bris de glace.

Il est intéressant de comparer les informations résultant de cette analyse à la moyenne nationale. Les données étant issues de contrats « aggravés », il parait incontestable que la sinistralité est plus forte pour les données étudiées que pour la moyenne nationale. En effet, la fréquence de sinistralité en responsabilité civile matérielle pour les contrats aggravés s'élève à 10,01%, alors qu'elle n'atteint que 3,2% au niveau national. Il en est de même pour la fréquence des vols, qui s'élève à 0,43% au niveau national et à 1,10% dans le cas de cette étude, ainsi que pour la fréquence des bris de glace (6,69% en France, 8,37% pour les contrats aggravés) et pour la fréquence de la responsabilité civile corporelle (0,34% en France, 0,43% pour les risques aggravés). Seule la garantie dommages au véhicule présente un taux de sinistralité plus faible que la moyenne nationale (8,69% en France, 4,42% pour les données de cette étude). Le coût moyen de chacune de ces garanties correspond aux coûts moyens nationaux, cela peut être expliqué par le fait que les caractéristiques du véhicule d'un assuré ne dépendent pas des antécédents d'assurance de ce dernier. Par conséquent, les coûts de réparation ou d'indemnisation sont les mêmes qu'au niveau national.

Par ailleurs, la proportion de la charge totale due à la garantie responsabilité civile est beaucoup plus importante qu'au niveau national. Cela est explicable par la nature des contrats aggravés : 49% des assurés en portefeuille ont souscrit uniquement la garantie responsabilité civile, d'où le fait que la plupart des sinistres déclarés sont associés à la responsabilité civile.

L'objectif de se mémoire étant de comparer deux types de méthodes de modélisation : le modèle linéaire généralisé et des méthodes de Machine Learning, il est indispensable d'avoir suffisamment de données à disposition. Or, les sinistres liés à une catastrophe naturelle, de la protection juridique, un incendie ou un vol représentent très peu de dossiers et les données sur les coûts sont trop peu nombreuses pour appliquer un modèle linéaire généralisé ou des méthodes de Machine Learning. Ces garanties ne seront donc pas étudiées dans le cadre de ce mémoire. Cependant, la modélisation de la prime pure peut être effectuée en appliquant d'autres méthodes. Par ailleurs, l'étude des sinistres de responsabilité civile matérielle porte essentiellement sur la fréquence. En effet, la plupart des sinistres entraînent l'application de la convention IRSA. Par conséquent, lorsque l'assuré est responsable, le montant d'indemnisation est réglementaire donc connu. Du fait de la fréquence très faible et des coûts pouvant s'élever à plusieurs millions d'euros, l'étude de la responsabilité civile corporelle est indispensable à l'assureur. Cependant, ces sinistres doivent faire l'objet d'une modélisation spécifique et approfondie. Par conséquent, la modélisation de la prime pure pour la garantie responsabilité civile ne sera pas étudiée pour ce mémoire.

L'objet de ce mémoire est la comparaison de méthodes de modélisation de la prime pure. Ces méthodes sont appliquées à deux garanties : le bris de glace et les dommages du véhicule.

Sinistralité par courtier :

Pour justifier la pertinence de concaténer les bases de cinq courtiers afin de proposer un tarif commun, il est important de vérifier que la sinistralité est homogène pour l'ensemble de ces courtiers.

La comparaison des coûts et des fréquences de sinistralité en fonction des différents courtiers est possible car les franchises et les plafonds appliqués sont les mêmes pour les polices d'assurance automobile « risque aggravé » des différents courtiers.

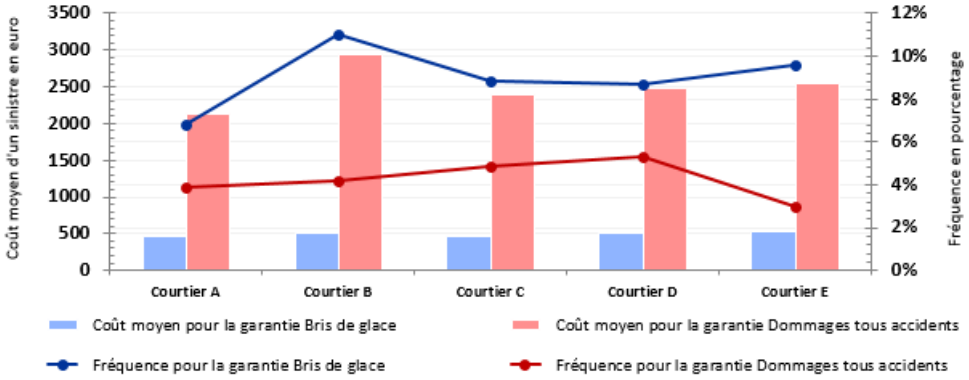

Figure 20 Coût moyen et fréquence des sinistres bris de glace et dommages par courtier apporteur

Il apparait ainsi que le coût moyen des sinistres bris de glace est identique pour les cinq courtiers. Le coût moyen et la fréquence de la garantie dommages sont assez semblables selon les courtiers. Seuls les partenaires A et B présentent des différences relatives à la fréquence de survenance des sinistres bris de glace, ce qui n'est ne pose pas de problème car les statistiques suivantes ont pour but de montrer que les écarts observés sont dus aux caractéristiques des assurés de ces deux courtiers et non à d'autres variables exogènes.

Sinistralité par zone définie grâce au zonier :

Le but d'un modèle mathématique est de capter les principaux facteurs de risques et les tendances de sinistralité liées à ces risques. Cependant, un des enjeux majeurs des modèles est d'éviter le surapprentissage. Il a lieu lorsque le modèle prend un cas particulier pour une généralité. Ce phénomène peut se produire lorsque les variables sont trop granulaires car il peut exister des modalités pour lesquelles aucun sinistre n'a eu lieu. Par exemple, une analyse de la sinistralité basée sur l'ensemble des communes françaises conduirait à considérer qu'une majorité d'entre elles ont un risque nul, ce qui n'est pas le cas en réalité. Par conséquent, la variable zone (issue du zonier) a été ajoutée à partir du code postal du lieu de garage du véhicule afin de discriminer la zone géographique tout en réduisant le nombre de modalités.

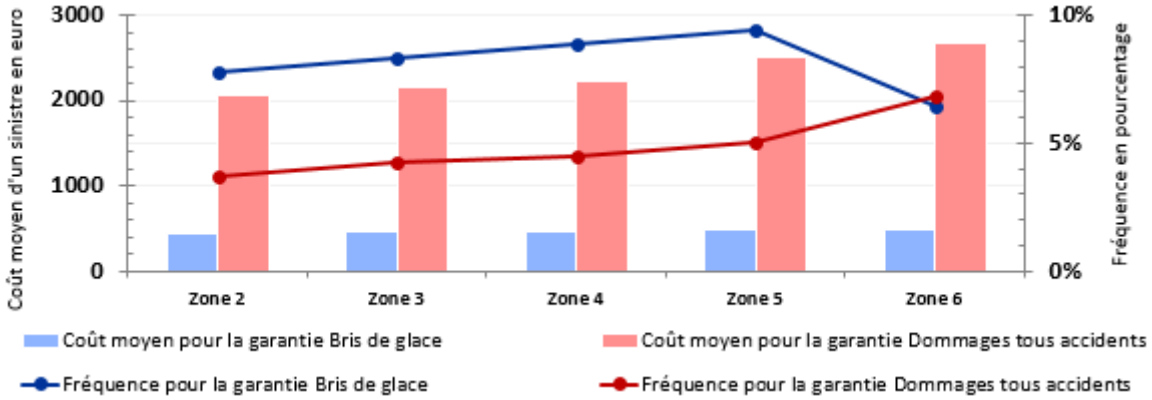

Figure 21 Coût moyen et fréquence des sinistres par garantie et par zone

Il est important de remarquer que globalement, le coût moyen et la fréquence des sinistres augmentent lorsque la zone est plus risquée (zone 6). Cela prouve que le zonier est adapté au portefeuille étudié et qu'il n'est pas nécessaire de l'ajuster.

Sinistralité par type de véhicule :

Les véhicules de la base de données ayant la même répartition selon les courtiers, il est important d'étudier leur sinistralité.

Comme l'a démontré l'analyse descriptive de la base, la répartition de chaque modalité des variables n'est pas équirépartie. Toujours dans le but d'éviter le surapprentissage pour les modalités sous-représentées et dans le but de réduire le nombre de modalités, des regroupements ont été effectués. Ceux-ci sont basés sur une première analyse de la sinistralité afin de lisser les résultats et ont été effectuées par groupe homogènes. Cette étape n'étant pas présentée dans le mémoire, les résultats groupés sont directement exposés.

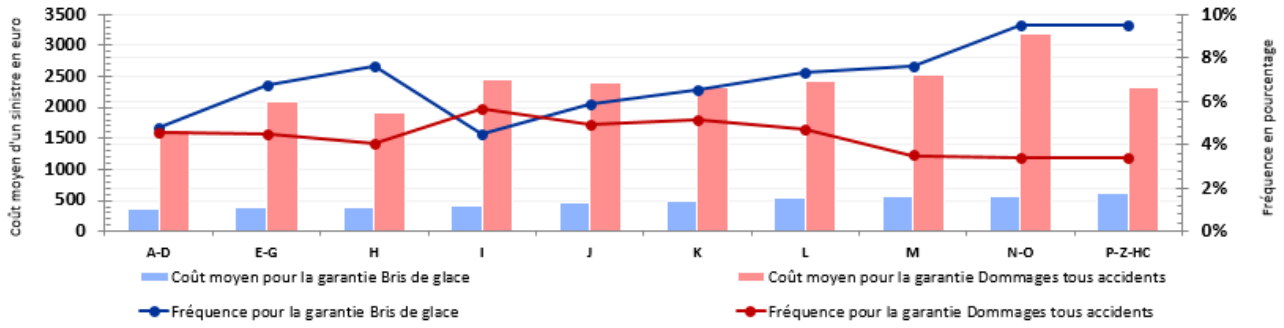

Figure 22 Coût moyen et fréquence des sinistres par garantie et par classe SRA des véhicules

Il est à noter que globalement, le coût moyen des sinistres augmente avec la classe du véhicule. En effet, par définition, la classe définie par le SRA correspond à la valeur de l'automobile et de ses pièces de remplacement.

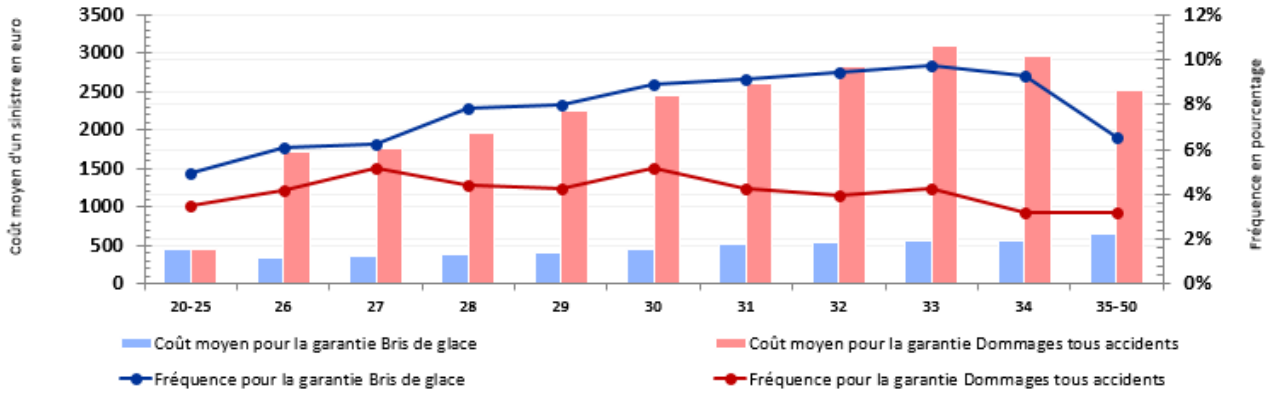

Figure 23 Coût moyen et fréquence des sinistres par garantie et par groupe SRA des véhicules

De manière générale, les coûts des sinistres augmentent en fonction des groupes SRA. Il en est de même pour la fréquence de survenance des bris de glace. À l'inverse, la fréquence des dommages a tendance à diminuer lorsque le groupe augmente.

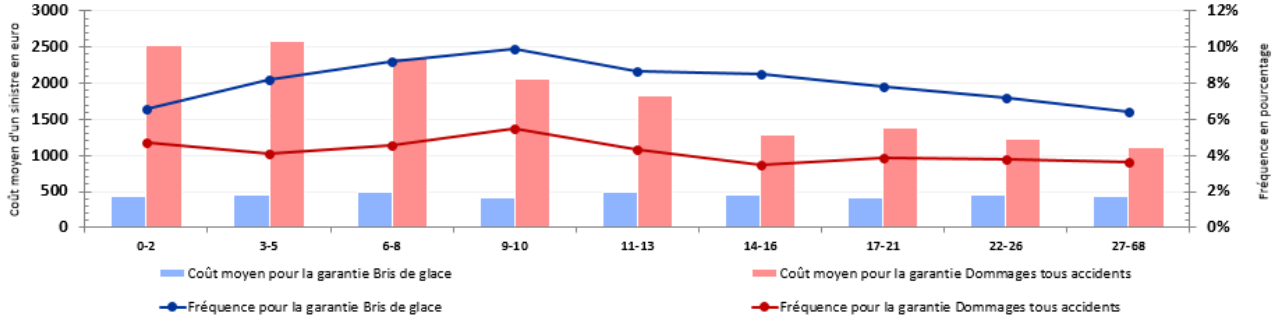

Figure 24 Coût moyen et fréquence des sinistres par garantie et par âge du véhicule

Contrairement au groupe et à la classe SRA où le coût moyen augmente avec ces deux derniers, le coût moyen baisse si le véhicule est vieux. En effet, en cas de remplacement ou de réparation d'une voiture neuve, il est évident que le coût sera plus important que si la voiture était ancienne et avait perdu de la valeur au cours des années. En revanche, les courbes représentant la fréquence de sinistralité ne semblent pas suivre de tendance particulière.

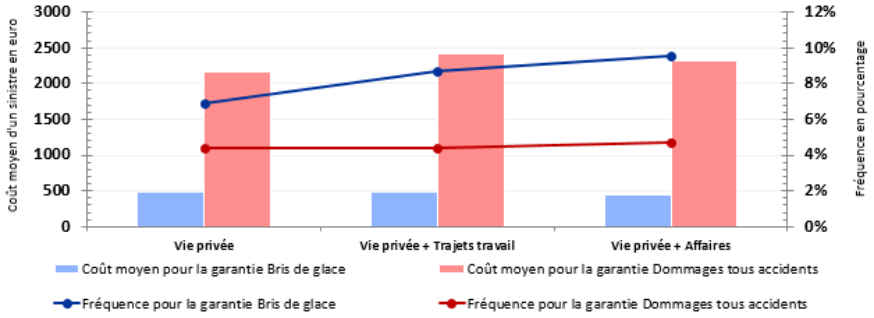

Figure 25 Coût moyen et fréquence des sinistres par garantie et par usage

L'usage du véhicule a seulement un impact sur la fréquence de sinistralité des bris de glace. En effet, les conducteurs utilisant leur véhicule pour un usage professionnel conduisent plus et sont donc plus exposés au risque que les autres conducteurs. En revanche, l'usage n'a aucun effet sur la fréquence des dommages ni sur le coût des sinistres car les frais sont liés au type de véhicule et non à son usage.

Sinistralité par type de conducteur :

Ce paragraphe présente les statistiques de sinistralité selon les caractéristiques du conducteur traditionnellement utilisées pour tarifer un produit d'assurance automobile. Puis, les spécificités du portefeuille seront analysées. Un portefeuille d'assurance « classique » ne possède que très peu d'individus ayant fait l'objet d'une résiliation chez un assureur précédent ou possédant un CRM élevé, ce qui n'est pas le cas des assurés de produits « aggravés ».

Les résultats issus de l'analyse en fonction de l'âge du conducteur et de son ancienneté de permis aboutissent aux mêmes conclusions. En effet, en général, les conducteurs passent leur permis quand ils sont jeunes. Par conséquent, ces deux variables sont liées.

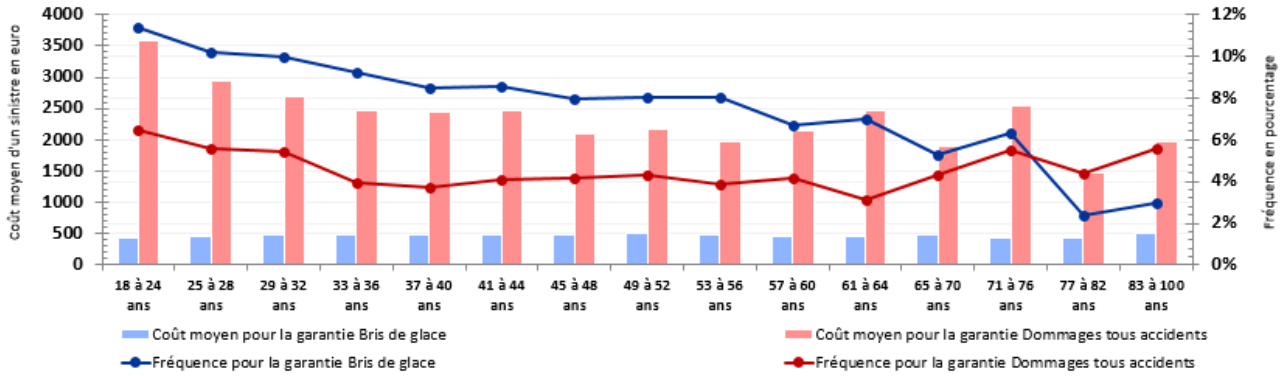

Figure 26 Coût moyen et fréquence des sinistres par garantie et par âge du conducteur

Ainsi, il est à noter que plus le conducteur est âgé et plus il a de l'expérience dans la conduite, moins il a de sinistres. Néanmoins, cette observation n'est plus valable pour la garantie dommages au véhicule lorsque le conducteur a plus de 65 ans ou plus de 53 ans d'expérience de conduite. Par ailleurs, le coût des dommages baisse également en fonction de ces deux variables tandis que le coût du bris de glace ne semble pas particulièrement impacté par celles-ci.

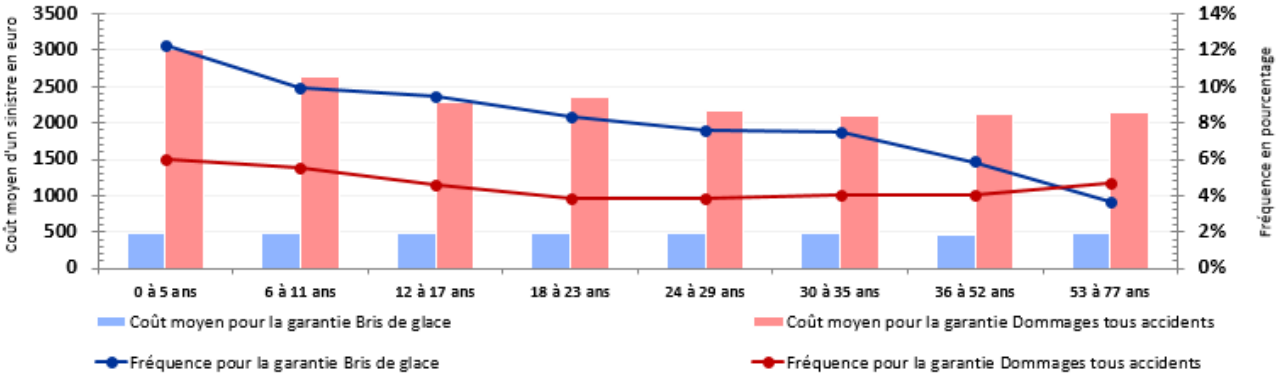

Figure 27 Coût moyen et fréquence des sinistres par garantie et par ancienneté d'obtention du permis de conduire

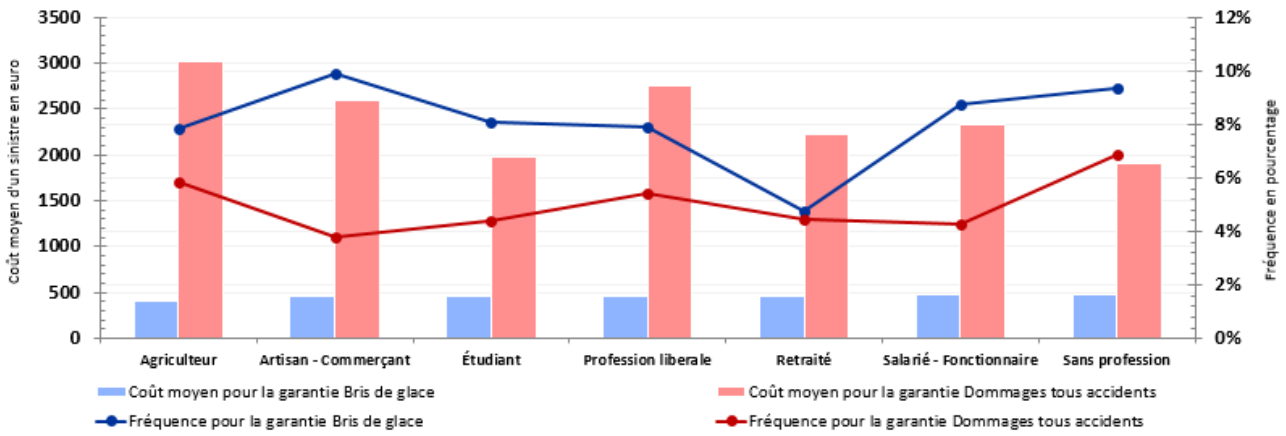

Figure 28 Coût moyen et fréquence des sinistres par garantie et par catégorie socio-professionnelle

Pour l'ensemble des garanties, les conducteurs exerçant une profession libérale ou n'ayant pas de profession présentent les taux de sinistralité les plus élevés. À l'inverse, les retraités ont un taux de sinistralité moins élevé. Par ailleurs, le coût moyen des bris de glace varie très peu en fonction de la catégorie socio-professionnelle.

Après avoir analysé les variables descriptives communes à tous les contrats d'assurance automobile, il est intéressant d'étudier les variables caractéristiques des risques aggravés.

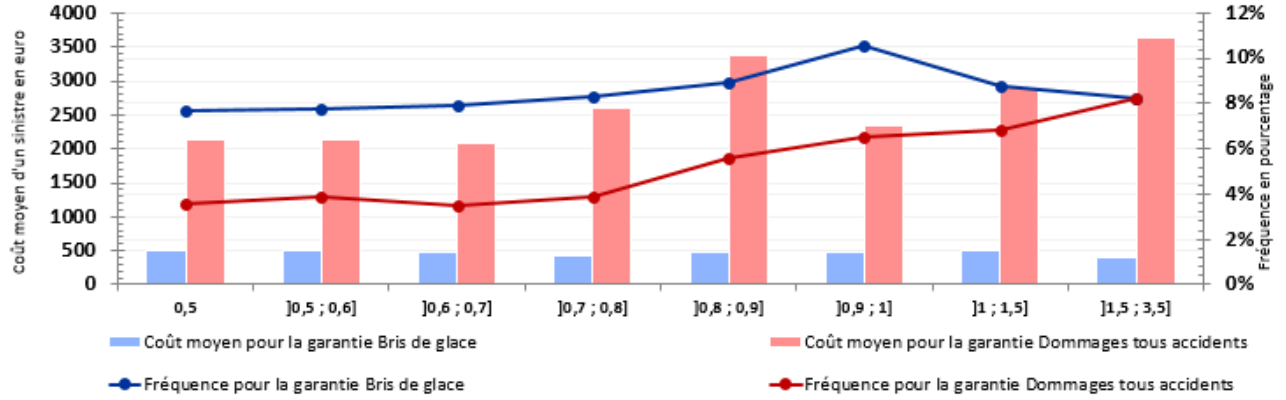

Figure 29 Coût moyen et fréquence des sinistres par garantie et par coefficient de réduction-majoration

Le coefficient de réduction-majoration étant lié à l'expérience de conduite de l'assuré et à sa sinistralité responsable passée, il parait évident que la fréquence des dommages au véhicule baisse lorsque le CRM diminue.

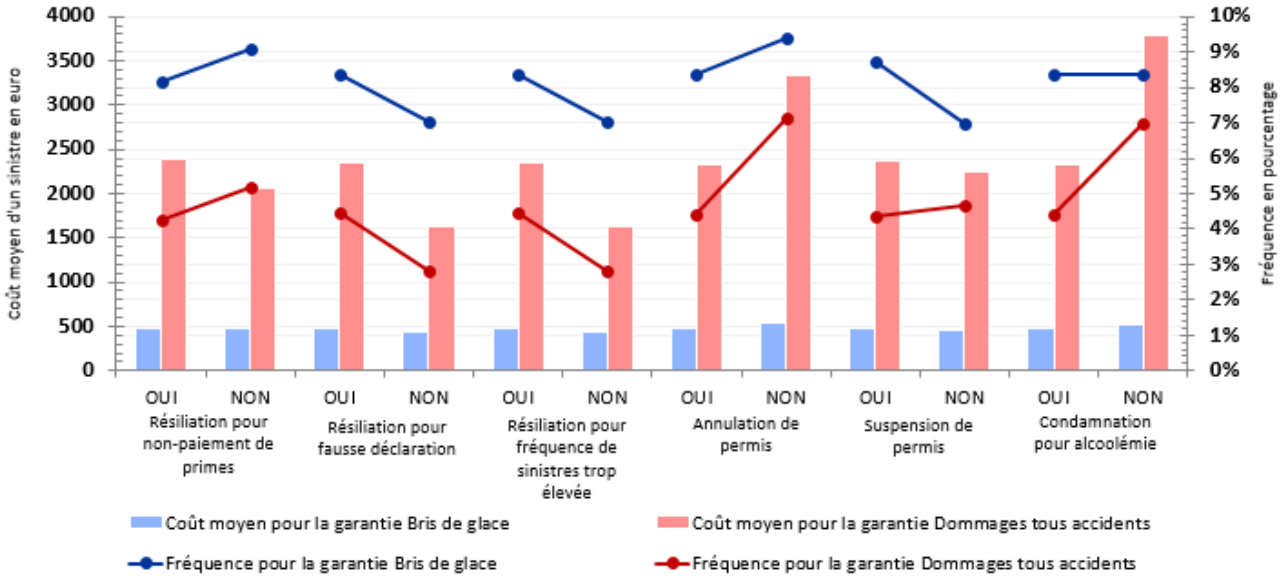

Figure 30 Coût moyen et fréquence des sinistres par garantie et par cause de résiliation

Les deux motifs entraînant une forte hausse de la fréquence et du coût moyen des deux garanties étudiées sont une annulation de permis et une condamnation pour alcoolémie.

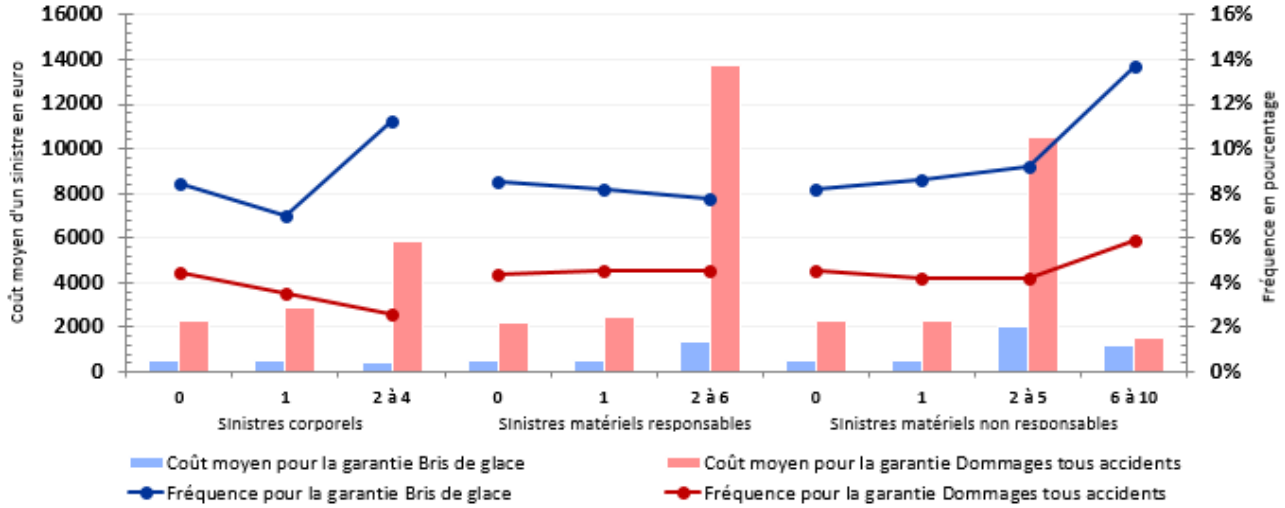

Figure 31 Coût moyen et fréquence des sinistres par garantie et par antécédents de sinistralité

Le nombre de sinistres corporels et matériels responsables et non responsables sur les trois dernières années est demandé aux assurés lors de la souscription d'un contrat. Il est incontestable que d'après le diagramme ci-dessus, plus l'assuré a des antécédents importants, plus ses coûts de sinistres sont élevés. De plus, la fréquence des deux garanties étudiées augmente en fonction du nombre de sinistres matériels non responsables.

L'analyse descriptive des variables et de la sinistralité permet d'avoir un aperçu général des facteurs présentant des taux de sinistralité ou des coûts moyens élevés. Ainsi, les assurés vivant dans des zones classées 5 ou 6 d'après le zonier ont une fréquence et un coût de sinistralité plus élevés. De manière générale, les caractéristiques du véhicule ont un impact sur le coût des sinistres tandis que les caractéristiques des conducteurs influent sur la fréquence. Ainsi, plus la classe et le groupe SRA sont grands, plus la fréquence et le coût moyen des sinistres le sont également. En revanche, le coût moyen des

vieux véhicules est inférieur à celui des véhicules neufs. Par ailleurs, les profils d'assurés ayant les plus hauts taux de sinistralité sont les artisans, les commerçants et les professions libérales, les jeunes conducteurs et ceux ayant obtenu leur permis de conduire récemment. De plus, les antécédents de sinistralité sont des éléments importants puisque plus le CRM est élevé, plus la fréquence et le coût moyen sont importants. Il en est de même pour les assurés ayant eu un grand nombre de sinistres au cours des trois dernières années et ceux ayant fait l'objet d'une annulation de permis*.

2.2.2 Dépendance des variables

2.2.2.1 Indépendance du coût et de la fréquence

Pour pouvoir appliquer le modèle « Coût moyen × fréquence » évoqué dans le premier chapitre, il est nécessaire que le coût et la fréquence soient indépendants.

Pour tester la relation entre deux variables, la corrélation est utilisée. Il s'agit d'une analyse bivariée mesurant le niveau de dépendance entre ces variables. La valeur du coefficient de corrélation varie entre -1 (les variables sont anti-corrélées) et 1 (les variables sont complètement corrélées). Si les variables sont indépendantes, alors le coefficient de corrélation est nul. Plusieurs mesures de la corrélation existent.

En notant $(x_1; \dots; x_n)$ les observations de la variable X et $(y_1; \dots; y_n)$ les observations de la variable Y, il est possible de définir trois coefficients de corrélation :

Coefficient de corrélation de Pearson :

Il mesure la présence ou l'absence de relation linéaire entre les deux variables et est égal à la covariance des variables divisée par le produit de leur écart-type.

$$r(X;Y) = \frac{\mathrm{cov}(X;Y)}{\sigma_X \sigma_Y} = \frac{\sum(x_i y_i) - \frac{\sum x_i \sum y_i}{n}}{\sqrt{\sum {x_i}^2 - \frac{(\sum x_i)^2}{n}}\sqrt{\sum {y_i}^2 - \frac{(\sum y_i)^2}{n}}}$$

Coefficient de corrélation de Kendall :

Contrairement au coefficient de corrélation de Pearson, il s'agit non plus de rechercher une relation entre les valeurs des variables, mais entre leur rang, ce qui permet de mettre en évidence des relations non-linéaires. Les paires d'observations $(x_i; y_i)$ et $(x_j; y_j)$ sont dites concordantes si $x_i < x_j \Rightarrow y_i < y_j$ ou si $x_i > x_j \Rightarrow y_i > y_j$. À l'inverse, elles sont dites discordantes si $x_i < x_j \Rightarrow y_i > y_j$ ou si $x_i > x_j \Rightarrow y_i < y_j$.

$$\tau(X;Y) = \frac{nombre\ de\ paires\ concordantes - nombre\ de\ maires\ discordantes}{\frac{n(n-1)}{2}}$$

Coefficient de corrélation de Spearman :

Ce coefficient se base sur celui de Pearson, mais est calculé pour le rang des variables.

$$\rho(X;Y) = \frac{\mathrm{cov}(\mathrm{rg}_X; \mathrm{rg}_Y)}{\sigma_{\mathrm{rg}_X}\sigma_{\mathrm{rg}_Y}} = 1 - \frac{6\sum_{i=1}^{n}\left(\mathrm{rg}_{x_i} - \mathrm{rg}_{y_i}\right)^2}{n^3 - n}$$

* Le courtier B est le courtier qui possède le plus de clients en situation de malus, ayant eu une annulation de permis et de jeunes conducteurs. Par conséquent, il s'agit du courtier dont la fréquence de sinistralité est la plus grande.

Le calcul de ces trois coefficients sur les garanties étudiées est résumé dans le tableau ci-dessous. Les coefficients sont très faibles pour les deux garanties, ce qui permet d'en déduire qu'il n'existe pas de relation entre le coût moyen et la fréquence des sinistres.

Garantie	Bris de glace	Dommages tous accidents
Pearson	0,0111	0,125
Kendall	-0,0116	0,0567
Spearman	-0,0171	0,0849

Tableau 6 Valeurs des coefficients de corrélation

Néanmoins, il est à noter que l'indépendance entraîne la non-corrélation mais que la réciproque n'est pas toujours vérifiée. Dans le but d'approfondir la vérification de l'hypothèse d'indépendance, il est possible d'utiliser la théorie des copules.

La fonction de répartition bivariée de $(X;Y)$ est caractérisée par $F(x;y) = \mathbb{P}(X \leq x; Y \leq y)$. Les lois marginales de X et Y sont respectivement $F_1(x) = \mathbb{P}(X \leq x)$ et $F_2(y) = \mathbb{P}(Y \leq y)$. Si les deux variables X et Y sont indépendantes, alors $F(x;y) = F_1(x)F_2(y)$. Une copule est une fonction de répartition bivariée C ayant pour marginales la loi uniforme sur $[0;1]$. D'après le théorème de Sklar, en reprenant les notations précédentes, si F est une fonction de répartition bivariée de marginales F_1 et F_2, alors il existe une copule C telle que pour tout $(x;y) \in \mathbb{R}^2 : F(x;y) = C\big(F_1(x);F_2(y)\big)$ (De plus, si F_1 et F_2 sont continues, alors C est unique). Par conséquent, la copule indépendante, également nommée copule produit, est définie par $C\big(F_1(x);F_2(y)\big) = F_1(x)F_2(y)$. Une copule peut être représentée graphiquement grâce à ses lignes de niveaux. Les lignes de niveau de la copule indépendante sont représentées sur la figure suivante.

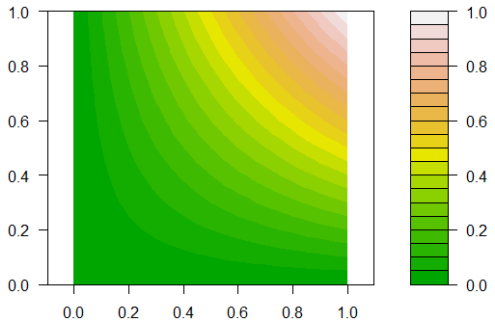

Figure 32 Lignes de niveau de la copule indépendante

Pour les deux garanties étudiées, il est possible d'estimer la copule empirique $C_n(u;v) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\left\{\frac{\mathrm{rg}_{X_i}}{n}<u;\frac{\mathrm{rg}_{Y_i}}{n}<v\right\}}$. Cela permet de tracer les lignes de niveaux ci-dessous :

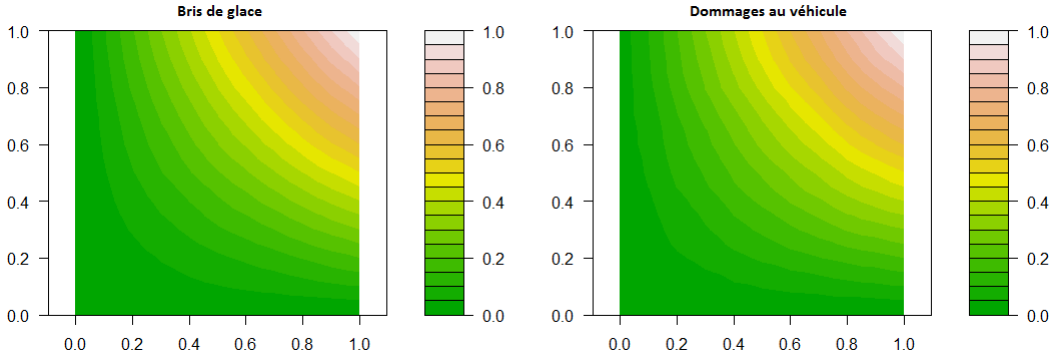

Figure 33 Lignes de niveau des copules empiriques estimées pour les garanties bris de glace et dommages tous accidents

Sur les deux graphiques, les lignes de niveau sont similaires à celles d'une copule indépendante, ce qui permet de conclure que l'hypothèse d'indépendance entre le coût et la fréquence est respectée.

2.2.2.2 Variables explicatives

Afin de modéliser le coût moyen des sinistres ou la fréquence grâce au modèle linéaire généralisé, il est nécessaire que les variables explicatives soient indépendantes. En effet, garder plusieurs variables fortement liées peut s'avérer problématique lors de la modélisation car cela peut engendrer des problèmes de stabilité du modèle (la variance des résultats d'une modélisation avec des variables dépendantes est plus grande que la variance d'une modélisation où ce n'est pas le cas, les résultats sont alors instables, ce qui les rend difficiles à interpréter).

L'intensité d'une relation entre deux variables qualitatives peut être mesurée grâce au V de Cramer.

Le test du $\mathcal{X}^2$ permet de déterminer s'il existe une liaison entre deux variables, mais il ne donne pas l'intensité de cette liaison, contrairement au V de Cramer qui indique si celle-ci est forte ou modérée. Il est défini par :

$$V = \sqrt{\frac{\mathcal{X}^2}{N\,(\min(l,c)-1)}}$$

avec :

- $\mathcal{X}^2$: test du $\mathcal{X}^2$: $\mathcal{X}^2 = \sum_{i=1}^{l}\sum_{j=1}^{c}\frac{\left(n_{i,j}-\frac{n_{i.}\,n_{.j}}{N}\right)^2}{\frac{n_{i.}\,n_{.j}}{N}}$ où $\frac{n_{i.}\,n_{.j}}{N}$ représente l'effectif théorique sous l'hypothèse d'indépendance et $n_{i,j}$ l'effectif réellement observé
- N : nombre d'observations
- l : nombre de modalités de la première variable
- c : nombre de modalités de la deuxième variable

Par définition, les valeurs de V sont comprises entre 0 (indépendance) et 1(colinéarité). Deux variables sont considérées fortement corrélées lorsque V est supérieur à 0,3.

Ainsi, pour chacune des 21 variables étudiées, le V de Cramer peut être calculé. Comme les variables jouent un rôle symétrique, seule l'analyse du triangle inférieur de la matrice est suffisante. Les résultats obtenus peuvent être représentés sous forme graphique.

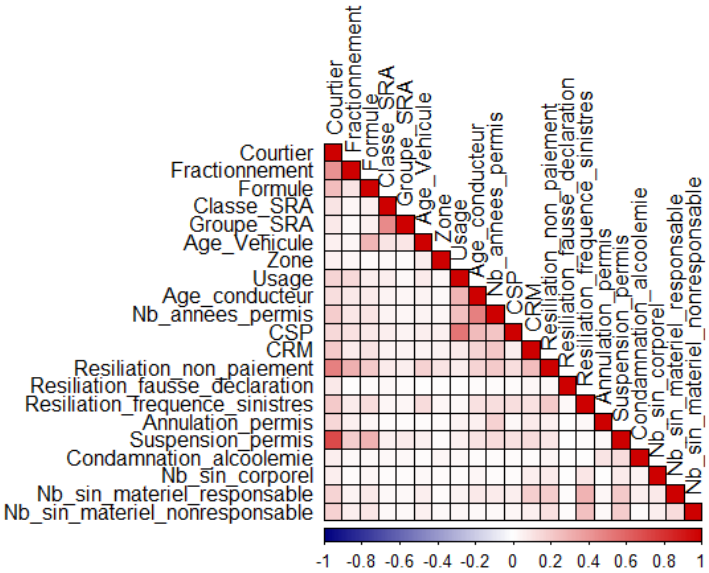

Figure 34 Représentation graphique des coefficients du V de Cramer

D'après la représentation graphique du V de Cramer, il apparait que certaines variables sont fortement dépendantes :

- V(Courtier ;Suspension de permis) = 0,70
- V(Usage ;CSP) = 0,53
- V(Courtier :Résiliation pour non-paiement) = 0,50
- V(Classe SRA ; Groupe SRA) = 0,45
- V(Courtier ;Fractionnement) = 0,42
- V(Fractionnement ;Résiliation pour non-paiement) = 0,31

En effet, comme l'ont montré les statistiques descriptives, les assurés ayant eu une suspension de permis sont, pour la plupart, apportés par le courtier A et ceux n'ayant pas payé leur prime par le courtier E. De plus, il est normal que les variables *fractionnement* et *résiliation pour non-paiement* soient dépendantes. En effet, les assurés ne payant pas leurs primes sont en général des personnes ayant des problèmes financiers. Par conséquent, ils préfèrent étaler leurs dépenses mensuellement. Il est nécessaire de ne pas prendre en compte la variable *Courtier* qui est dépendante de trois autres variables. Ce choix parait raisonnable car l'objectif de la Direction des Partenariats est de proposer un tarif commun à tous les courtiers. Par ailleurs, la variable *fractionnement* n'est également pas conservée car elle est dépendante d'une autre variable et car la sinistralité est due au comportement des conducteurs, à leurs antécédents et au type de véhicule conduit. Par conséquent, la période de paiement de la prime n'a, en théorie, pas d'incidence sur la sinistralité. Cependant, les assureurs préfèrent que le versement de la prime soit annuel afin d'avoir plus de flux financiers et pour limiter le risque de non-paiement de la prime. C'est pourquoi, il est fréquent que les assureurs majorent les assurés désirant étaler leur prime. L'usage du véhicule et la catégorie socio-professionnelle sont fortement liés. En effet, les retraités et les personnes sans profession assurent généralement leur véhicule pour un usage exclusivement privé. À l'inverse, les professions libérales, les artisans et les commerçants utilisent leur véhicule dans un but professionnel. Puisque la variable *CSP* comporte plus de modalités que la variable *usage*, et puisque d'après les analyses descriptives, l'usage du véhicule ne semble pas avoir d'impact majeur sur la sinistralité, la variable *usage* est retirée des variables explicatives. Par ailleurs, la classe du véhicule est fixée par la valeur de celui-ci et de ses pièces détachées. Le groupe est relié à la dangerosité du véhicule, et est calculé notamment grâce à la puissance fiscale, la vitesse maximale et la cylindrée. Or, ces trois éléments déterminent pour les constructeurs automobiles une grosse partie du prix de vente. Ainsi, par construction de ces deux variables, il est naturel qu'elles ne soient pas indépendantes. Cependant, elles renseignent sur des aspects différents. Par conséquent, il peut être intéressant de tout de même les conserver toutes les deux. Pour justifier ce choix, le V de Cramer a été calculé pour l'ensemble des véhicules répertoriés par le SRA, issus du véhiculier. Il vaut 0,29, ce qui correspond à une liaison modérée.

Ainsi, trois variables sont supprimées (courtier, fractionnement et usage). Par conséquent, 18 variables seront utilisées pour la modélisation du coût et de la fréquence grâce au modèle linéaire généralisé.

2.2.3 Seuil d'écrêtement des sinistres graves

En assurance automobile, comme pour d'autres types d'assurance IARD, une petite partie des sinistres peut représenter une part importante de la charge. Cependant, il n'est pas possible de modéliser simultanément les sinistres graves et les sinistres attritionnels. En effet, bien que les sinistres graves aient une fréquence très faible, leur coût est très élevé. C'est la raison pour laquelle la même méthodologie ne peut pas être appliquée pour modéliser deux types de sinistres dont le comportement est différent.

Parmi les deux garanties étudiées, seule la garantie dommages présente des coûts de sinistres relativement élevés. En effet, tandis que les coûts maximaux de la garantie bris de glace correspondent au prix de remplacement d'un pare-brise et de l'ensemble des vitres, les coûts des dommages peuvent s'élever à

plusieurs dizaines de milliers d'euros dès lors que le véhicule doit être remplacé ou que sa valeur Argus est importante.

Il est donc nécessaire de fixer un seuil à partir duquel les sinistres sont considérés comme graves. Pour cela, la théorie des valeurs extrêmes est utilisée.

Le choix du seuil d'écrêtement des sinistres graves est très important. En effet, si le seuil choisi est trop faible, des sinistres de coûts moyens seraient pris en compte dans la modélisation des très gros sinistres, ce qui engendrerait une sous-estimation des sinistres extrêmes. Si au contraire le seuil choisi est trop élevé, des sinistres d'un coût relativement élevés seraient considérés comme attritionnels, et seraient mal pris en compte lors de la modélisation, ce qui entrainerait une sous-estimation du montant des sinistres attritionnels. De plus, dans ce cas, les données seraient insuffisantes pour modéliser les sinistres extrêmes de façon robuste.

La théorie des valeurs extrêmes permet de modéliser le comportement des valeurs extrêmes. Elle a été introduite par le mathématicien Gumbel et vise à modéliser la distribution des excédents, c'est-à-dire les valeurs dépassant un seuil.

Pour une variable aléatoire Y dont la fonction de répartition est F et un seuil u fixé, on note la fonction de répartition des excès $F_u(x) = \mathbb{P}[Y - u \leqslant x | Y > u] = \frac{F(u+x)-F(u)}{1-F(u)}$. Cette valeur peut être estimée grâce au théorème suivant :

Théorème 2 Théorème de Pickands, Balkena et De Haan :

Soit F_u la fonction de répartition des excès et F la fonction de répartition d'une variable aléatoire.

Il existe k et $\sigma(u)$ tels que

$$\lim_{u \to \infty} \sup_{x > 0} \left| F_u(x) - G_{k,\sigma(u)}(x) \right| = 0$$

où $G_{k,\sigma(u)}(x)$ est la fonction de répartition d'une loi de Pareto généralisée, et définie par :

$$G_{k,\sigma(u)}(x) = \begin{cases} 1 - \left(1 + \dfrac{kx}{\sigma(u)}\right)^{-\frac{1}{k}} & \text{si } k \neq 0 \\ 1 - e^{-\frac{x}{\sigma(u)}} & \text{si } k = 0 \end{cases}$$

Ainsi, pour u suffisamment grand, la loi des excès $F_u(x)$ peut être approximé par une loi de Pareto généralisée $G_{k,\sigma(u)}(x)$. Trois paramètres doivent alors être déterminés :

- u : seuil d'écrêtement
- k : paramètre mesurant l'épaisseur de la queue de la loi de Pareto généralisée
- $\sigma(u)$: paramètre d'échelle, fonction strictement positive

Pour déterminer le seuil d'écrêtement u, il est possible d'utiliser la fonction *Moyenne des excès* (en anglais : *Mean Excess*). Cette fonction est définie par :

$$ME(u) = \mathbb{E}[Y - u | Y > u]$$

Si Y suit une loi de Pareto généralisée et si $k < 1$, alors $ME(x) = \frac{k}{1-k}u + \frac{\sigma}{1-k}$.

En traçant le *Mean Excess Plot*, c'est-à-dire la représentation graphique de $(u; ME(u))$, on trouve graphiquement que le seuil u à retenir est celui à partir duquel la fonction *moyenne des excès* est une fonction affine.

Le *Mean Excess Plot* appliqué aux coûts des sinistres de la garantie Dommages est le suivant :

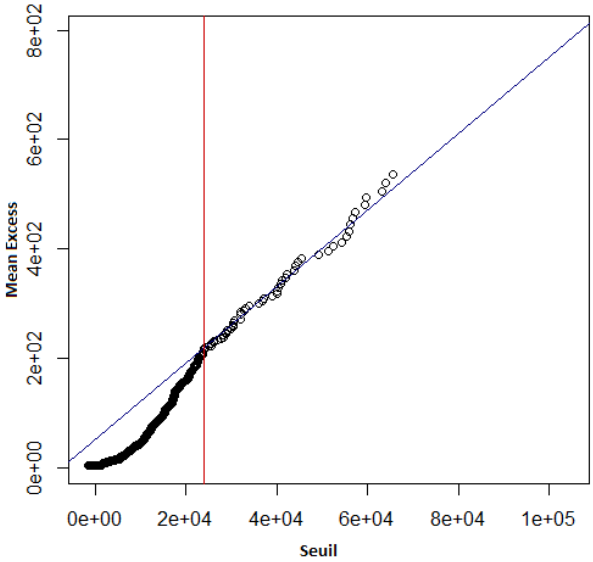

Figure 35 Mean Excess Plot des coûts de sinistres Dommages

D'après la représentation graphique, le seuil u doit être fixé à 24 000 €. Les dommages dont le coût est supérieur à ce seuil ne représentent que 0,39% de l'ensemble des sinistres dommages, soit moins d'une centaine. Ces sinistres n'étant pas assez nombreux pour faire l'objet d'une modélisation détaillée, ils sont écrêtés et ne sont pas étudiées dans la suite de ce mémoire, dont la vocation est la comparaison de différentes méthodes de modélisation : les modèles linéaires généralisés et des méthodes de Machine Learning.

Afin de tout de même proposer une tarification complète des garanties, la prime est calculée de manière non segmentée, c'est-à-dire que la charge de ces sinistres est répartie de manière homogène pour l'ensemble des assurés. Le coût moyen des sinistres graves de la base de données s'élève à 40 107 €mais leur fréquence de survenance n'est que de 0,016%, ce qui fixe la prime pure forfaitaire des sinistres graves pour la garantie dommages à 6,51 €. Néanmoins, le coût et la fréquence des sinistres graves étant très volatiles, il est indispensable de fixer des chargements de sécurité suffisamment importants pour rester solvable. Par ailleurs, quand les coûts dépassent 100 000 €, ce qui n'est pas le cas ici car il s'agit d'un portefeuille « risque aggravé », mais cela peut se produire pour les portefeuilles de voitures de collection, des traités de réassurance sont conclus avec un réassureur.

Chapitre 3 : Modèle linéaire généralisé

Ce chapitre du mémoire est consacré à la modélisation de la prime pure à l'aide des modèles linéaires généralisés. La première partie de ce chapitre expose les éléments théoriques au sujet de ces modèles (calibration du modèle, sélection des variables, validation du modèle). Puis, ces modèles sont appliqués aux données étudiées afin de modéliser la fréquence et le coût moyen, dans le but d'en déduire la prime pure. Comme le démontre l'analyse descriptive faite dans le deuxième chapitre, la prime doit être calculée de manière distincte pour les deux garanties.

3.1 Théorie sur le modèle linéaire généralisé

3.1.1 Notations

Le modèle linéaire généralisé consiste à étudier la liaison entre une variable aléatoire dite « réponse » Y et un ensemble de variables explicatives déterministes $(X_1; \dots; X_k)$. Il est très répandu en assurance dommages, et particulièrement en assurance automobile, pour déterminer les tarifs à appliquer. Ainsi, l'objectif du modèle linéaire généralisé est d'estimer la prime pure π définie par :

$$\pi(X_1; \dots; X_k) = \mathbb{E}[S|(X_1; \dots; X_k)] = \mathbb{E}[N|(X_1; \dots; X_k)] \times \mathbb{E}[C|(X_1; \dots; X_k)].$$

S désigne la charge des sinistres, N la fréquence de sinistralité et C le coût moyen des sinistres.

Le coût et la fréquence se modélisant de manière indépendante, la théorie sur le modèle est exprimée selon une variable aléatoire notée Y.

3.1.2 Modèle linéaire classique

Le modèle linéaire généralisé est une extension du modèle linéaire classique. Il convient donc de rappeler les principes de ce dernier.

Le modèle linéaire classique est utilisé pour trouver une relation linéaire entre la variable aléatoire à expliquer Y et les variables explicatives déterministes $(X_1; \dots; X_k)$. Ce modèle peut s'écrire grâce à la formule suivante :

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \varepsilon$$

avec :

- $(\beta_0; \beta_1; \dots; \beta_k)$: paramètres du modèle à estimer
- ε : résidu du modèle : variable aléatoire représentant l'écart entre la valeur observée et la valeur estimée avec $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ où σ correspond à l'écart-type des erreurs et vérifiant l'hypothèse d'homoscédasticité.

Ainsi, $Y|X$ suit une loi normale de moyenne $\beta_0 + \sum_{i=1}^{k} \beta_i X_i$ et de variance σ^2. Or, en pratique, cette hypothèse de normalité n'est pas vérifiée. Cette contrainte est allégée dans les modèles linéaires généralisés en supposant que $Y|X$ suit une loi appartenant à la famille exponentielle.

3.1.3 Les trois composantes d'un modèle linéaire généralisé

3.1.3.1 La variable réponse, composante aléatoire

La variable réponse est la variable aléatoire que l'on souhaite expliquer et dont l'espérance est recherchée. Dans le cadre de la tarification d'un produit d'assurance, il peut s'agir du montant des sinistres ou du nombre de sinistres. Afin de modéliser cette variable aléatoire, qui peut être binaire, discrète ou continue, des réalisations notées $(y_1; \dots y_n)$ sont observées.

La loi de Y doit appartenir à la famille exponentielle, c'est-à-dire que sa fonction de densité (par rapport à la mesure de comptage sur $\mathbb{N}$ ou à la mesure de Lebesgue sur $\mathbb{R}$) s'écrit sous la forme :

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

avec :

- $a(\phi)$: fonction non nulle
- $b(\theta)$: fonction de classe $\mathcal{C}^2$ et connexe
- $c(y,\phi)$: fonction
- ϕ : paramètre d'échelle
- θ : paramètre naturel car il est lié aux deux premiers moments de la loi

De plus, l'espérance et la variance de Y vérifient les relations suivantes, dont la démonstration est faite en annexe 3.D :

$$\begin{cases} \mathbb{E}[Y] = \mu = b'(\theta) \\ \mathbb{V}[Y] = \sigma^2 = b''(\theta) \times a(\phi) \end{cases}$$

Le tableau ci-dessous présente les principales lois appartenant à la famille exponentielle. Pour toutes ces lois, la fonction a est définie par $a(\phi) = \phi$. Les quatre premières lois correspondent à des lois de probabilité discrètes, les quatre suivantes sont des lois de probabilité continues.

Loi	θ	ϕ	$b(\theta)$	$c(y,\phi)$	$\mathbb{E}(Y)$	$\mathbb{V}(Y)$
Bernoulli $\mathcal{B}(1;p)$	$\ln\left(\frac{p}{1-p}\right)$	1	$\ln(1+e^\theta)$	0	p	$p(1-p)$
Binomiale $\mathcal{B}(n;p)$	$\ln\left(\frac{p}{1-p}\right)$	1	$n\ln(1+e^\theta)$	$\ln\binom{n}{y}$	np	$np(1-p)$
Binomiale négative $\mathcal{BN}(r;p)$	$\ln(p)$	1	$-r\ln(1-e^\theta)$	$\ln\binom{y+r-1}{y}$	np	$np(1-p)$
Poisson $\mathcal{P}(\lambda)$	$\ln(\lambda)$	1	$e^\theta = \lambda$	$-\ln(y!)$	$\frac{r(1-p)}{p}$	$\frac{r(1-p)}{p^2}$
Normale $\mathcal{N}(\mu;\sigma^2)$	μ	σ^2	$\frac{\theta^2}{2}$	$-\frac{1}{2}\left(\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)\right)$	μ	σ^2

Inverse gaussienne $\mathcal{IG}(\mu;\sigma^2)$	$-\frac{1}{2\mu^2}$	σ^2	$-\sqrt{-2\theta}$	$-\frac{1}{2}\left(\ln(2\pi\phi y^3)+\frac{1}{\phi y}\right)$	μ	$\frac{\mu^3}{\sigma^2}$
Exponentielle $\mathcal{E}(\lambda)$	$-\lambda$	1	$-\ln(-\theta)$	0	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma $\gamma(\alpha;\beta)$	$-\frac{\beta}{\alpha}$	$\frac{1}{\alpha}$	$-\ln(-\theta)$	$(\frac{1}{\phi}-1)\ln(y)-\ln\left(\Gamma(\frac{1}{\phi})\right)$ *	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$

Tableau 7 Exemples de lois appartenant à la famille exponentielle

Une autre loi usuellement utilisée pour modéliser une variable continue est la loi log-normale. Cependant, il est à noter que cette dernière ne fait pas partie de la famille exponentielle. De ce fait, les modèles linéaires généralisés ne peuvent pas être directement appliqués : il est nécessaire de considérer une régression normale sur le logarithme de la variable à expliquer. Cela se justifie par la définition de la loi log-normale :

$$Y \sim \mathcal{LN}(\mu;\sigma^2) \Leftrightarrow \ln(Y) \sim \mathcal{N}(\mu;\sigma^2)$$

Néanmoins, le principal défaut de cette méthode est qu'il est nécessaire de multiplier les prédictions par $e^{\frac{\sigma^2}{2}}$ afin de passer des prédictions faites à partir du modèle sur le logarithme de Y aux prédictions sur Y. En effet, $\mathbb{E}[Y] = e^{\mu+\frac{\sigma^2}{2}} \neq e^{\mu} = e^{\mathbb{E}[\ln(Y)]}$

3.1.3.2 Les variables explicatives, composante déterministe

Comme pour le modèle linéaire classique, une combinaison linéaire des variables explicatives est recherchée : $\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$.

Le choix des variables explicatives est important. En effet, il ne faut pas en prendre trop pour que le modèle soit utilisable en pratique, cependant, il est nécessaire d'en sélectionner suffisamment afin que celui-ci soit robuste et pertinent. Ainsi, il convient de sélectionner les variables apportant le plus d'informations au modèle. Plusieurs critères, détaillés dans la partie suivante peuvent être utilisés afin de sélectionner les meilleures variables.

3.1.3.3 La fonction de lien

Contrairement au modèle linéaire classique qui cherche à modéliser directement la variable à expliquer, les modèles linéaires généralisés modélisent une fonction de l'espérance de cette variable. Cette fonction, notée g, matérialise la relation entre l'espérance de la variable réponse et des variables explicatives. Elle est appelée *fonction de lien* et doit être monotone et dérivable : $g(\mathbb{E}[Y]) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$.

Les fonctions de lien les plus courantes sont les suivantes :

- Fonction identité : $g(x) = x$. Donc $\mathbb{E}[Y] = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$, ce qui correspond au modèle linéaire classique.

* $\Gamma(\alpha)$ est la fonction gamma et est définie par $\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1} exp(x)\, dx$, $\forall\, \alpha > 0$

- Fonction canonique : $g(x) = b'^{-1}(x)$. $b(x)$ est la fonction associée à la famille exponentielle. Ainsi, $g(\mathbb{E}(Y)) = b'^{-1}(b'(\theta)) = \theta = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$. Le paramètre naturel est modélisé linéairement.
- Fonction réciproque : $g(x) = -\frac{1}{x}$.
- Fonction logarithme népérien : : $g(x) = \ln(x)$. Donc $\mathbb{E}(Y) = e^{\beta_0} \times e^{\beta_1 X_1} \times \ldots \times e^{\beta_k X_k}$. Il s'agit d'un modèle multiplicatif.
- Fonction logit : $g(x) = \ln\left(\frac{x}{1-x}\right)$, ce qui correspond à une régression logistique. Ce modèle est adapté aux cas où x est compris entre 0 et 1, par exemple pour une probabilité de survenance d'un sinistre ou d'un rachat.

3.1.4 Calibration d'un modèle linéaire généralisé

L'objectif est d'estimer les paramètres $(\beta_0; \beta_1; \ldots; \beta_k)$ du modèle à partir des données observées $(x_{1,i}; \ldots; x_{k,i}; y_i)_{i=1}^{n}$. L'espérance des observations est estimée en calculant $\mathbb{E}(Y|X) = g^{-1}(\widehat{\beta_0} + \widehat{\beta_1} X_1 + \ldots + \widehat{\beta_k} X_k)$. La méthode utilisée est celle du maximum de vraisemblance.

3.1.4.1 Sélection de la loi

Tandis que les lois discrètes sont utilisées pour modéliser la fréquence, les lois continues servent à modéliser le coût des sinistres.

Afin de sélectionner la loi de Y, plusieurs méthodes peuvent être utilisées. Tout d'abord, une sélection graphique peut être effectuée. Pour cela, la distribution empirique est tracée et comparée à différentes distributions théoriques. Cette méthode graphique permet de présélectionner les lois les plus pertinentes.

Afin de confirmer le choix de la loi, il est nécessaire de calculer la log-vraisemblance des modèles et de sélectionner celui maximisant cette dernière.

3.1.4.2 Sélection des variables

Comme évoqué précédemment, il est nécessaire de sélectionner les meilleures variables, c'est-à-dire celles qui sont significatives pour le modèle.

Deux types de méthodes permettent de sélectionner les variables : les tests de *type III* et la recherche automatique du meilleur modèle.

Les tests de *type III* permettent de comparer deux modèles emboîtés. Ces modèles, notés M_1 et M_2, partagent les mêmes catégorisations et les mêmes distributions. La seule différence est que le modèle M_2 possède des variables explicatives en plus de celles de M_1. Soit k_1 et k_2 le nombre de paramètres respectivement des modèles M_1 et M_2 et soit $q = k_2 - k_1 > 0$.

Le test consiste à étudier les hypothèses :

$$\begin{cases} H_0 : \text{les paramètres supplémentaires du modèle } M_2 \text{ sont tous nuls} \\ H_1 : \text{au moins un des paramètres supplémentaires de } M_2 \text{ est non nul} \end{cases}$$

Trois tests peuvent être utilisés :

- Le test de Wald vérifie si l'écart entre les coefficients estimés sans contrainte $\hat{\beta}$ et les coefficients estimés avec contrainte $\widehat{\beta_{H_0}}$ est nul.

- Le test de vraisemblance (LRT pour Likelyhood Ratio Test en anglais) teste si la différence entre la vraisemblance au point $\hat{\beta}$ et celle au point $\hat{\beta}_{H_0}$ vaut 0.
- Le test du score vérifie que la pente de la log-vraisemblance en $\hat{\beta}$ est nulle.

Sous l'hypothèse H_0, les statistiques associées à ces trois tests suivent asymptotiquement une loi du $\mathcal{X}^2$ à q degrés de liberté. Ces trois tests sont équivalents asymptotiquement et sont détaillés en annexe 3.E. L'hypothèse H_0 est rejetée si la statistique est supérieure au fractile d'ordre $1-\alpha$.

Une autre méthode de sélection des variables consiste à rechercher automatiquement le meilleur modèle existant en se basant sur deux critères : le critère d'information d'Akaike (AIC) et le critère d'information bayésien (BIC). Ils sont définis par

$$\begin{cases} AIC = -2\mathcal{LL} + 2q \\ BIC = -2\mathcal{LL} + q\log n \end{cases}$$

avec :

- $\mathcal{LL}$: log-vraisemblance maximisée du modèle
- q : nombre de paramètres du modèle
- n : nombre d'observations

Ils permettent de pénaliser les modèles avec un grand nombre de variables afin d'avoir un compromis entre le biais (qui diminue lorsque le nombre de paramètres augmente) et la parcimonie (décrire les données avec le moins de variables explicatives possible).

Chaque combinaison de variables explicatives choisies peut constituer un modèle. Le meilleur modèle est celui dont l'AIC ou le BIC est minimal. Cependant, lorsque le nombre de variables explicatives est important, comme c'est le cas en assurance automobile, il n'est pas possible de tester toutes les combinaisons possibles. Trois méthodes permettent de sélectionner les meilleures variables, sans tester toutes les combinaisons :

- Recherche descendante (backward en anglais) : à partir d'un modèle complet, les variables sont enlevées au fur et à mesure selon un critère d'ajustement (variable la moins explicative selon un test de Student, variable dont la suppression améliore le plus les critères AIC ou BIC, etc…). Le processus s'arrête lorsque plus aucune variable ne peut être supprimée.
- Recherche ascendante (forward en anglais) : à partir d'un modèle sans variable, l'ajout d'une variable explicative est testé en utilisant un critère d'ajustement. Ce processus est répété jusqu'à ce que toutes les variables significatives aient été sélectionnées, c'est-à-dire jusqu'à ce que plus aucune variable ne puisse améliorer le modèle.
- Recherche stepwise : il s'agit d'une combinaison des recherches ascendante et descendante. À chaque étape de la recherche ascendante, les variables sélectionnées sont réévaluées grâce à une recherche descendante. Ce processus est réitéré jusqu'à ce que plus aucune variable ne puisse être ajoutée ou supprimée.

3.1.4.3 Estimation des paramètres du modèle

En partant de la formule $g(\mathbb{E}[Y|X]) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$, il est possible d'en déduire une estimation de l'espérance : $\mathbb{E}(Y|X) = g^{-1}\left(\widehat{\beta_0} + \widehat{\beta_1} X_1 + \ldots + \widehat{\beta_k} X_k\right)$ grâce à la méthode du maximum de vraisemblance. La log-vraisemblance est définie par :

$$log\ \mathcal{L} = \sum_{i=1}^{n} \left[\frac{Y_i\theta_i - b(\theta_i)}{a(\psi)} + c(Y_i, \psi) \right]$$

Il n'existe pas de solution explicite à ce problème. Néanmoins, comme l'ont démontré Schwrorer et Hovey en 2004, la solution peut être calculée par itérations successives grâce au théorème suivant :

Théorème 3 : Solution du maximum de vraisemblance (Démonstration en annexe 3.F en utilisant l'algorithme de Fisher[*]*) :*

La solution du maximum de vraisemblance se calcule par approximations successives par la formule :

$$\beta^{k+1} = \left(X'W^kX\right)^{-1}X'W^kz^k$$

avec :

- X *:matrice des données explicatives à* n *lignes et* k *colonnes*
- $W^k = \mathrm{diag}\left(\frac{\left(\left(g^{-1}\right)'\right)^2\left(X_i^T\beta^k\right)}{b''(\theta_i)}\right)$ *: matrice des poids*
- $z^k = X\beta^k + \mathrm{diag}\left(\frac{1}{(g^{-1})'\left(X_i^T\beta^k\right)}\right)\left(y - b'(\theta)\right)$

De plus, l'estimateur du maximum de vraisemblance $\hat{\beta}$ converge presque surement vers β et suit asymptotiquement une loi normale. Par conséquent, en notant $z_{1-\frac{\alpha}{2}}$ le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi normale centrée réduite, l'intervalle de confiance du coefficient $\widehat{\beta_J}$ vaut $\left[\widehat{\beta_J} - z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\mathbb{V}}[\widehat{\beta_J}]}; \widehat{\beta_J} + z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\mathbb{V}}[\widehat{\beta_J}]}\right]$

3.1.4.4 Utilisation d'une variable offset

Une variable *offset* est une variable explicative dont le coefficient β_i est connu et ne doit pas être estimé. L'offset est notamment utilisée pour modéliser le nombre de sinistres en tenant compte de l'exposition au risque. En effet, pour modéliser le nombre de sinistres, il est nécessaire de prendre en considération l'exposition, c'est-à-dire la durée durant laquelle un assuré est resté dans le portefeuille. Un assuré ayant eu un sinistre en 6 mois est statistiquement deux fois plus risqué qu'un assuré ayant eu un sinistre pendant une année complète. Ainsi, il convient de diviser le nombre de sinistres par l'exposition. En notant d l'exposition, le modèle devient : $g\left(\mathbb{E}\left[\frac{Y}{d}|X\right]\right) = \beta_0 + \beta_1X_1 + \ldots + \beta_kX_k$. En considérant une fonction de lien logarithmique, la régression se réécrit : $\ln(\mathbb{E}[Y|X]) = \beta_0 + \beta_1X_1 + \ldots + \beta_kX_k + \ln(w)$. Par conséquent l'offset joue un rôle de variable explicative pour laquelle le coefficient vaut 1. En pratique, l'offset est utilisée afin de normaliser le nombre de sinistres par rapport à une durée définie.

3.1.5 Validation du modèle

Après avoir modélisé la variable réponse Y, il est nécessaire de vérifier la qualité de l'ajustement des estimations.

Pour cela, la déviance est calculée. Cette méthode consiste à comparer le modèle estimé à un modèle « saturé », c'est-à-dire un modèle linéaire généralisé de même loi et de même fonction de lien mais avec

[*] Il est également possible d'approximer les coefficients grâce à l'algorithme de Newton-Raphson. Cependant, les travaux de A. Schworer et P. Hovey ont démontré qu'utiliser la méthode de Fisher permet d'obtenir des résultats plus robustes et dont la convergence est assurée.

autant de paramètres que d'observations. Le modèle saturé estime donc exactement les variables. Par définition, la déviance est définie par le double de la différence entre les log-vraisemblances de ces deux modèles : $D = 2(\mathcal{LL}_{saturé} - \mathcal{LL}_{estimé})$. Elle est positive et est petite pour un modèle de bonne qualité. De plus, D suit asymptotiquement une loi du $\mathcal{X}^2$ à $n - k$ degrés de liberté, ce qui permet de construire un test de rejet ou d'acceptation du modèle.

La déviance est un indicateur global pouvant être complété par une analyse plus fine observation par observation : l'analyse des résidus. Les résidus sont calculés pour chaque donnée i et correspondent aux écarts entre les valeurs à expliquer et leurs estimations. Les résidus bruts $y_i - \hat{y}_i$ ne sont pas facilement interprétables car leur valeur doit être comparée aux valeurs attendues y_i et car leur variance varie selon les modèles utilisés. Par conséquent, des résidus standardisés sont utilisés. Les deux principaux sont les résidus de Pearson ${r_i}^P$ et les résidus de déviance ${r_i}^D$, définis par :

$$\begin{cases} {r_i}^P = \dfrac{y_i - \hat{y}_i}{\sqrt{\widehat{\mathbb{V}}[\hat{y}_i]}} \\ {r_i}^D = signe(y_i - \hat{y}_i)\sqrt{D_i} \end{cases}$$

avec :

- D_i : contribution de la i^ème^ donnée à la déviance : $D_i = 2\left(\mathcal{LL}_{i,saturé} - \mathcal{LL}_{i,estimé}\right)$
- $\widehat{\mathbb{V}}[\hat{y}_i]$: variance estimée du modèle pour la i^ème^ donnée

Un autre type de résidus introduit par Peter K. Dunn et Gordon K. Smyth peut être utilisé : les résidus quantiles normalisés. Ils sont définis par : ${r_i}^Q = F^{-1}_{\mathcal{N}(0;1)}\left(\hat{F}(y_i)\right)$ où $F^{-1}_{\mathcal{N}(0;1)}$ désigne la fonction quantile* de la loi normale centrée réduite et $\hat{F}$ la fonction de répartition estimée de la variable réponse. Leur définition vient du fait que si le modèle est correct, les $\hat{F}(y_i)$ sont uniformément distribués sur l'intervalle unité. Par conséquent, il est possible d'appliquer le théorème d'inversion** et d'en déduire que les résidus quantiles normalisés suivent une loi normale centrée réduite.

L'analyse de ces différents résidus est généralement effectuée visuellement, grâce à des graphiques. La représentation graphique des résidus en fonction des prédictions associés à ces résidus ne doit pas faire ressortir de tendance particulière : les résidus doivent être indépendants. Sinon, cela signifie qu'un effet explicatif des variables n'est pas capté par le modèle. De plus, si le modèle est correct, les résidus suivent une loi normale centrée réduite. Il est donc possible de comparer la distribution empirique des résidus à la distribution théorique d'une loi normale centrée réduite.

Il est possible de définir des seuils de valeurs à partir desquels les résidus sont anormalement élevés (positivement ou négativement). Pour cela, le seuil z_α est défini par : $\mathbb{P}(Z \notin [-z_\alpha; z_\alpha]) = \alpha$ avec $Z \sim \mathcal{N}(0;1)$. Ainsi, chaque résidu a une probabilité α d'être en dehors de l'intervalle $[-z_\alpha; z_\alpha]$ et la probabilité qu'au moins un résidu sur les n se trouve hors de $\left[-z_{\frac{\alpha}{n}}; z_{\frac{\alpha}{n}}\right]$ est inférieure ou égale à α.

---

* La fonction quantile, également appelée fonction inverse généralisée de la fonction de répartition F_X d'une variable aléatoire X est définie par $F_X^{-1}(p) = inf\{t \in \mathbb{R} : \ F_X(t) \geq p\}$.

** Théorème d'inversion : Soit une variable aléatoire U suivant une loi uniforme sur l'intervalle $[0;1]$ et soit F_X une fonction de répartition dont la fonction quantile est F_X^{-1}. Alors $F_X^{-1}(U)$ a pour fonction de répartition F_X.

3.2 Modélisation des garanties

Cette partie du mémoire consiste à appliquer des modèles linéaires généralisés aux données de sinistralité des garanties bris de glace et dommages tous accidents.

Le coût moyen et la fréquence de ces deux garanties sont indépendants, par conséquent, il est possible d'utiliser la méthode « coût × fréquence » présentée dans le premier chapitre.

3.2.1 Modélisation pour la garantie bris de glace

3.2.1.1 Modélisation de la fréquence

Choix de la loi :

Afin de sélectionner la loi utilisée pour modéliser la fréquence, une première approche graphique peut être envisagée. Pour cela, la distribution du nombre de sinistres est tracée, puis, elle est comparée à la distribution théorique de deux lois dont les paramètres sont les paramètres naturels associés à ces lois.

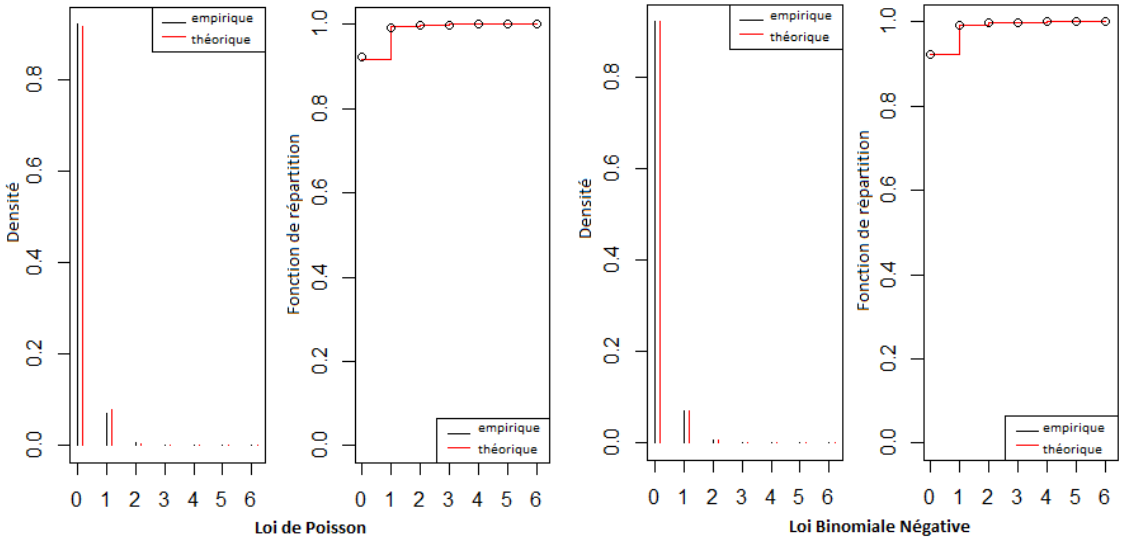

Figure 36 : Comparaison de la distribution empirique du nombre de sinistres Bris de glace avec deux lois théoriques : Poisson et Binomiale négative

D'après l'analyse élémentaire des graphiques, il ressort que la loi binomiale négative est celle permettant une meilleure estimation. En effet, la loi de Poisson surestime le nombre de sinistres.

Pour compléter cette analyse, il est possible d'étudier l'adéquation des modèles aux différentes lois. La meilleure correspond à celle dont la déviance et le critères AIC sont minimaux.

Loi	Déviance	AIC
Poisson	40 545	68 528
Binomiale négative	34 899	60 023

Tableau 8 Comparaison des critères d'ajustement pour le modèle de fréquence de la garantie bris de glace

Ces résultats permettent de confirmer que la loi la plus adaptée à la modélisation de la fréquence des sinistres bris de glace est la loi binomiale négative.

Sélection des variables :

Après avoir sélectionné la loi permettant la meilleure modélisation de la variable réponse (le nombre de sinistres) il est nécessaire de sélectionner les variables explicatives. Pour cela, un test de Wald est réalisé. Les résultats sont exposés dans le tableau ci-dessous. La variable est significative si la p-value est inférieur à 0,05. Les variables sélectionnées pour être explicatives dans le modèle linéaire généralisé sont présentées en vert, celles qui ne sont pas significatives en rouge :

- p-value(Classe SRA) = 0
- p-value(Age du conducteur) 0
- p-value(Nombre d'années de permis) = 0
- p-value(Annulation de permis) = $8{,}15 \times 10^{-9}$
- p-value(CRM) = $1{,}46 \times 10^{-7}$
- p-value(Zone) = $2{,}53 \times 10^{-6}$
- p-value(Age du véhicule) = $4{,}73 \times 10^{-6}$
- p-value(Résiliation non-paiement) = $2{,}19 \times 10^{-5}$
- p-value(Nombre de sinistres matériels non responsables) =$6{,}25 \times 10^{-5}$
- p-value(CSP) = $1{,}73 \times 10^{-4}$
- p-value(Groupe SRA) = $9{,}63 \times 10^{-4}$
- p-value(Condamnation pour alcoolémie) = $1{,}06 \times 10^{-3}$
- p-value(Nombre de sinistres matériels responsables) = $3{,}23 \times 10^{-3}$
- p-value(Résiliation fréquence de sinistres) = $9{,}51 \times 10^{-3}$
- p-value(Nombre de sinistres corporels) = $8{,}25 \times 10^{-1}$
- p-value(Suspension de permis) = $9{,}42 \times 10^{-1}$
- p-value(Résiliation fausse déclaration) = $9{,}55 \times 10^{-1}$

D'après cette analyse, trois variables doivent être retirées à cause de leur manque de significativité : une résiliation à la suite d'une fausse déclaration, une suspension de permis ou le nombre de sinistres corporels.

Une sélection stepwise selon le critère AIC a également été réalisée et permet d'obtenir les mêmes conclusions, permettant ainsi de diminuer l'AIC de 60 023 à 56 478.

Coefficients du modèle :

La fonction de lien utilisée est la fonction logarithmique. La constante β_0 est égale à $-2{,}4522$. Pour déterminer la fréquence de référence, il est nécessaire de passer sous forme exponentielle cette constante β_0. La fréquence de référence ainsi obtenue vaut $e^{-2{,}4522} = 0{,}086$, ce qui correspond à la fréquence mesurée grâce à l'analyse descriptive. Cela permet ainsi de vérifier la cohérence des résultats.

Les graphiques suivants présentent une partie des coefficients obtenus à l'aide du modèle linéaire généralisé, ainsi que leur intervalle de confiance à 95%.

Certains regroupements de modalités ont été effectués. Ceux-ci sont réalisés afin de mieux calibrer le modèle et sont effectués lorsque deux coefficients sont très similaires. Cela permet à la fois de réduire le nombre de coefficients et d'améliorer le modèle en diminuant la variance car le calcul des coefficients est effectué sur un plus grand nombre de données.

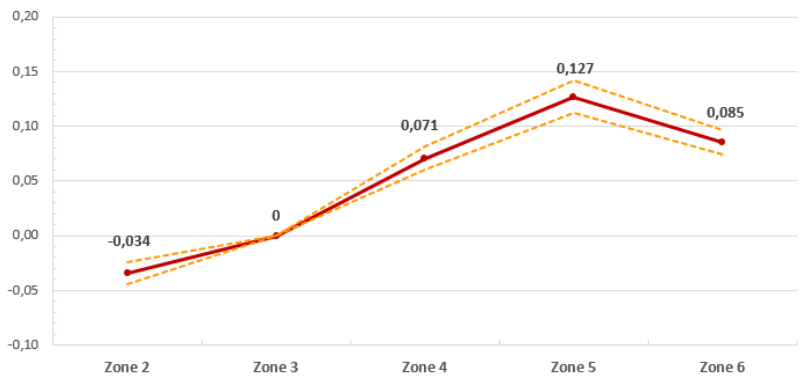

Figure 37 Coefficients estimés par le modèle linéaire généralisé pour la variable Zone

La zone 3 est la zone servant de référence car il s'agit de la zone d'habitation d'une grosse proportion des assurés. Pour trouver les minorations ou les majorations de fréquence de sinistralité selon les différentes variables, il est nécessaire de passer les coefficients sous forme exponentielle. Ainsi, habiter dans une zone peu risquée (zone 2) diminue la fréquence de sinistralité de 3,4% tandis qu'habiter en zone 5 augmente la fréquence de 13,5%.

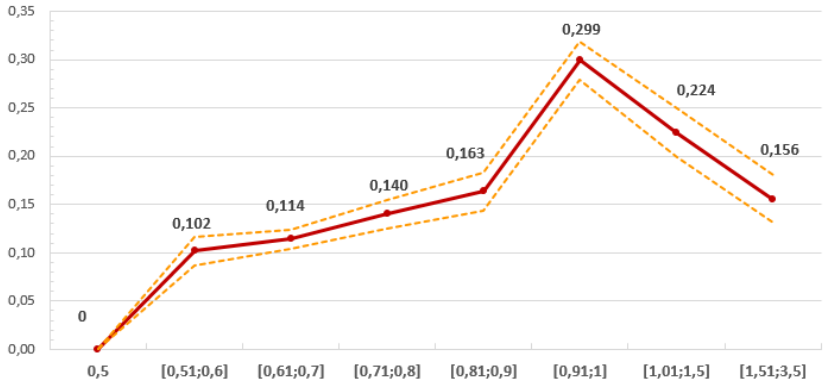

Figure 38 Coefficients estimés par le modèle linéaire généralisé pour la variable CRM

Une particularité du portefeuille étudié et des assurés en situation d'aggravation est le fait que très peu d'entre eux ont un bonus maximal. Le graphique ci-dessus présente les coefficients associés à chaque groupe de CRM. Les coefficients correspondent à la tendance déterminée grâce à l'analyse descriptive faite dans le chapitre 2. Il est à noter que le fait d'avoir un CRM compris entre 0,9 et 1 vient augmenter la fréquence de 35% par rapport à celle servant de référence.

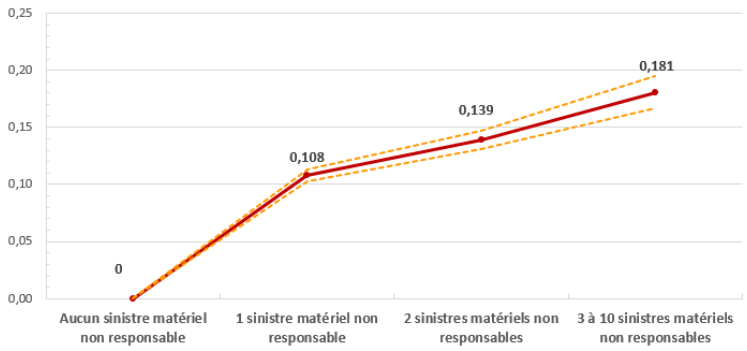

Figure 39 Coefficients estimés par le modèle linéaire généralisé pour la variable Antécédant de sinistres non responsables

Comme l'ont démontré les statistiques descriptives, les antécédents de sinistralité non-responsable ont une forte influence sur la fréquence des sinistres bris de glace (qui est une garantie associée à des sinistres non-responsables). Ainsi, plus le nombre de sinistres précédent est élevé, plus le coefficient augmente.

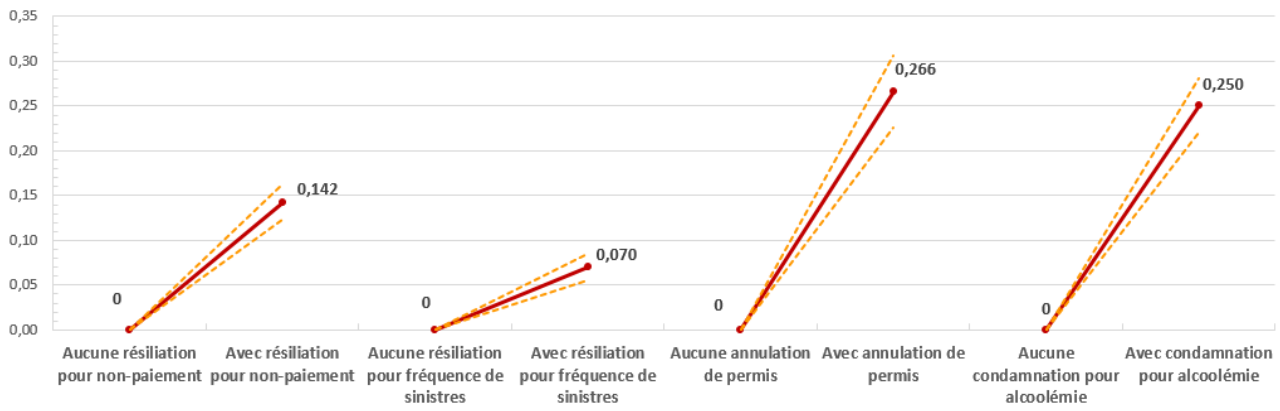

Figure 40 Coefficients estimés par le modèle linéaire généralisé pour les variables résiliations et condamnations

Une autre particularité d'un portefeuille de risque aggravé est le cadre de souscription des personnes ayant fait l'objet d'une résiliation ou d'une condamnation liée à la conduite. D'après les coefficients issus du modèle, le fait d'avoir eu une résiliation pour non-paiement de prime ou à la suite d'un nombre de sinistres trop important, une annulation de permis ou une condamnation pour alcoolémie au volant augmente fortement la fréquence de sinistralité, atteignant 30% d'augmentation pour les conducteurs ayant eu une annulation de permis.

Analyse des résidus :

Les tendances des coefficients issus du modèle linéaire généralisé sont en adéquation avec les observations élémentaires faites grâce à l'analyse descriptive. Cependant, pour valider le modèle, il est nécessaire d'analyser les résidus.

Tout d'abord, les résidus de Pearson sont calculés. Ils sont regroupés en paquets (*crunched residuals* en anglais). Ces regroupements sont effectués à partir des valeurs prédites triées par ordre croissant, de manière à être homogènes en termes de risque prédit. Pour vérifier leur normalité, la densité et la fonction de répartition des résidus de Pearson sont tracées et comparées avec celles d'une loi normale centrée réduite.

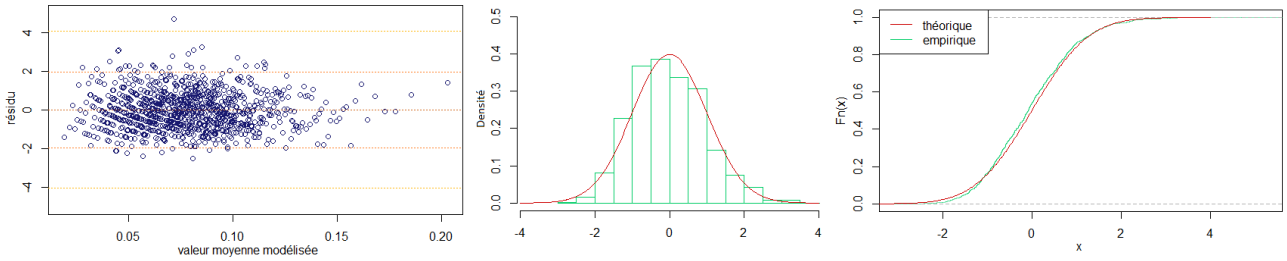

Figure 41 : Résidus groupés de Pearson du modèle de fréquence de la garantie Bris de glace en fonction de la valeur moyenne modélisé) et comparaison de leur densité et leur fonction de répartition empirique à la loi normale centrée réduite

La plupart des valeurs des résidus de Pearson sont comprises entre -1,96 et 1,96 (quantile à 95% de la loi normale centrée réduite) et une seule dépasse 4,05 (seuil à 95% pour les 1000 classes de risque). De plus, la moyenne des résidus vaut $-0{,}03$ ce qui prouve qu'ils sont centrés. Afin de confirmer la normalité, un test de Shapiro-Wilk, dont la p-valeur vaut 0,327, est effectué. Aucune sur-dispersion des résidus ne se distingue.L'analyse de la distribution des résidus de Pearson permet donc de confirmer la normalité et l'homoscédasticité de ces derniers et la validité du modèle linéaire généralisé pour la fréquence de la garantie Bris de glace.

Afin de compléter l'analyse de la validation du modèle, les résidus quantiles normalisés sont calculés et représentés graphiquement. En raison du nombre d'observations très important, des couleurs sont utilisées afin de déterminer la densité des points : la zone rouge correspond à la zone dans laquelle la plupart des résidus sont situés, la zone bleue celle où il y en a le moins.

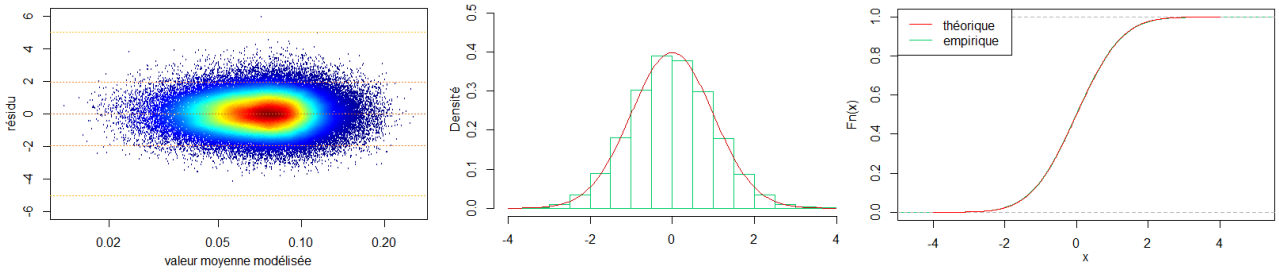

Figure 42 : Résidus quantiles normalisés du modèle de fréquence de la garantie Bris de glace en fonction de la valeur moyenne modélisée (échelle logarithmique) et comparaison de leur densité et leur fonction de répartition empirique à la loi normale centrée réduite

Aucune tendance sur les résidus n'apparait sur les graphiques. Par ailleurs, la distribution des résidus quantiles randomisés normalisés semble approximer parfaitement celle d'une loi normale centrée réduite. Cela permet de conclure que la modélisation de la fréquence par un modèle linéaire généralisé pour la garantie bris de glace est correcte.

3.2.1.2 Modélisation du coût

Après avoir modélisé la fréquence de sinistralité, c'est-à-dire la probabilité d'avoir un ou plusieurs sinistres au cours d'une année, il est nécessaire de modéliser le coût moyen de ces sinistres.

Choix de la loi :

Avant de commencer à modéliser le coût, il est nécessaire de sélectionner la loi permettant de modéliser la variable à expliquer.

Une analyse graphique de la distribution des sinistres permet d'avoir un premier aperçu de la distribution du coût des sinistres et de la loi la plus adaptée à leur modélisation. Pour cela, la distribution empirique des coûts de sinistres est comparée avec la distribution théorique de deux lois : la loi Gamma et la loi Log-normale.

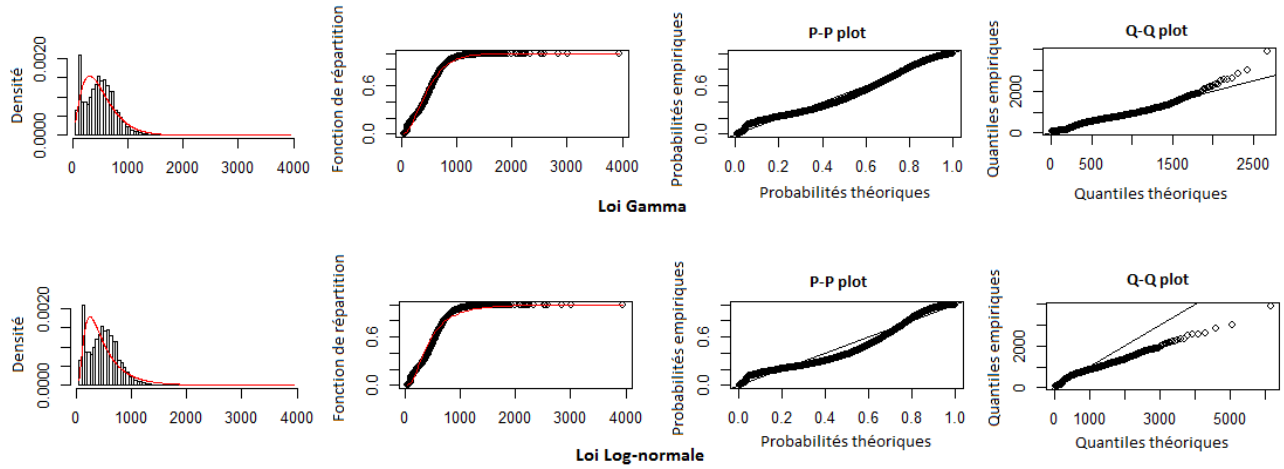

Figure 43 : Comparaison de la distribution empirique du coût de sinistres Bris de glace avec deux lois théoriques : Gamma et Log-normale

Aucune de ces deux lois ne s'adapte réellement à la distribution des données. En effet, un pic est compris entre 100 et 200 €. Ce pic peut correspondre par exemple à des remplacements de vitres latérales. Cependant, aucune information spécifiant la pièce indemnisée (pare-brise, vitres latérales, lunette arrière), n'est disponible dans les bases. D'après le P-P plot et le Q-Q plot, la loi Gamma semble néanmoins plus adaptée à la modélisation du coût des sinistres. Pour confirmer cela, la déviance et le critères AIC sont calculés.

Loi	Déviance	AIC
Gamma	3 199	12 6293
Log-normale	3 893	13 6847

Tableau 9 Comparaison des critères d'ajustement pour le modèle de coût de la garantie bris de glace

Les deux critères sont minimaux pour la loi Gamma, ce qui permet de confirmer qu'il s'agit de la loi la plus adaptée.

Sélection des variables :

Les statistiques descriptives ont permis d'émettre l'hypothèse que le coût moyen des sinistres bris de glace varie très peu en fonction de la plupart des variables. En effet, l'analyse élémentaire réalisée dans le deuxième chapitre montre que seules les variables liées au véhicule ont une influence importante sur le coût des sinistres bris de glace. Cette hypothèse semble être rationnelle puisque le coût de remplacement d'un pare-brise ou des autres vitres dépend du modèle de la voiture et non des caractéristiques du conducteur.

La significativité des variables a été testée grâce au test de Wald. Seules sept variables sont significatives pour la modélisation du coût :

- La classe SRA (p-value = 0)
- L'âge du véhicule (p-value = 0)
- La résiliation pour fréquence de sinistralité trop élevée (p-value = $2{,}78 \times 10^{-3}$)
- Le groupe SRA (p-value = $4{,}25 \times 10^{-3}$)
- La catégorie socio-professionnelle (p-value = $8{,}84 \times 10^{-3}$)
- La résiliation pour non-paiement (p-value = $1{,}09 \times 10^{-2}$)
- Le nombre de sinistres matériels non responsables (p-value = $1,38 \times 10^{-2}$)

La variable expliquant le mieux le coût des sinistres est la classe SRA. En effet, par définition, cette variable est fixée selon le prix des pièces de remplacement.

Coefficients du modèle :

Le coefficients β_0 du modèle linéaire généralisé vaut $6{,}1553$, ce qui fixe la prime de base à 471€ qui est le montant correspondant exactement au coût médian des sinistres bris de glace.

Les coefficients du modèle sont présentés pour la variable la plus significative : la classe SRA pour laquelle des regroupements de modalités ont été effectués.

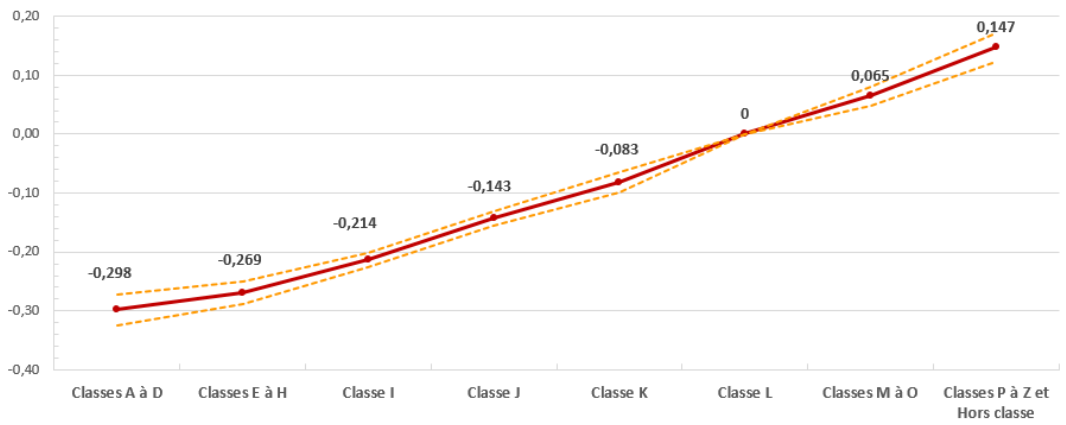

Figure 44 Coefficients estimés par le modèle linéaire généralisé pour la variable classe SRA

Les coefficients déterminés grâce au modèle linéaire généralisé augmentent linéairement selon la classe du véhicule. Les voitures dont la classe est petite ont un coût moyen 24% inférieur au coût de référence. À l'inverse, les voitures haut de gamme ont un coût moyen de bris de glace supérieur de 16%.

Analyse des résidus :

Afin de vérifier que le modèle linéaire généralisé modélise correctement les coûts des sinistres bris de glace, il est nécessaire d'analyser les résidus afin de comparer les valeurs prédites aux valeurs observées.

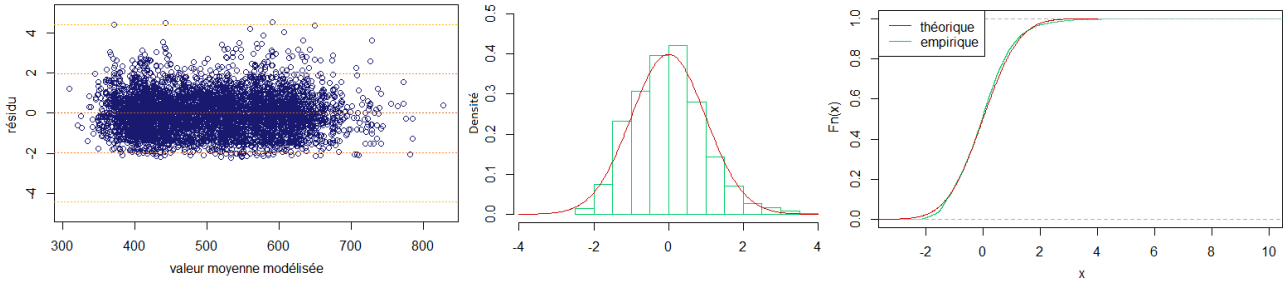

Figure 45 : Résidus de Pearson du modèle de coût de la garantie Bris de glace en fonction de la valeur modélisée et comparaison de leur densité et leur fonction de répartition empirique à la loi normale centrée réduite

La moyenne des résidus de Pearson vaut 0,004 ce qui prouve qu'ils sont centrés. Par ailleurs, la plupart d'entre eux sont bien compris dans l'intervalle [-1,96 ;1,96]. La p-valeur du test de Shapiro-Wilk étant égale à 0,061, la normalité des résidus est vérifiée. Par ailleurs, aucune tendance ne se dégage de l'analyse graphique des résidus, ce qui permet de vérifier l'hypothèse d'homoscédasticité.

3.2.2 Modélisation pour la garantie dommages tous accidents

La modélisation des garanties bris de glace et dommages tous accidents étant similaires, les résultats de la modélisation de la fréquence et du coût sont présentés plus succinctement pour la garantie dommages.

3.2.2.1 Modélisation de la fréquence

Choix de la loi :

Le choix de la loi permettant de modéliser le nombre de sinistres se fait de la même manière que pour la garantie bris de glace.

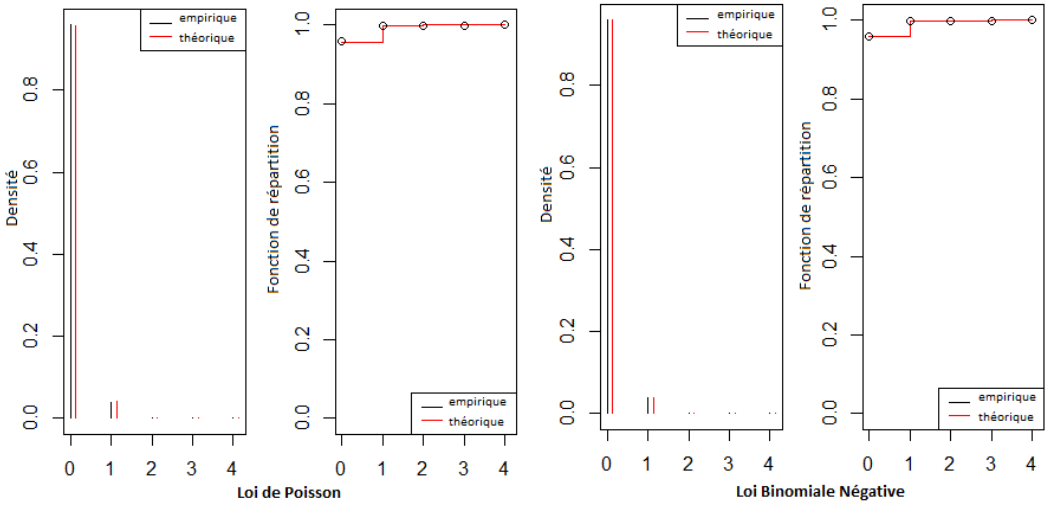

Figure 46 Comparaison de la distribution empirique du nombre de sinistres Dommages tous accidents avec deux lois théoriques : Poisson et Binomiale négative

Comme pour la modélisation de la fréquence des sinistres bris de glace, la loi de Poisson surestime le nombre de sinistres.

D'après la comparaison entre la densité et la fonction de répartition empiriques avec la densité et la fonction de répartition théorique d'une loi binomiale négative, cette dernière semble s'adapter à la répartition du nombre de sinistres.

Cette analyse graphique est complétée par la comparaison des valeurs de la déviance et du critère AIC.

Loi	Déviance	AIC
Poisson	14 500	19 995
Binomiale négative	10 899	17 305

Tableau 10 Comparaison des critères d'ajustement pour le modèle de fréquence de la garantie dommages tous accidents

Par conséquent, la loi permettant de mieux modéliser la fréquence de sinistralité de la garantie dommages est la loi binomiale négative.

Sélection des variables :

La sélection des variables a été réalisée grâce au test de Wald, dont les résultats ont été confirmés par une sélection stepwise selon le critère AIC. Les variables significatives pour modéliser la fréquence des sinistres dommages sont :

- Coefficient de réduction-majoration
- Classe SRA du véhicule
- Âge du conducteur
- Catégorie socio-professionnelle
- Ancienneté du permis de conduire
- Nombre de sinistres matériels non responsables
- Groupe SRA du véhicule
- Âge du véhicule
- Résiliation pour non-paiement de prime
- Nombre de sinistres matériels responsables
- Zone
- Condamnation pour alcoolémie au volant
- Suspension de permis
- Annulation de permis

Ces quatorze variables permettent de baisser le critère AIC à 16 617.

Coefficients du modèle :

Comme pour la garantie bris de glace, la fonction de lien choisie est la fonction logarithmique.

Le coefficient β_0 issu du modèle linéaire généralisé vaut $-3{,}179$, ce qui signifie que la fréquence de base est égale à 4,16%, ce qui est cohérent avec l'analyse faite dans le deuxième chapitre.

Les coefficients des variables définissant une aggravation du risque sont présentés sur la figure page suivante.

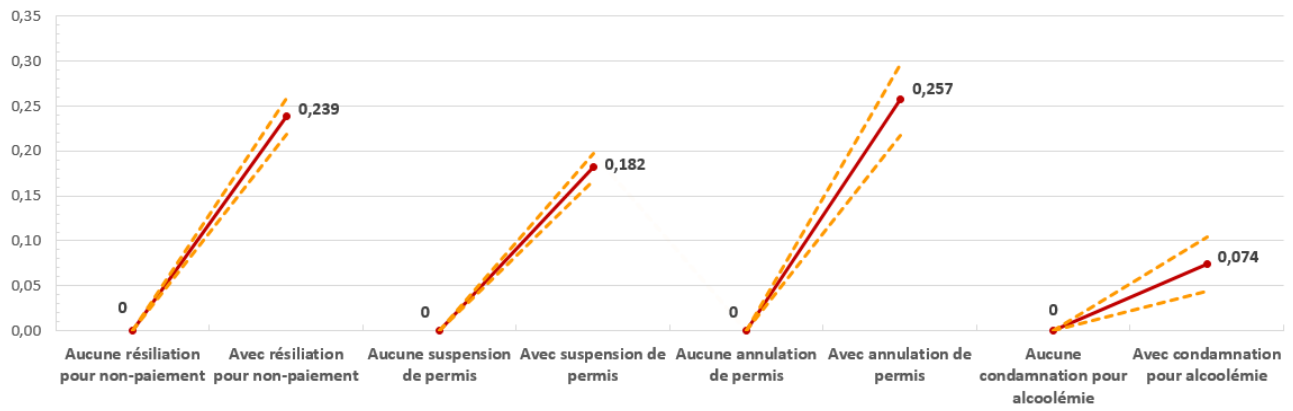

Figure 47 : Coefficients estimés par le modèle linéaire généralisé pour les variables Résiliations et condamnations

Les variables caractérisant un risque aggravé engendrent une hausse de la fréquence de dommages. Par exemple, un conducteur ayant eu une annulation de permis a un coefficient majorant de 29,3%

Analyse des résidus :

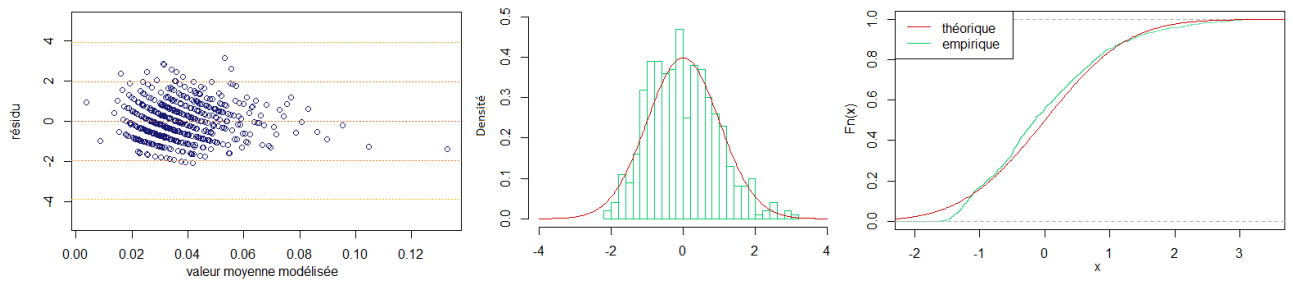

Figure 48 Résidus groupés de Pearson du modèle de fréquence de la garantie Dommages tous accidents en fonction de la valeur moyenne modélisée et comparaison de leur densité et leur fonction de répartition empirique à la loi normale centrée réduite

Les résidus de Pearson sont répartis de manière linéaire. Cependant, cela est dû à la construction du graphe et le modèle peut être jugé comme correct lorsque la dispersion des points est symétrique autour de 0, ce qui est le cas. La moyenne des valeurs des résidus de Pearson vaut-0,008 donc ces derniers sont centrés. Cependant, la densité et la fonction de répartition empiriques semblent présenter quelques problèmes d'ajustement à une loi normale centrée réduite. Pour confirmer ce constat, un test de Shapiro-Wild est effectué. La p-valeur vaut 0,02, ce qui est inférieur au seul 5%, donc la répartition gaussienne des résidus de Pearson n'est pas vérifiée.

Néanmoins, malgré quelques écarts empêchant la validation du test, les résidus de Pearson ne semblent pas présenter d'anomalies particulières. Pour cette raison, la validation du modèle est complétée par l'étude des résidus quantiles normalisés.

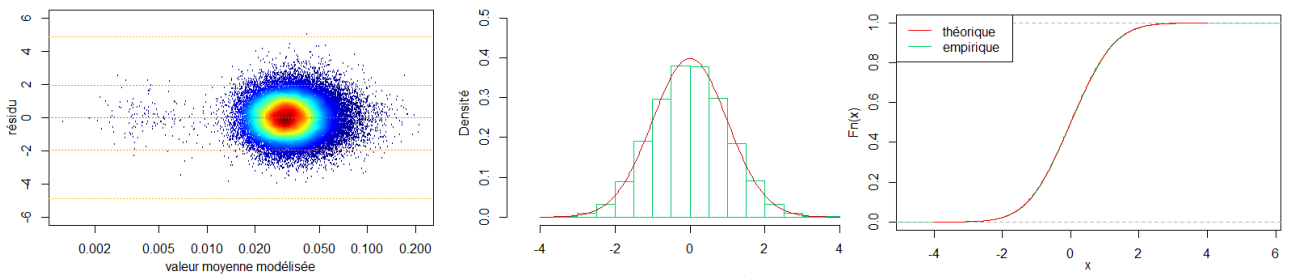

Figure 49 Résidus quantiles normalisés du modèle de fréquence de la garantie Bris de glace en fonction de la valeur moyenne modélisée (échelle logarithmique) et comparaison de leur densité et leur fonction de répartition empirique à la loi normale centrée réduite

La plupart des résidus quantiles normalisés sont centrés autour de 0 avec des valeurs comprises entre -1,96 et 1,96. Leur moyenne vaut 0,003. Par ailleurs, un test de Shapiro-Wilk (dont la p-valeur vaut 0,58) est effectué afin de confirmer l'adéquation à la loi normale centrée réduite.

Cela confirme que le modèle linéaire généralisé permet de modéliser correctement la fréquence des dommages.

3.2.2.2 Modélisation du coût

Choix de la loi :

Comme pour la modélisation du coût des sinistres bris de glace, la loi log-normale sous-estime le coût des sinistres et la loi la plus adaptée est la loi Gamma.

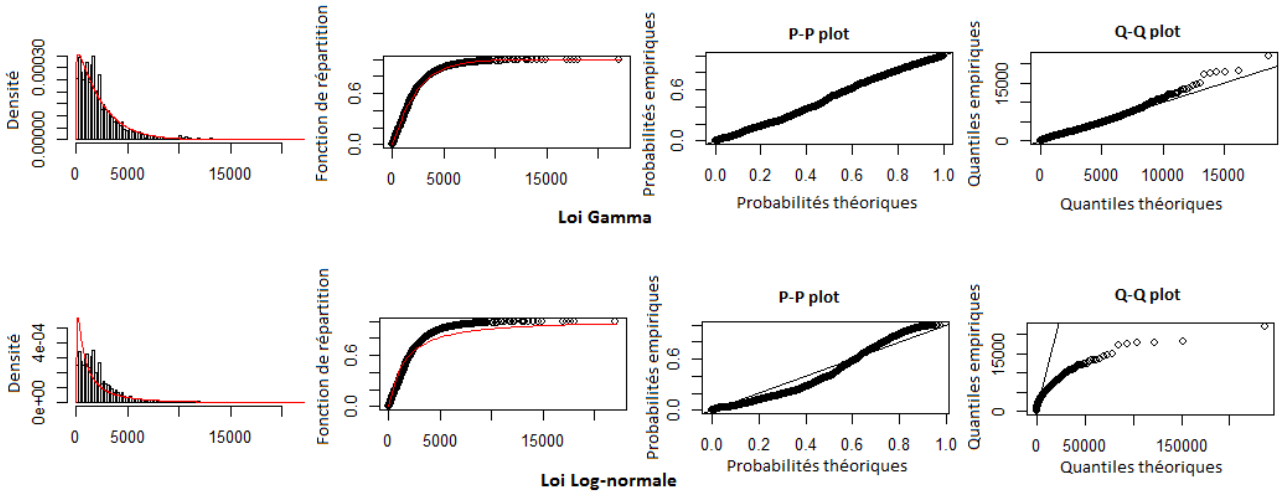

Figure 50 : Comparaison de la distribution empirique du coût de sinistres Dommages tous accidents avec deux lois théoriques : Gamma et Log-normale

De plus, les valeurs de la déviance et du critère AIC sont minimaux pour la loi Gamma.

Loi	Déviance	AIC
Gamma	2 131	39 518
Log-normale	4 571	44 777

Sélection des variables :

Un test de Wald et une analyse stepwise ont été effectué afin de conserver les variables significatives à la modélisation du coût des dommages. Seules sept variables sont significatives : la classe SRA, le groupe SRA, l'âge du conducteur, l'âge du véhicule, le coefficient de réduction-majoration, la zone d'habitation et le fait d'avoir eu une annulation de permis. Il est intéressant de noter que seules trois variables liées au risque aggravé sont significatives : l'âge du conducteur, son CRM et une annulation de permis. Il est également normal que les variables concernant le véhicule soient significatives puisqu'elles ont une influence sur les coûts de réparation.

Coefficients du modèle :

Le modèle linéaire généralisé permet de fixer un coût référent à 2 246 €, ce qui correspond approximativement au coût moyen des sinistres dommages. Certaines caractéristiques spécifiques au

risque aggravé engendrent une augmentation du coût des sinistres. Ainsi, les résultats du modèle estiment que le coût augmente de 39% pour les jeunes conducteurs, de 25% pour les conducteurs ayant un malus modéré et de 60% ceux dont le malus est supérieur à 1,5. De plus, le coût moyen des sinistres des conducteurs ayant eu une annulation de permis est supérieur de 30% à celui de base.

Analyse des résidus :

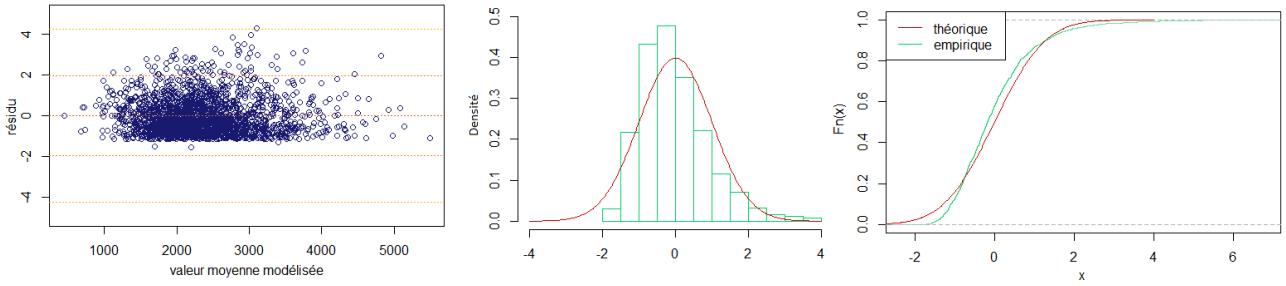

Figure 51 Résidus de Pearson du modèle de coût de la garantie Dommages tous accidents en fonction de la valeur modélisée et comparaison de leur densité et leur fonction de répartition empirique à la loi normale centrée réduite

Malgré le fait que la moyenne des résidus de Pearson vaut 0,0007, le modèle linéaire généralisé semble sous-estimer certains coûts car certaines valeurs sont supérieures à 1,96. Néanmoins, 95,6% des résidus sont situés dans l'intervalle de confiance à 95% de la loi normale.

Afin de compléter l'analyse, les résidus quantiles normalisés sont étudiés. Leur moyenne vaut -0,01 et le test de Shapiro-Wilk permet de confirmer leur normalité.

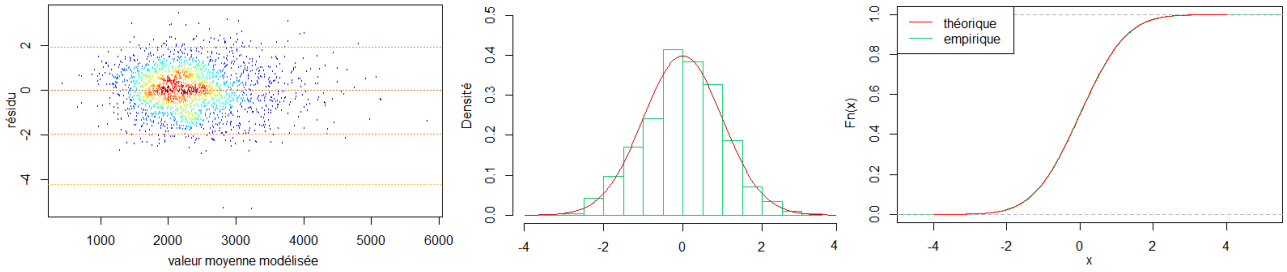

Figure 52 : Résidus quantiles normalisés du modèle de fréquence de la garantie Bris de glace en fonction de la valeur moyenne modélisée et comparaison de leur densité et leur fonction de répartition empirique à la loi normale centrée réduite

Cette analyse des résidus permet de conclure que le modèle linéaire généralisé permet de modéliser correctement le coût des sinistres dommages au véhicule.

Les différentes analyses effectuées permettent de contrôler que le modèle linéaire généralisé modélise correctement la fréquence et le coût des sinistres bris de glace et dommages au véhicule. En effet, les tendances constatées lors de la description de la sinistralité se retrouvent dans les coefficients des modèles. De plus, l'analyse des résidus permet de valider les différents modèles.

Cependant, les modèles linéaires généralisés présentent certaines limites. Leur principal défaut vient du fait qu'ils sont paramétriques et qu'une loi décrivant la variable Y doit être spécifiée. De plus, le choix des variables significatives peut s'avérer difficile. Face à cette limite, d'autres méthodes sont testées dans le quatrième chapitre afin d'essayer d'améliorer la modélisation de la fréquence et du coût des deux garanties étudiées.

Chapitre 4 : Modèles de Machine Learning

4.1 Théorie

L'apprentissage statistique (*Machine Learning* en anglais) est une branche de l'intelligence artificielle consistant à utiliser des ordinateurs pour prendre des décisions, sans que ceux-ci ne soient explicitement programmés, c'est-à-dire sans programmation impérative de type « if…, then…, else… ». Deux grands types d'algorithmes se distinguent :

- L'apprentissage supervisé : les données à prédire sont « étiquetées ». L'objectif consiste donc à trouver un lien entre la variable à expliquer et les variables explicatives. Lorsque la variable à prédire est quantitative, l'algorithme effectue une régression, lorsqu'elle est qualitative, l'algorithme procède à une classification.
- L'apprentissage non supervisé : la variable à expliquer n'est pas présente. Le but est d'étudier la structure des observations afin de les regrouper en classes homogènes.

Dans le cadre de la tarification en assurance automobile, les méthodes utilisées sont supervisées : la variable à expliquer est le nombre de sinistres ou le coût d'un sinistre. Il s'agit de variables quantitatives connues.

Les notations utilisées dans ce chapitre sont les mêmes que celles introduites précédemment.

4.1.1 Arbre de régression CART

L'algorithme CART (*Classification and regression trees* en anglais, *arbres de décision pour la classification et la régression* en français) a été développé en 1984 par Breiman, Friedman, Olshen et Stone. Il s'agit d'une méthode non paramétrique d'apprentissage statistique.

En théorie des graphes, un arbre est un graphe connexe (il existe une chaîne entre deux sommets quelconques), acyclique (ne contenant aucun cycle) et possédant une unique racine (sommet à partir duquel il existe un chemin unique vers tous les autres sommets du graphe).

Un arbre représente une hiérarchie, dont chaque élément est appelé nœud. Il se décompose en plusieurs catégories :

- Une racine : ensemble de l'échantillon d'apprentissage.
- Un tronc et des branches : règles de division permettant de segmenter les données. L'algorithme CART crée des arbres binaires, c'est-à-dire que chaque nœud non terminal possède deux branches.
- Des feuilles : nœuds terminaux, données segmentées regroupées de manière homogène.

4.1.1.1 Algorithme CART

Le principe de l'algorithme CART consiste à partitionner récursivement les variables explicatives $X = (X_1; \dots; X_k)$ de manière binaire et à trouver l'arbre optimal pour la prédiction. Plusieurs problématiques sont liées à la construction d'un arbre. Tout d'abord, le choix de la variable explicative est important car il faut sélectionner la variable la plus discriminante. De plus, il est nécessaire de savoir « découper » les différentes modalités de la variable. Enfin, la taille optimale de l'arbre doit être recherchée afin de trouver un équilibre entre arbre trivial et sur-apprentissage. Deux étapes sont donc nécessaires à la construction d'un arbre de régression : la construction d'un arbre maximal (saturé) et l'élagage.

Construction de l'arbre saturé :

La construction de l'arbre maximal se fait à partir de la base d'apprentissage. La première étape consiste à diviser la racine, c'est-à-dire l'ensemble des données, en deux sous-ensembles les plus homogènes possibles. Puis, ces deux sous-ensembles sont eux-mêmes divisés en deux classes homogènes, et ainsi de suite. Les divisions sont de la forme $x_i \in \mathcal{C} \mid x_i \notin \mathcal{C}$. L'enjeu consiste donc à déterminer la variable explicative i et les modalités $\mathcal{C}$ prises par cette variable permettant une séparation la plus homogène possible, c'est-à-dire à déterminer la règle de découpage. Pour cela, il est nécessaire d'introduire les fonctions d'hétérogénéité. Ce sont des fonctions permettant de mesurer l'hétérogénéité au sein d'un nœud. L'objectif de l'algorithme CART consiste donc à réduire cette erreur lors de chaque division binaire.

Certaines notations sont introduites :

- t : nœud
- $D(t) = \{(x_{1,i}; \dots; x_{k,i}; y_i) \in t\}$: données appartenant au nœud t
- $N(t) = \mathrm{card}(D(t))$: nombre de valeurs appartenant au nœud t
- t_G et t_D : nœud droit et nœud gauche, descendants du nœud t
- $\bar{y}(t) = \frac{1}{N(t)} \sum_{D(t)} y_i$: moyenne empirique des valeurs de la variable à prédire, pour le nœud t

La fonction d'hétérogénéité correspond à la variance intra-nœud. Elle est définie par :

$$r(t) = \frac{1}{N(t)} \left(\sum_{D(t)} \left(y_i - \bar{y}(t)\right)^2 - \left(\sum_{D(t_G)} \left(y_i - \bar{y}(t_G)\right)^2 + \sum_{D(t_D)} \left(y_i - \bar{y}(t_D)\right)^2 \right) \right)$$

L'optimalité est obtenue lorsque $r(t)$ est maximale.

Pour diviser un nœud en deux, il est nécessaire de déterminer le critère de séparation de la variable. Trois cas peuvent alors se présenter :

- La variable explicative est qualitative et possède j modalités : le nombre de partitions possibles vaut $2^{j-1} - 1$
- La variable explicative est quantitative et discrète avec j modalités : le nombre de partitions vaut $j - 1$ car les modalités sont ordonnées. Le critère de séparation vaut la moyenne de deux modalités consécutives.
- La variable explicative est continue : ce cas se traite de la même façon que le cas précédent. Comme chacune des observations peut être différente, il y a autant de modalités possibles que d'observations, ce qui signifie que le nombre de partitions maximal vaut $n - 1$

La maximisation de la fonction d'hétérogénéité se fait par calcul de celle-ci pour l'ensemble des partitions possibles.

L'algorithme CART est un algorithme récursif : l'étape précédente permet d'obtenir deux branches (droite et gauche). La maximisation de la fonction d'hétérogénéité est ensuite effectuée sur chacune de ces deux branches, et ainsi de suite jusqu'à ce que plus aucune segmentation ne soit possible, c'est-à-dire lorsque tous les nœuds sont terminaux. Un nœud est terminal lorsque toutes les données contenues dans ce nœud sont telles que $y_i = \bar{y}(t)$ ou lorsqu'il n'existe plus de variable explicative à tester. Chaque nœud terminal est une feuille et est caractérisée par un unique chemin partant de la racine. Ce chemin correspond à une règle. L'ensemble des règles pour toutes les feuilles constitue le modèle.

Algorithme 1 : CART : construction d'un arbre saturé :

Entrée :	Échantillon d'apprentissage
Initialisation :	Racine ← données de l'échantillon
Itérations :	Répéter pour chaque nœud de l'arbre : Déterminer toutes les partitions possibles (pour les k variables explicatives et pour chaque potentielle division) Pour chaque partition : Calculer la fonction d'hétérogénéité $r(t)$ Fin pour Choisir la partition qui maximise $r(t)$ Segmenter selon la partition sélectionnée : division en nœud gauche et en nœud droit Jusqu'à : tous les nœuds sont terminaux
Sortie :	Arbre binaire saturé

Cet algorithme permet d'obtenir un arbre saturé, c'est-à-dire que plus aucune segmentation n'est possible. Cependant, en pratique, une condition d'arrêt est généralement spécifiée. Les conditions d'arrêts classiques consistent à fixer un nombre d'observations appartenant à un nœud au-delà duquel ces nœuds ne sont plus découpés ou à fixer la profondeur maximale d'un arbre.

Élagage de l'arbre :

Le modèle ainsi obtenu est adapté aux données d'apprentissage, par conséquent, il n'est pas nécessairement un bon estimateur pour d'autres données. En effet, l'erreur de prédiction diminue avec la complexité du modèle, c'est-à-dire lorsque l'arbre est saturé. Cependant, si le modèle est trop adapté aux données d'apprentissage, il perd sa qualité de prédicteur et est donc un mauvais modèle pour d'autres données. Choisir le modèle optimal revient à choisir celui parmi les modèles construits sur les données d'apprentissage qui minimise l'erreur calculée sur les données de validation.

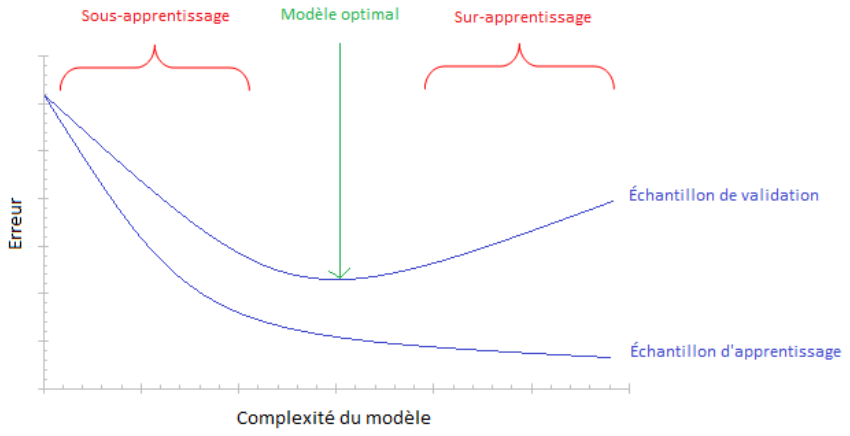

Figure 53 Choix du meilleur modèle

L'optimalité consiste donc à trouver le bon compromis entre biais et variance. En effet, l'arbre maximal a un biais faible mais sa variance est élevée. Pour éviter le sur-apprentissage et afin de rendre le modèle plus robuste et plus performant, il est nécessaire d'élaguer, c'est-à-dire chercher le meilleur sous-arbre, grâce à l'échantillon de validation, ce dernier étant totalement indépendant de la construction de l'arbre saturé.

Soit T un arbre. La fonction $f_T: \mathbb{R}^k \to \mathbb{R}$ associe à un vecteur $(x_1; \dots; x_k)$ de variables explicatives la valeur estimée grâce au modèle CART $y = f_T(x_1; \dots; x_k)$.

Classiquement, l'erreur est définie comme la somme des écarts au carré :

$$R(T) = \sum_{i=1}^{n} \left(y_i - f_T(x_{1,i}; \dots; x_{k,i})\right)^2$$

Potentiellement, tous les sous-arbres binaires de T_{max} (l'arbre saturé) et contenant la racine pourraient être l'arbre optimal. Ainsi, le but de l'algorithme d'élagage est de construire la suite de tous les meilleurs arbres à i feuilles (où i varie de 1 au nombre de feuilles de T_{max}). Ainsi, pour chaque sous-arbre de T_{max} à i feuilles construit grâce à la base d'apprentissage, l'algorithme sélectionne celui minimisant l'erreur sur la base de validation. Finalement, l'arbre optimal est celui parmi tous les sous-arbres minimisant l'erreur sur la base de validation. Cet algorithme est détaillé ci-dessous.

Algorithme 2 : CART : Élagage :

Entrées : Arbre saturé ; échantillon de validation

Initialisation : $T_{max} \leftarrow$ arbre saturé # construit grâce à l'échantillon d'apprentissage

$K \leftarrow |T_{max}|$ # nombre de feuilles de T_{max}

Tous les sous-arbres à i feuilles sont testés

Itérations : Pour $i \leftarrow K - 1$ à 1 :

$$T_i \leftarrow \underset{\substack{T \in \text{arbres construits} \\ \text{sur la base d'apprentissage} \\ \text{ayant } K-1 \text{ feuilles}}}{\operatorname{argmin}} R(T)$$

L'erreur $R(T)$ est testée pour les données de l'échantillon de validation

Fin pour

Cela permet d'obtenir une suite $T_1 \subset \dots \subset T_{K-1} \subset T_{max}$

$$k^* \leftarrow \underset{k \in \{1;\dots;K\}}{\operatorname{argmin}} R(T_k)$$

Arbre optimal $\leftarrow T_{k^*}$

Sortie : Arbre binaire optimal

Cet algorithme permet d'obtenir un arbre plus robuste, avec un nombre de feuilles optimal par rapport à la base de validation, indépendante de la construction du modèle. Cependant, la complexité de cet algorithme est extrêmement élevée car le nombre de sous-arbres potentiel croit exponentiellement en fonction du nombre de données. Néanmoins, une énumération implicite suffit à atteindre un arbre optimal. Pour cela, la fonction de pénalisation est définie : $R_\alpha(T) = R(T) + \alpha|T|$. Le premier terme est le critère d'erreur précédemment défini, il décroît en fonction du nombre de feuilles contrairement au deuxième terme qui croît avec le nombre de feuilles. Le paramètre α permet d'ajuster la pénalité. Ainsi, plus sa valeur est grande, plus les modèles comportant un grand nombre de feuilles, c'est-à-dire les modèles les plus complexes, sont pénalisés.

Breiman a démontré que pour chaque valeur de α, il existe un unique sous-arbre de T_{max} noté T_α tel que $T_\alpha = \underset{T \subset T_{max}}{\operatorname{argmin}} R_\alpha(T)$. De plus, il existe une suite finie croissante de réels positifs noté $(\alpha_k)_{k \in \{1;\dots;K-1\}}$ telle que $\forall k \in \{1;\dots;K-1\}$, si $\alpha \in [\alpha_k;\alpha_{k+1}[$ alors $T_\alpha = T_{\alpha_k}$. Par conséquent, construire la suite $T_1 \subset \dots \subset T_{K-1} \subset T_K$ grâce au critère $R(T)$ afin de déterminer le nombre de feuilles optimal k^* est équivalent à construire la suite d'arbres emboités $T_{\alpha_1} \subset \dots \subset T_{\alpha_{K-1}} \subset T_{\alpha_K}$ avec le critère $R_\alpha(T)$ afin d'en déduire la pénalité de complexité optimale α^* et l'arbre associé à celle-ci. Ce procédé évite de tester tous les sous-arbres potentiels en sélectionnant seulement une partie d'entre eux. En effet, les arbres qui ne sont pas sélectionnés sont ceux dont le critère de pénalité est plus grand, donc ils ne sont pas compétitifs par rapport à ceux sélectionnés. L'algorithme d'élagage est présenté ci-dessous :

Algorithme 3 : CART : Élagage :

Entrées : Arbre saturé ; échantillon de validation

Initialisation : $T_{max} \leftarrow$ arbre saturé # construit grâce à l'échantillon d'apprentissage

$K \leftarrow |T_{max}|$ # nombre de feuilles de T_{max}

$T_{\alpha_K} \leftarrow T_{max}$ # $\alpha_K = 0$

Itérations : Pour $i \leftarrow K-1$ à 1 :

$$\alpha_i \leftarrow \min_{\{t:noeud\ de\ T\}} \frac{R(t) - R\left(T^t_{\alpha_{i+1}}\right)}{\left|T^t_{\alpha_{i+1}}\right| - 1}$$

$T_i \leftarrow$ sous arbre élagué des branches t telles que $R_{\alpha_i}(T) = R(t) + \alpha_i$

Fin pour

$$\alpha^* \leftarrow \underset{\alpha \in \{\alpha_1;\dots;\alpha_K\}}{\operatorname{argmin}} R(T_\alpha)$$

Arbre optimal $\leftarrow T_{\alpha^*}$

Sortie : Arbre binaire optimal

Validation croisée :

En pratique, les données ne sont pas simplement séparées en deux échantillons : celui d'apprentissage et celui de validation. L'objectif de la validation croisée est que l'ensemble des données (excepté l'échantillon de test) servent à la fois à l'apprentissage et à la validation afin de rendre le modèle plus robuste et plus stable. Une des méthodes les plus utilisées est la méthode *k-fold*. Elle consiste à diviser aléatoirement les données (hors échantillon test) en k sous-échantillons de même taille. Chacun de ces échantillons est utilisé exactement $k-1$ fois pour la modélisation et 1 fois pour la validation. Cela permet à la fois de réduire le risque de sous-apprentissage puisque toutes les données sont utilisées pour l'apprentissage et de réduire celui de sur-apprentissage en utilisant tous les échantillons pour la validation.

4.1.1.2 Avantages et inconvénients des arbres

Théoriquement, les arbres de régression présentent certains avantages par rapport au modèle linéaire généralise :

- L'algorithme CART est une méthode non paramétrique : contrairement au modèle linéaire généralisé, aucune hypothèse n'est supposée quant à la distribution des données.
- Les regroupements de variables quantitatives sont effectués par l'utilisateur pour les modèles linéaires généralisés. L'algorithme CART permet de regrouper ces variables de manière optimale. Cela engendre plus de précision quant à la qualité de ces regroupements. En effet, les arbres permettent de traiter à la fois des variables explicatives catégorielles et continues.
- Les modèles linéaires généralisés déterminent de manière unique les coefficients β. Or, une variable peut avoir des effets différents selon les valeurs prises par une seconde variable. Les arbres permettent donc de regrouper les modalités d'une variable selon des critères différents selon les nœuds car le critère de séparation est réévalué pour chacun des nœuds.
- La sélection des variables est une phase importante lors de la calibration d'une modèle linéaire généralisé. Cette étape n'est pas nécessaire pour les arbres car, lors de la création des nœuds, les variables sont hiérarchisées selon l'importance de leur effet sur la variable à prédire. Cela permet de recueillir par ordre décroissant les variables les plus importantes.
- Le temps de calcul nécessaire à la construction d'un arbre CART (apprentissage et élagage) est relativement rapide. La complexité de l'algorithme est de l'ordre de $\mathcal{O}(kn\ln(n))$ dans le meilleur des cas (l'arbre est parfaitement équilibré) et de $\mathcal{O}(kn^2)$ dans le pire des cas (l'arbre est totalement déséquilibré*)
- Comme pour les modèles linéaires généralisés, les résultats issus de l'algorithme CART sont facilement interprétables. En effet, chaque prédiction peut être retrouvée en retraçant l'unique chemin de la racine à la feuille associée à la prédiction, en répondant aux questions successives de chaque nœud.

Cependant, les arbres présentent des limites non négligeables :

- Les arbres manquent de robustesse car ils dépendent fortement des données. En effet, pour deux échantillons d'apprentissage issus de la base de données, les deux arbres obtenus comportent généralement des différences (variables explicatives et segmentations de celles-ci). L'ajout de quelques données dans la base d'apprentissage ne devrait modifier que de façon marginale le modèle pour l'améliorer. Or, cet ajout d'observations peut totalement changer la topologie du modèle.
- L'arbre détecte des optimums locaux et non des optimums globaux. En effet, contrairement à la sélection stepwise des variables effectuée pour les modèles linéaires généralisés, la sélection des variables s'effectue de manière séquentielle pour chaque nœud, ainsi, il n'est pas possible d'avoir une visibilité globale sur l'effet des variables.

* Également appelé arbre dégénéré ou arbre en peigne

4.1.2 Forêt aléatoire

L'inconvénient majeur des arbres CART est leur manque de robustesse. En effet, les estimations construites varient beaucoup selon la base de données ayant servi à la construction. Par conséquent, les estimations peuvent être mauvaises sur une autre base de données. Pour limiter ce problème, il est possible d'utiliser des méthodes d'agrégation. L'idée principale consiste à construire un grand nombre d'arbres puis à les réunir afin d'obtenir une meilleure estimation. Cette méthode permet de réduire la variance de l'estimation et ainsi d'obtenir des résultats plus robustes.

Il existe différentes méthodes d'agrégation. Seule la méthode des forêts aléatoires est étudiée dans ce mémoire. Tout comme l'algorithme CART, elle a été introduite par Breiman en 2001. Il s'agit d'une méthode non-paramétrique.

Le principe des forêts aléatoires repose sur l'agrégation d'une combinaison d'arbres. L'idée principale de l'algorithme repose sur le fait que plutôt que d'avoir un estimateur très complexe, il est plus judicieux d'en construire plusieurs d'une qualité individuelle moins développée, mais dont l'agrégation apporte une grande efficacité. Les arbres sont construits de manière à être le plus décorrélés possible, de sorte qu'ils soient indépendants entre eux, ce qui permet de réduire la variance du modèle. Pour cela, l'algorithme procède à un double échantillonnage : à la fois sur les observations (*tree bagging*[*]), mais aussi sur les variables (*feature sampling).*

La première étape concerne la création de B échantillons bootstrappés. La technique de Bootstrap consiste à rééchantillonner les données, c'est-à-dire à créer de « nouveaux échantillons » à partir de l'échantillon de données D. Ces échantillons sont de taille n et sont constitués à partir des observations de l'échantillon initial par tirage aléatoire équiprobable avec remise.

Puis, à partir de chacun de ces échantillons, un arbre est créé. La particularité des forêts aléatoires concerne la sélection des variables. En effet, pour chaque nœud, plutôt que de sélectionner la division optimale parmi toutes les divisions de variables explicatives possibles, le choix de la division se limite à un nombre m de variables explicatives. Ces m variables sont tirées aléatoirement (sans remise) parmi les k variables explicatives. Par défaut, le nombre de variables est fixé par $m = \left\lfloor \frac{k}{3} \right\rfloor$ pour une régression.

Les forêts aléatoires comportent donc deux sources d'aléas : l'aléa dû au Bootstrap et l'aléa du choix des variables à chaque nœud. Le fait de ne prendre que m variables explicatives parmi les k à disposition permet de construire B arbres les plus indépendants possible. Couplé au fait de choisir B grand, cela permet de réduire significativement la variance. En effet, la variance d'une forêt aléatoire est égale à $\mathbb{V}[\text{forêt}] = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$ où ρ désigne la corrélation entre les différents arbres et σ^2 leur variance. En augmentant le nombre d'arbres B grâce à l'étape Bootstrap, le second terme $\frac{1-\rho}{B}\sigma^2$ diminue. Pour réduire le premier terme, il est nécessaire de construire des arbres indépendants, de telle sorte que $\rho \sim 0$, ce qui est rendu possible grâce à l'étape de sélection aléatoire des variables.

La dernière étape consiste à agréger les B arbres obtenus. La variable d'intérêt est obtenue en calculant la moyenne des différentes estimations des B arbres :

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} \widehat{f_b}(x_1; \dots; x_k)$$

[*] Bagging = Bootstrap Aggregation

La figure ci-dessous présente de manière synthétique les différentes étapes de construction évoquées précédemment. Comme le montrent les différentes flèches, l'algorithme est parallélisable, ce qui permet de réduire le temps de calcul.

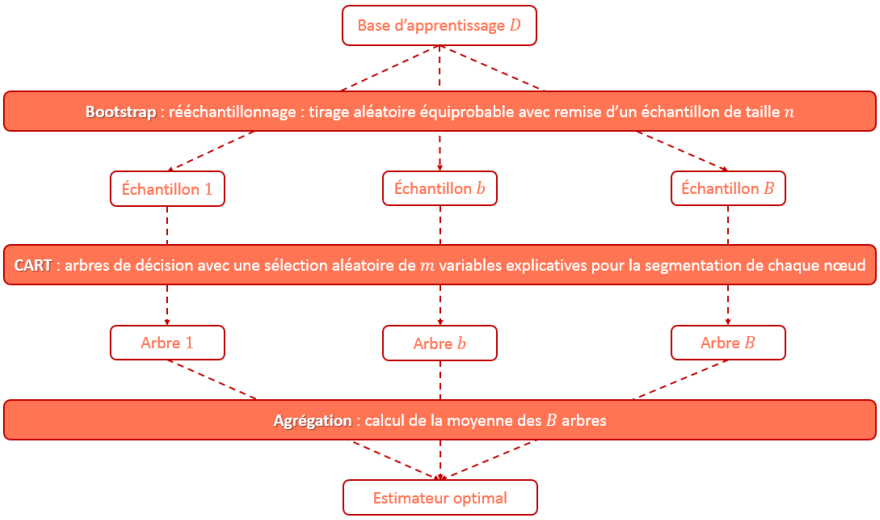

Figure 54 : Schéma synthétique de l'algorithme Forêt aléatoire (Random Forest)

L'algorithme Random Forest permet à la fois de construire un prédicteur et de calculer une estimation de l'erreur de généralisation : l'erreur OOB (*Out-of-Bag* en anglais, littéralement « *en dehors du Bootstrap* » en français). Pour chaque donnée $(x_{1;i}; \dots; x_{k;i}; y_i)$ de l'échantillon d'apprentissage D, la prédiction de Y_i est calculée uniquement à partir des arbres construits grâce aux échantillons Bootstrap ne contenant pas cette donnée. En comparant la prédiction obtenue à la valeur réellement observée, il est possible de calculer l'erreur quadratique due à cette donnée. Puis, après avoir fait ceci pour l'ensemble des données, l'erreur globale Out-of-Bag peut être calculée :

$$err_{OOB} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \frac{1}{B_i}\sum_{\substack{arbres\ b\ construits \\ sur\ les\ échantilllons \\ Bootstrap\ ne\ contenant \\ pas\ l'observation\ i}} \widehat{f_b}(x_{1;i}; \dots; x_{k;i})\right)^2$$

Comme pour la validation croisée, les données servant à calculer l'erreur n'ont pas été rencontrées au cours de l'étape de modélisation. Cependant, l'avantage par rapport aux estimateurs CART est qu'il n'est pas nécessaire de découper l'échantillon d'apprentissage puisque celui-ci se fait de manière implicite lors de l'étage de Bootstrap.

Un des principaux avantages des arbres CART est leur facilité d'interprétation. Cette caractéristique se perd pour les forêts aléatoires du fait de l'agrégation des modèles. Cependant, pour palier à cette contrainte, Breiman introduit un indice permettant de mesurer l'importance des variables. Pour chaque échantillon OOB, les valeurs de la variable dont l'importance est recherchée sont permutées aléatoirement. L'importance de la variable est définie comme la différence entre l'erreur OOB calculée sur l'échantillon perturbé et celle sur l'échantillon original. Plus cette différence est grande, plus la variable a de l'importance car en brisant le lien entre cette variable et la variable explicative, une erreur importante est commise. À l'inverse, si les permutations n'ont aucun effet sur l'erreur, la variable n'a que très peu d'importance.

Pour conclure cette présentation théorique, l'algorithme Random Forest est récapitulé page suivante.

Algorithme 4 : Forêt aléatoire (Random Forest)

Entrée : Données D

Initialisation : $B \leftarrow$ nombre d'échantillons Bootstrap

$m \leftarrow$ nombre de variables choisies pour la segmentation de chaque nœud

Itérations : Pour b $\leftarrow 1$ à B :

$D_b \leftarrow$ échantillon Bootstrap des données # n tirages aléatoires équiprobables avec remise sont effectués sur l'échantillon des données D

$w_b = \begin{pmatrix} w_{b;1} \\ \vdots \\ w_{b;n} \end{pmatrix} \leftarrow$fréquence de tirage dans l'échantillon b de chacune des observation de D

$T_b \leftarrow \mathrm{CART}(D_b; m)$ # À chaque nœud de l'arbre construit grâce à l'échantillon Bootstrap, m variables sont sélectionnées aléatoirement pour découper le nœud. Permet d'obtenir l'estimateur $\widehat{f_b}$.

Pour $j \leftarrow 1$ à k :

$OOB_b \leftarrow D \setminus D_b$ # échantillon out-of-bag

$$err^j_{OOB_b} \leftarrow \frac{1}{|OOB_b|} \sum_{i=1}^{|OOB_b|} \left(y_i - \widehat{f_b}(x_1; \dots; x_j; \dots; x_k)\right)^2$$

$\widetilde{OOB_b} \leftarrow OOB_b$ avec permutation aléatoire des valeurs de la variable X_j

$$err^j_{\widetilde{OOB_b}} \leftarrow \frac{1}{|\widetilde{OOB_b}|} \sum_{i=1}^{|\widetilde{OOB_b}|} \left(y_i - \widehat{f_b}(x_1; \dots; x'_j; \dots; x_k)\right)^2$$

Fin pour

Fin pour

Pour $j \leftarrow 1$ à k : # Calcul de l'importance des variables

$$IV(X_j) \leftarrow \sum_{b=1}^{B} \left(err^j_{\widetilde{OOB_b}} - err^j_{OOB_b}\right)$$

Fin pour

Agrégation : estimation finale de la forêt obtenue par moyenne

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} \widehat{f_b}(x_1; \dots; x_k)$$

Calcul de l'erreur globale de la forêt ($B_i \leftarrow$ nombre d'échantillons ne contenant pas i

$$err_{OOB} \leftarrow \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \frac{1}{B_i} \sum_{b \text{ tel que } w_{b;i}=0} \widehat{f_b}(x_1; \dots; x_k)\right)^2$$

Sorties : Estimateur $\hat{f}$; erreur Out-of-Bag err_{OOB} ; importance des variables

4.2 Modélisation

Cette partie du mémoire est consacrée à la présentation des résultats des algorithmes CART et Forêt aléatoire appliqués aux données afin de modéliser la fréquence et le coût moyen des garanties bris de glace et dommages tous accidents.

4.2.1 Modélisation pour la garantie bris de glace

4.2.1.1 Modélisation de la fréquence

Arbre de CART :

La première étape correspond à la construction de l'arbre saturé. Les branches de cet arbre sont représentées à titre informatif sur la figure ci-dessous.

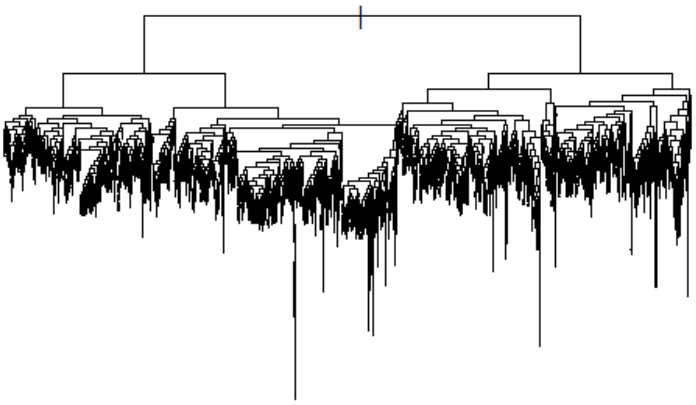

Figure 55 Arbre de régression saturé pour la modélisation de la fréquence des bris de glace

Cet arbre est composé de 7 526 feuilles. Outre son illisibilité, il présente le défaut d'être sur-appris en étant adapté aux données d'apprentissage. Ainsi, comme évoqué dans la partie précédente, afin de rendre le modèle plus robuste, il est nécessaire d'élaguer l'arbre. Pour cela, la cross-validation est mise en œuvre.

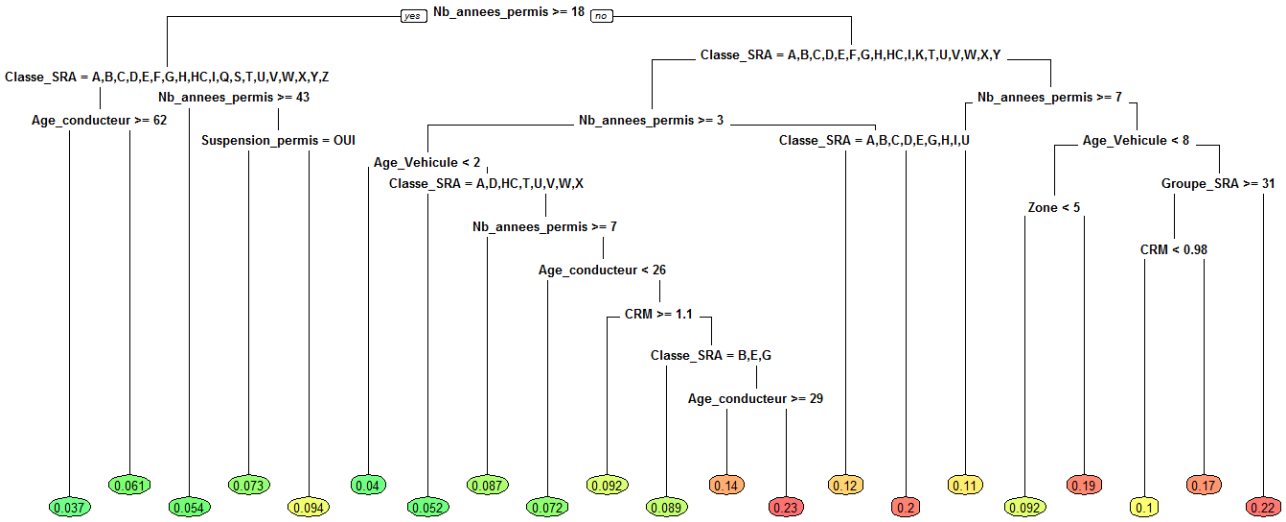

Figure 56 Arbre CART optimal pour la modélisation de la fréquence de Bris de glace

L'arbre optimal contient 21 feuilles, ce qui correspond à une segmentation assez peu développée. Les variables les plus influentes dans le modèle sont l'ancienneté de permis et la classe SRA du véhicule. La fréquence augmente lorsque l'ancienneté du permis de conduire baisse.

Forêt aléatoire :

Il est possible de tracer l'erreur en fonction du nombre d'arbres B. Cette erreur diminue lorsque le nombre d'arbres augmente, cependant, elle se stabilise à partir d'une certaine valeur correspondant au nombre d'arbres optimal.

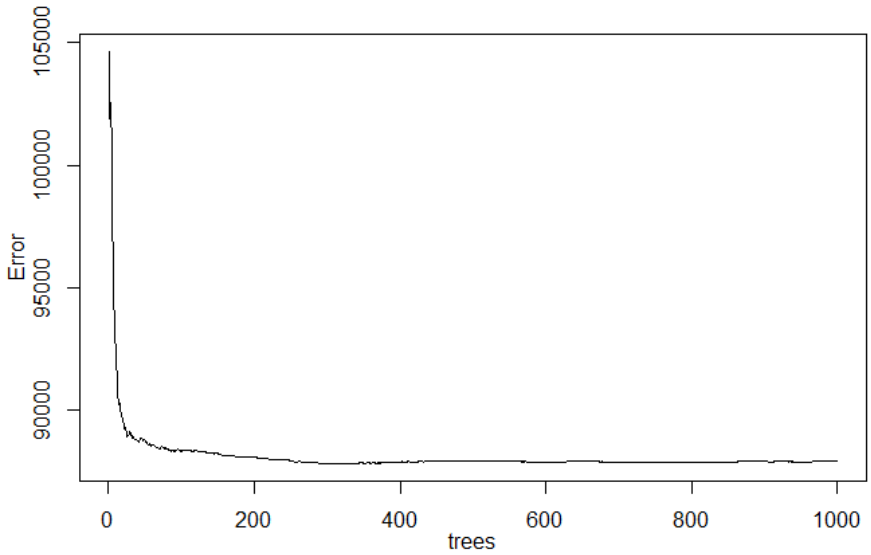

Figure 57 Évolution de l'erreur en fonction du nombre d'arbres

Bien que les forêts aléatoires soient moins facilement interprétables que les modèles linéaires généralisés et les arbres CART, il est tout de même possible de déterminer l'importance des variables explicatives. Le graphique ci-dessous représente la diminution moyenne de l'impureté des nœuds selon les différentes variables.

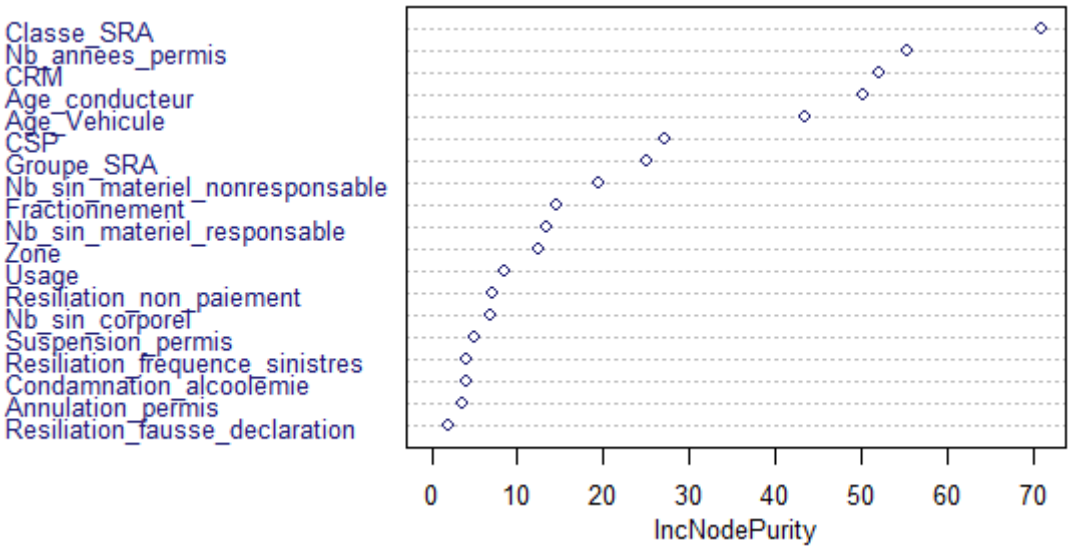

Figure 58 Importance des variables explicatives dans la forêt aléatoire modélisant la fréquence des Bris de glace

Comme pour l'arbre CART, les deux variables les plus importantes sont la classe SRA du véhicule et l'ancienneté du permis de conduire du conducteur.

Malgré l'absence de coefficients permettant d'expliquer la modélisation réalisée, il est possible d'avoir un aperçu des tendances modélisées. Pour cela, la réponse moyenne pour différentes modalités est tracée.

Tout d'abord, le graphique ci-dessous présente la fréquence moyenne modélisée grâce à la forêt aléatoire selon la classe SRA du véhicule, c'est-à-dire selon la variable la plus importante.

Figure 59 Fréquence moyenne modélisée pour la garantie Bris de glace selon la classe SRA du véhicule

De manière générale, la fréquence augmente selon la classe du véhicule, puis, elle diminue pour les véhicules très chers, c'est-à-dire ceux dont la classe dépasse U. Néanmoins, contrairement aux coefficients issus du modèle linéaire généralisé, ces valeurs doivent être interprétées avec parcimonie car elles dépendent de la classe SRA mais aussi de toutes les autres variables explicatives.

Le portefeuille étudié étant issu de produits d'assurance aggravés, il est nécessaire d'étudier la modélisation selon les variables décrivant une aggravation.

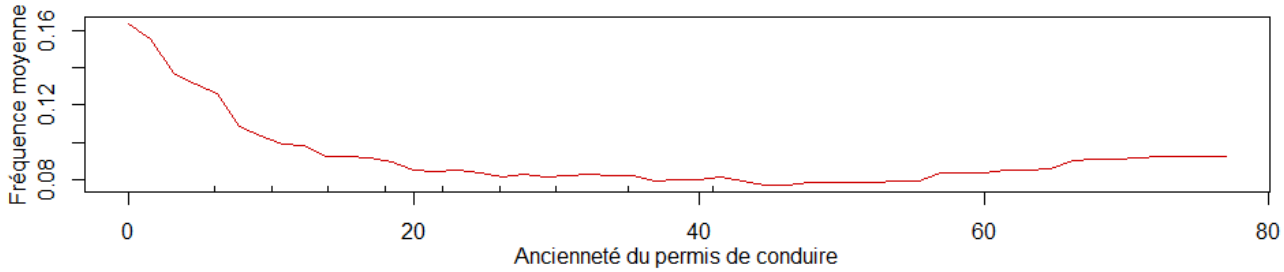

Figure 60 Fréquence moyenne modélisée pour la garantie Bris de glace selon l'ancienneté du permis

La fréquence de survenance d'un sinistre bris de glace décroit significativement en fonction de l'ancienneté du permis de conduire de l'assuré, et donc indirectement de son expérience de conduite.

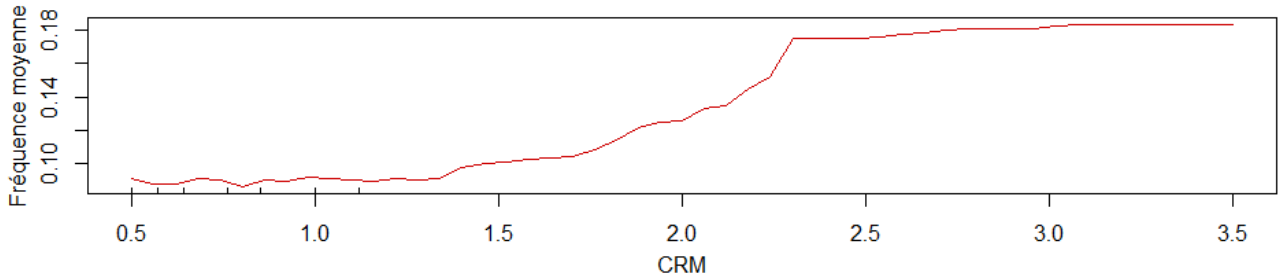

Figure 61 Fréquence moyenne modélisée pour la garantie Bris de glace selon le coefficient de réduction-majoration

La fréquence modélisée grâce à la forêt aléatoire est relativement stable pour les assurés possédant un bonus automobile. Plus le malus est élevé, plus la fréquence modélisée accroit fortement.

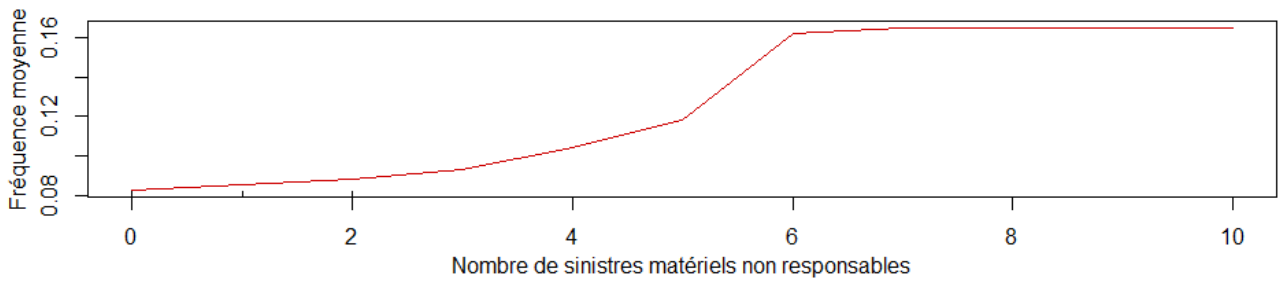

Figure 62 Fréquence moyenne modélisée pour la garantie Bris de glace selon les antécédents de sinistres matériels non responsables

Plus les assurés ont eu un nombre de sinistres matériels non responsables important avant la souscription de leur contrat, plus la forêt aléatoire modélise une fréquence de sinistralité moyenne importante.

L'arbre et la forêt aléatoire semblent modéliser les mêmes tendances remarquées lors de l'analyse de la sinistralité dans le deuxième chapitre ainsi que celles issues du modèle linéaire généralisé. Leur performance est étudiée dans le dernier chapitre de ce mémoire.

4.2.1.2 Modélisation du coût

Après avoir modélisé la fréquence de survenance d'un sinistre bris de glace, il est nécessaire de modéliser son coût afin d'en déduire la prime pure.

Arbre CART :

Tout d'abord, l'algorithme CART est mis en œuvre afin de déterminer un arbre de régression modélisant le coût moyen des sinistres bris de glace.

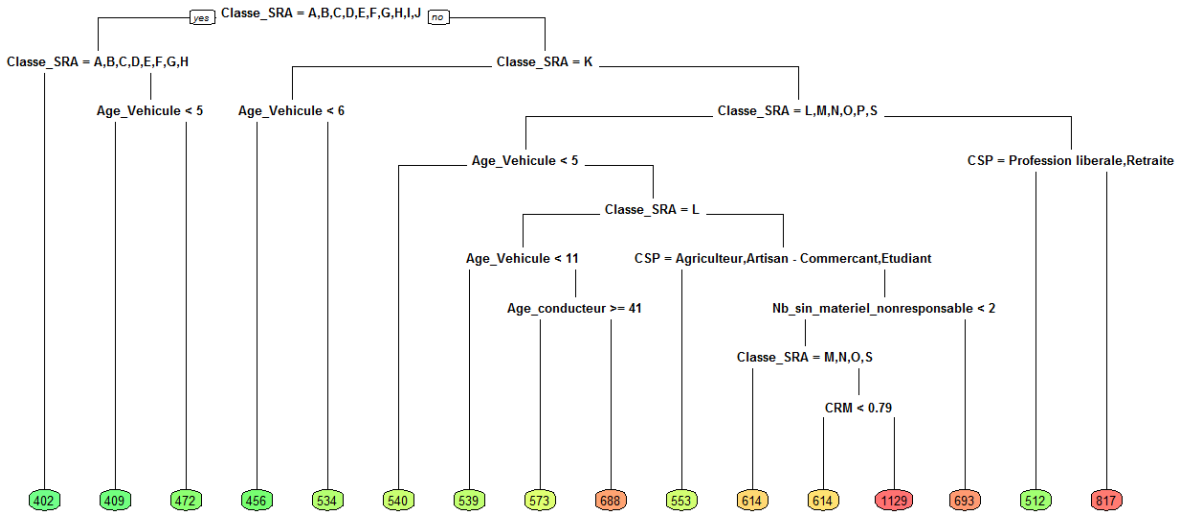

Figure 63 Arbre CART optimal pour la modélisation du coût des sinistres Bris de glace

L'arbre optimal ne comporte que 16 feuilles, ce qui correspond à une segmentation encore plus faible que celle utilisée pour la fréquence. La classe SRA est la variable la plus importante, en effet, comme évoqué dans le troisième chapitre, par définition, elle décrit la valeur du véhicule et de ses pièces de remplacement. Les coûts les moins importants correspondent à des petites classes, et réciproquement, plus la classe du véhicule est élevée, plus le coût du sinistre est important.

Forêt aléatoire :

L'algorithme forêt aléatoire est implémenté pour la modélisation du coût des sinistres bris de glace.

La figure ci-dessous présente l'importance des variables explicatives.

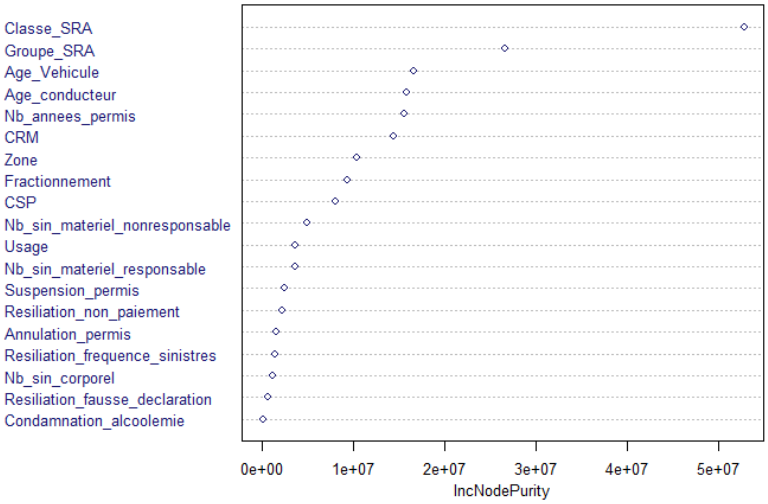

Figure 64 Importance des variables explicatives dans la forêt aléatoire modélisant le coût des sinistres Bris de glace

Ce graphique montre distinctement que les variables les importantes sont les caractéristiques du véhicule, et en particulier la classe définie par le SRA.

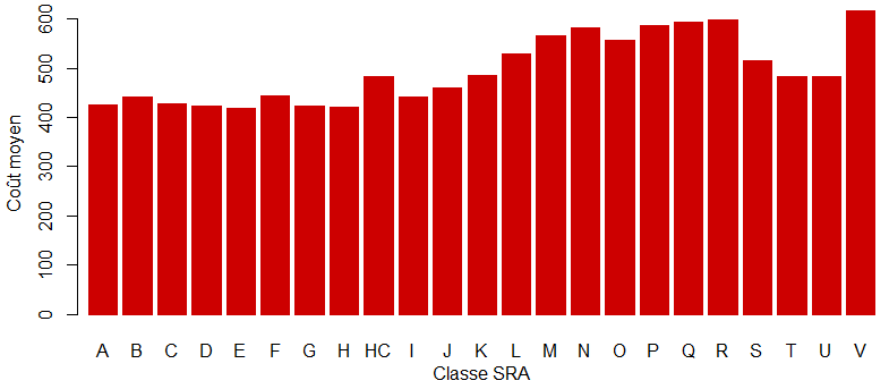

Figure 65 Coût moyen modélisé pour la garantie Bris de glace selon la classe du véhicule

Plus la classe du véhicule est élevée, plus le montant moyen du sinistre est important, ce qui est rationnel par définition de cette variable.

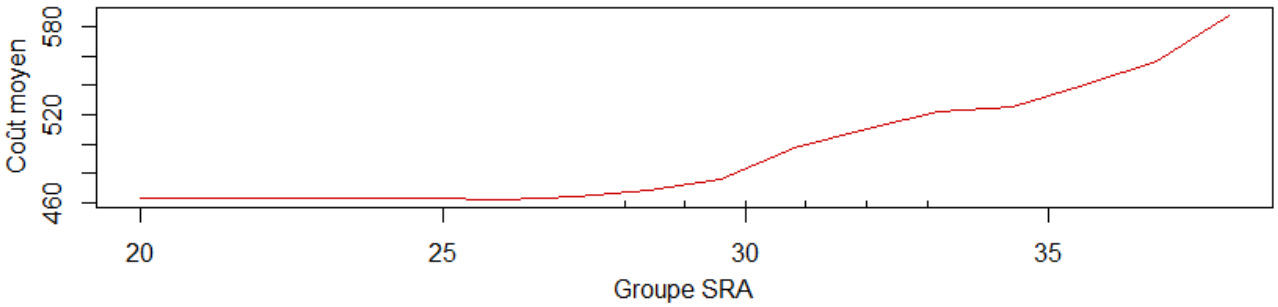

Figure 66 Coût moyen modélisé pour la garantie Bris de glace selon le groupe du véhicule

La deuxième variable la plus importante est également une caractéristique du véhicule. Lorsque le groupe augmente, le coût augmente également.

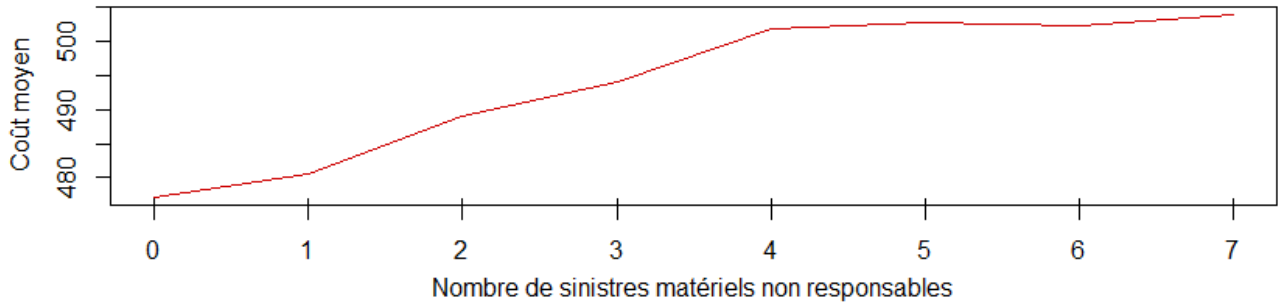

Figure 67 Coût moyen modélisé pour la garantie Bris de glace selon les antécédents de sinistres matériels non responsables

Les antécédents de sinistres représentent des variables importantes, notamment pour les produits aggravés. Le graphique ci-dessus nombre l'évolution du coût moyen modélisé par la forêt aléatoire en fonction du nombre de sinistres matériels non responsables, dont font partie les sinistres bris de glace.

Les arbres CART et les forêts aléatoires permettent de constater que les variables les plus importantes à la modélisation de la prime pure de la garantie bris de glace, tant pour la fréquence que pour le coût, sont les caractéristiques du véhicule. De plus, l'âge du conducteur, son ancienneté de permis et son CRM ont également une grande influence sur la modélisation.

4.2.2 Modélisation pour la garantie dommages tous accidents

Après avoir modélisé la fréquence de sinistralité et le coût moyen des sinistres de la garantie Bris de glace grâce à des arbres CART et des forêts aléatoires, ces mêmes méthodes sont appliquées afin de déterminer la prime pure de la garantie dommages.

4.2.2.1 Modélisation de la fréquence

Arbre CART :

L'arbre CART optimal permettant de modéliser la fréquence des sinistres dommages du véhicule est composé de 41 feuilles. Il est présenté sur la figure ci-dessous.

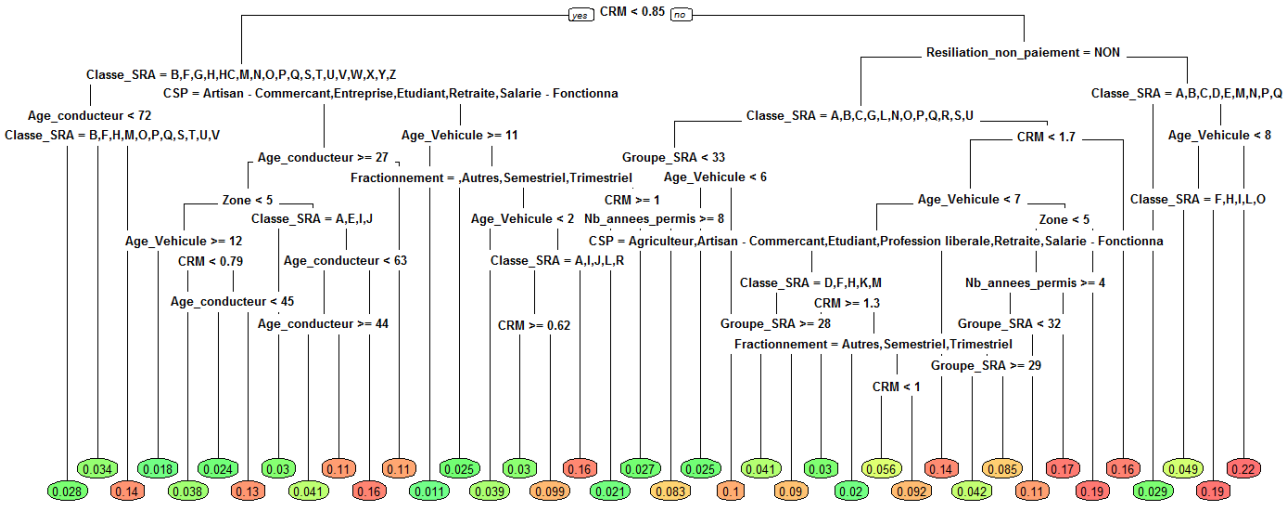

Figure 68 Arbre CART optimal pour la modélisation de la fréquence des sinistres Dommages

La segmentation de cet arbre est plus développée que celle des arbres des sinistres bris de glace.

La variable la plus discriminante est le coefficient de réduction-majoration. Elle est directement liée au risque du conducteur puisqu'elle est à la fois rattachée au nombre d'années depuis lesquelles l'assuré est titulaire du permis de conduire, mais aussi à ses antécédents de sinistralité responsable.

Forêt aléatoire :

La forêt aléatoire est une méthode d'agrégation permettant théoriquement d'obtenir de meilleurs résultats que ceux issus de l'algorithme CART. Elle est utilisée pour modéliser la fréquence des sinistres dommages.

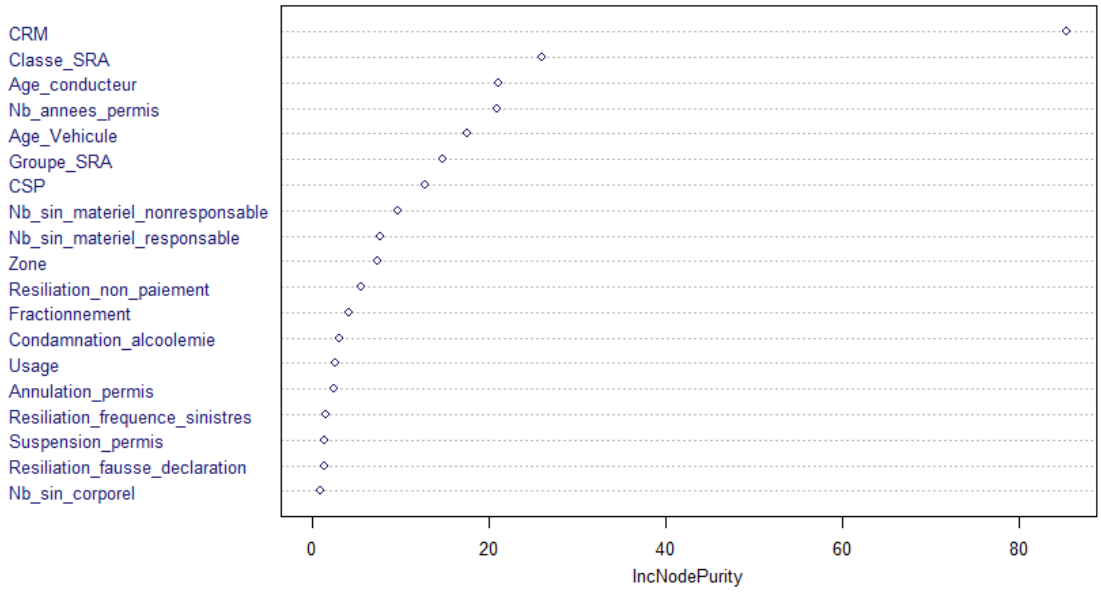

Figure 69 Importance des variables explicatives dans la forêt aléatoire modélisant la fréquence des sinistres Dommages

La variable la plus significative est incontestablement le coefficient de réduction-majoration. Il est étonnant de noter que d'après le graphe ci-dessus, les différentes causes de résiliation n'ont que très peu d'importance dans la modélisation de la fréquence des dommages. Cependant, cela peut être expliqué par la construction de la variable *importance des variables*. En effet, celle-ci est calculée en effectuant des permutations aléatoires dans les échantillons OOB. Comme les variables de résiliation, de suspension et d'annulation de permis ne comportent que deux modalités, dont la plupart valent *NON*, les permutations font moins varier l'erreur que si les variables étaient continues. Néanmoins, comme les arbres ne génèrent pas de coefficients explicites et ne sont pas facilement interprétables, cette hypothèse ne peut pas être vérifiée. Afin de tout de même avoir un aperçu de la modélisation générée par la forêt aléatoire, la valeur moyenne modélisée est tracée en fonction de plusieurs variables.

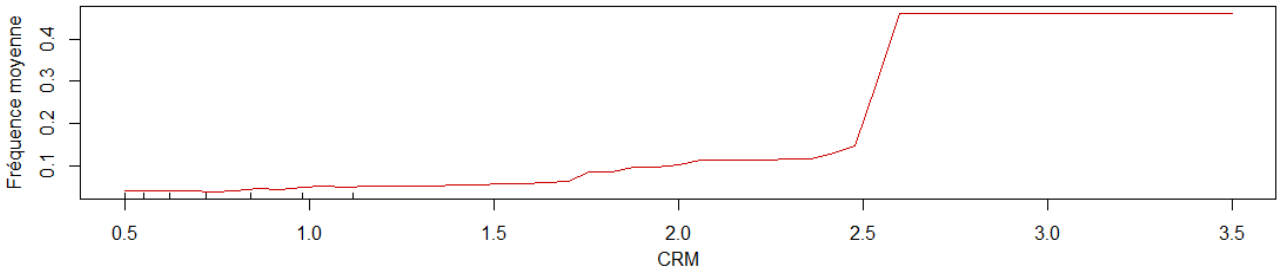

Figure 70 Fréquence moyenne modélisée pour la garantie Dommages selon le coefficient de réduction-majoration

Comme remarqué dans les précédentes parties, plus le CRM augmente, plus la fréquence de sinistralité de la garantie Dommages au véhicule augmente.

La garantie dommages est fortement liée aux caractéristiques du conducteur. Les deux figures suivantes mettent en avant deux d'entre elles : l'âge du conducteur et la durée depuis laquelle il est titulaire du permis.

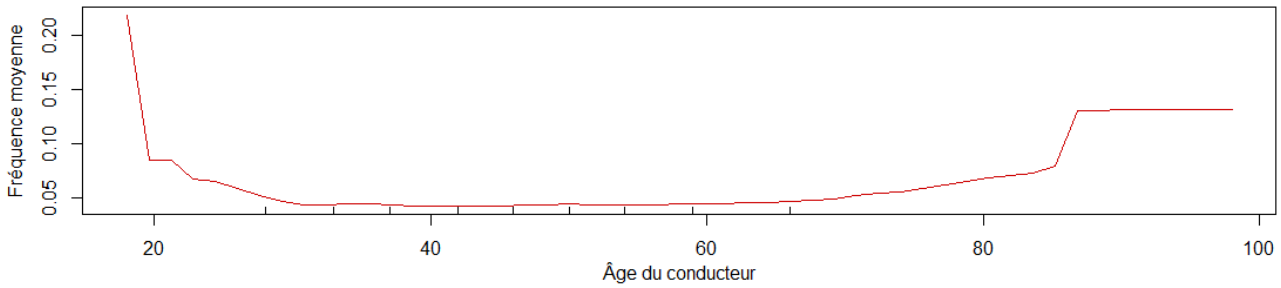

Figure 71 Fréquence moyenne modélisée pour la garantie Dommages selon l'âge du conducteur

La forêt aléatoire permet de modéliser des fréquences plus élevées pour les jeunes et les conducteurs âgés. Ces deux types d'assurés rentrent dans la catégorie « risque aggravé » dans les conditions de souscription fixées par la Direction des Partenariats.

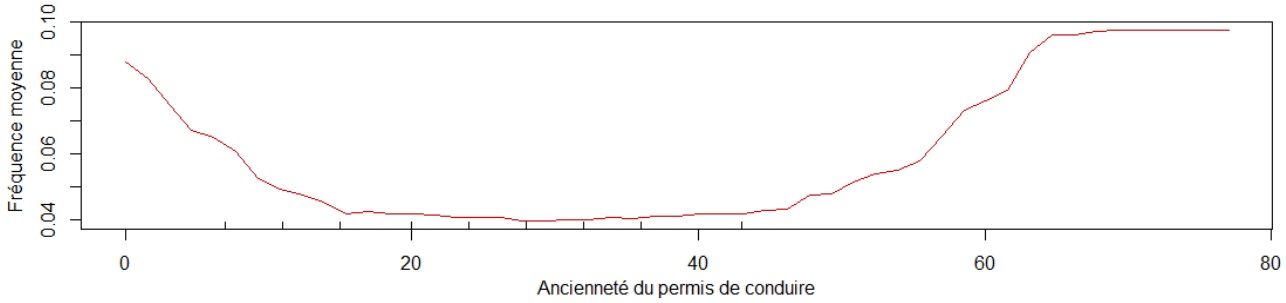

Figure 72 Fréquence moyenne modélisée pour la garantie Dommages selon l'ancienneté du permis

Plus le conducteur a de l'expérience de conduite, plus sa fréquence de sinistralité modélisée grâce à la forêt diminue. Puis, cette fréquence augmente lorsque l'assuré devient âgé, c'est-à-dire lorsqu'il possède le permis de conduire depuis longtemps.

Bien que les causes d'aggravation pour résiliation ou condamnation n'apparaissent pas comme des variables importantes pour les raisons évoquées précédemment, la forêt modélise des fréquences moyennes différentes selon les causes d'aggravation.

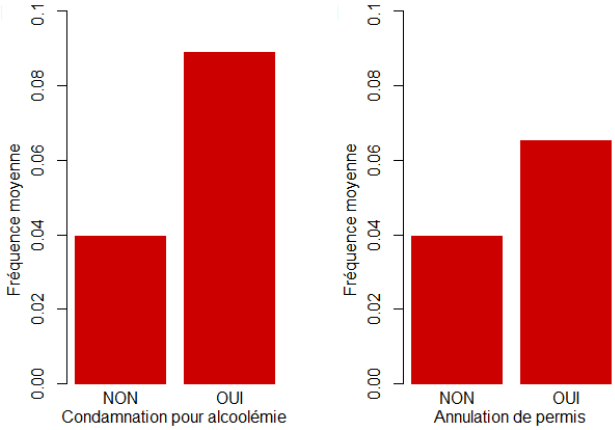

Figure 73 Fréquence moyenne modélisée pour la garantie Dommages selon les antécédents d'alcoolémie et d'annulation de permis

La cause de risque aggravé présentant la plus grande variation de fréquence modélisée est le fait d'avoir été condamné pour consommation d'alcool. En effet, la forêt aléatoire modélise pour ces assurés une fréquence moyenne de 8,3%. Le fait d'avoir eu une annulation de permis augmente également la fréquence moyenne modélisée par la forêt.

4.2.2.2 Modélisation du coût

Pour terminer ce chapitre consacré aux méthodes de Machine Learning, les deux algorithmes étudiés sont appliqués à la modélisation du coût des sinistres dommages.

Arbre CART :

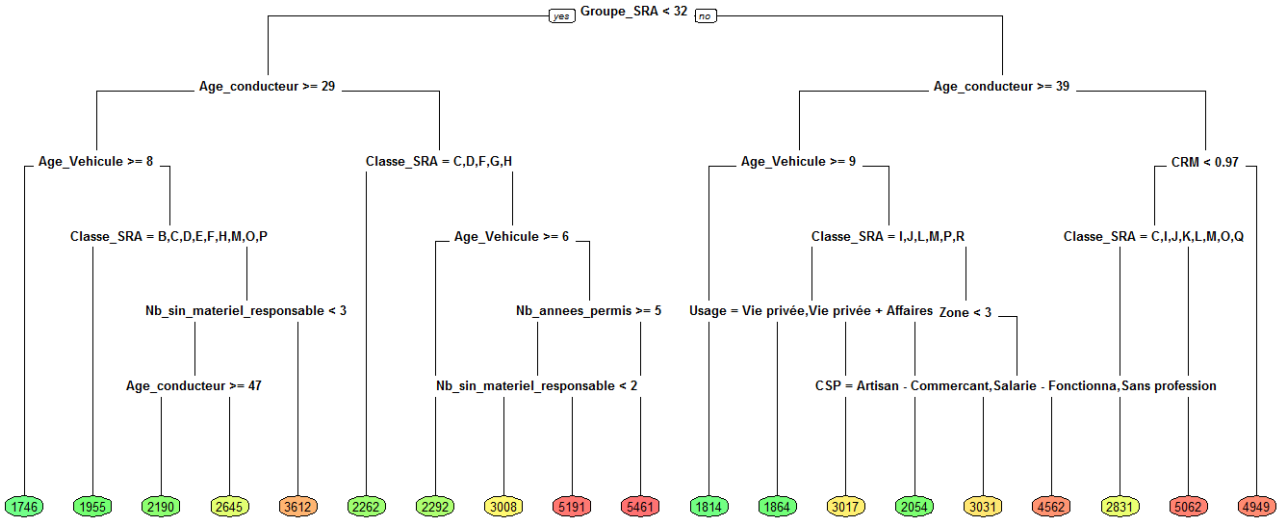

Figure 74 Arbre CART optimal pour la modélisation du coût des sinistres Dommages

L'arbre optimal ne comporte que 19 feuilles. Les variables ayant la plus forte influence sur la modélisation sont les caractéristiques du véhicule (groupe, classe, âge) ainsi que des variables décrivant la cause d'aggravation du risque (âge du conducteur, CRM, antécédents de sinistres responsables).

Forêt aléatoire :

Bien que les résultats d'une forêt aléatoire ne soient pas facilement interprétables, il est possible de visualiser les variables ayant le plus d'impact sur la modélisation.

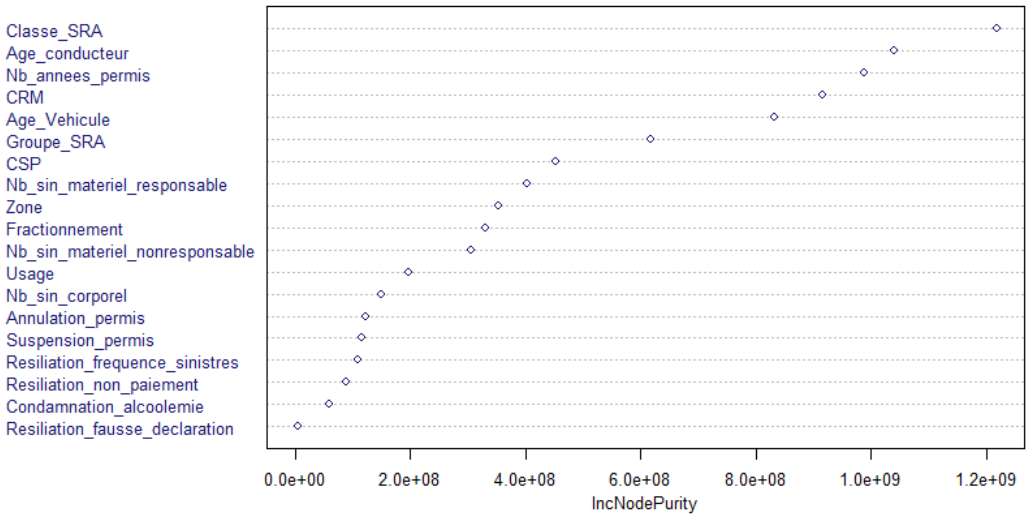

Figure 75 Importance des variables explicatives dans la forêt aléatoire modélisant le coût des sinistres Dommages

Comme pour la modélisation du coût des sinistres bris de glace, la variable la plus importante est la classe SRA du véhicule. En effet, il s'agit d'une variable décrivant le coût du véhicule et de ses pièces de remplacement, et par conséquent, il est naturel qu'elle soit fortement liée au coût des sinistres matériels du véhicule.

Afin d'avoir un aperçu de la modélisation effectuée par la forêt aléatoire, le coût moyen des sinistres est tracé selon plusieurs variables.

Le véhicule assuré a une grande importance sur le coût des sinistres puisque ses caractéristiques sont directement liées au coût de réparation ou à la valeur de remplacement du véhicule.

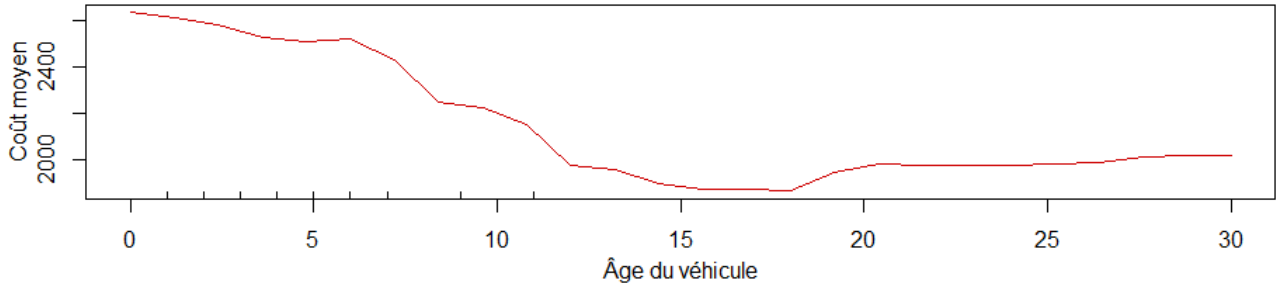

Figure 76 Coût moyen modélisé pour la garantie Dommages selon l'âge du véhicule

Globalement, le coût des sinistres dommages décroit en fonction de l'âge du véhicule.

Le coût moyen modélisé est ensuite analysé selon les variables décrivant le type d'aggravation du risque.

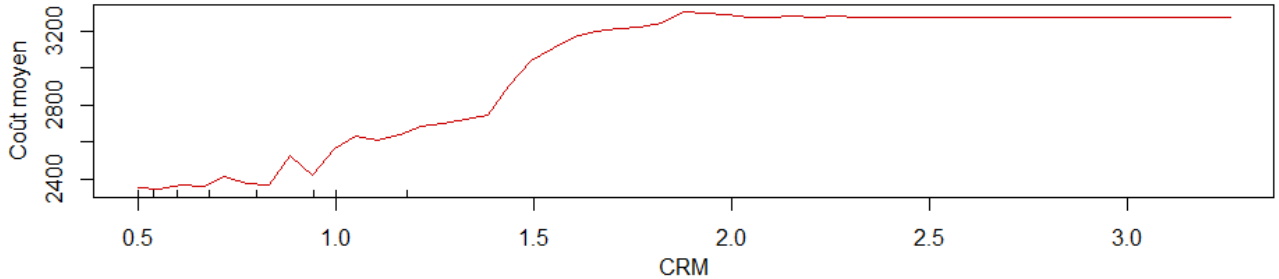

Figure 77 Coût moyen modélisé pour la garantie Dommages selon le coefficient de réduction-majoration

Plus l'assuré est en situation de malus, plus le coût moyen modélisé par la forêt augmente.

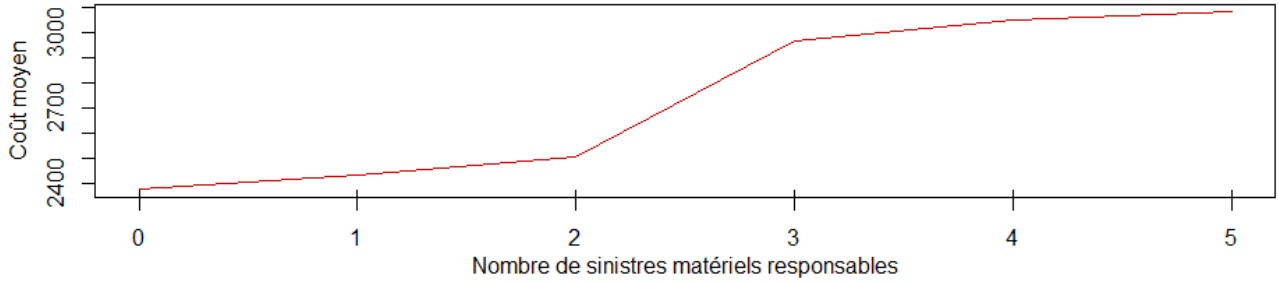

Figure 78 Coût moyen modélisé pour la garantie Dommages selon les antécédents de sinistres matériels responsables

Les sinistres responsables déclarés par l'assuré font référence à leur quantité durant les trois années avant la date d'effet du contrat et non pas à leur gravité, c'est-à-dire à leur coût. Cependant, plus l'assuré a des antécédents de sinistres responsables importants, plus le coût des sinistres est élevé.

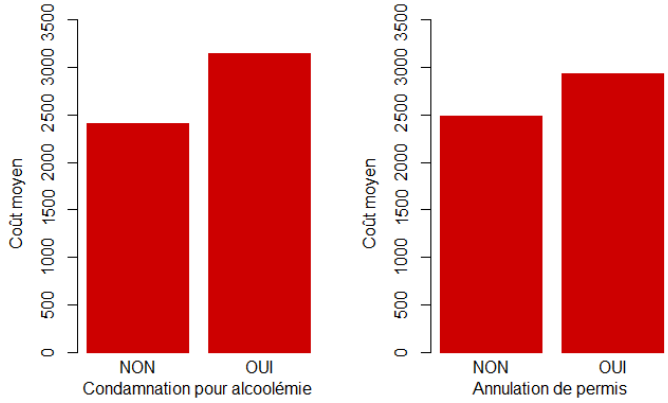

Figure 79 Coût moyen modélisé pour la garantie Dommages selon les antécédents d'alcoolémie et d'annulation de permis

Le coût moyen des individus ayant été condamné pour alcoolémie au volant ou ayant fait l'objet d'une annulation de permis présentent un coût moyen modélisé grâce à la forêt aléatoire plus important que les autres conducteurs.

Contrairement au modèle linéaire généralisé qui permet d'obtenir des coefficients explicites, les fréquences et les coûts présentés précédemment prennent en compte toutes les variables. Il n'est donc pas possible d'isoler l'effet d'une variable précise. Cependant, les arbres, tout comme les forêts sont des méthodes de Machine Learning dont l'un des avantages est la possibilité de connaître l'importance des variables.

Chapitre 5 : Analyse des résultats

Les chapitres 3 et 4 exposent la modélisation de la fréquence et du coût des sinistres des garanties Bris de glace et Dommages au véhicule selon différents procédés, une méthode classique, traditionnellement utilisée en tarification IARD : le modèle linéaire généralisé ; et deux méthodes de Machine Learning : l'arbre CART et la forêt aléatoire.

Après avoir détaillé la théorie relative à ces différentes méthodes ainsi que les résultats obtenus, il est nécessaire de comparer les modélisations réalisées afin de juger leur qualité et de déterminer si une méthode est plus efficace que les autres.

Ce chapitre se décompose en trois parties. La première permet d'analyser la conception des modèles en s'intéressant aux variables explicatives utilisées. La deuxième partie examine les résultats des modèles et vérifie que la charge totale des sinistres est couverte par la prime modélisée. Enfin, la dernière partie consiste à vérifier le caractère prédictif des modèles.

5.1 Variables explicatives utilisées

Les trois modèles implémentés n'utilisent pas les mêmes variables explicatives. Le modèle linéaire généralisé les sélectionne grâce à une analyse stepwise réalisée selon le critère AIC ou grâce à un test de *type III*. L'usage du véhicule et le fractionnement étant fortement corrélé à d'autres variables, ces deux variables ont été retirées pour les modèles linéaires généralisés. L'arbre CART choisit les variables de telle sorte qu'à chaque itération, la variable permettant la meilleure réduction de l'erreur est sélectionnée. Lors de l'élagage pour obtenir l'arbre optimal, seule les plus significatives, c'est-à-dire les plus proches de la racine, sont conservées. La forêt aléatoire permet d'avoir un aperçu des variables les plus importantes en calculant l'erreur commise sur l'échantillon OOB en réalisant des permutations aléatoires sur cette variable. Plus l'erreur est grande, plus la variable est significative.

Le tableau page suivante présente les variables retenues par les modèles linéaires généralisés et par les arbres CART, ainsi que l'importance des variables dans les forêts aléatoires. Les variables *usage du véhicule* et *fractionnement de la prime* ne sont sélectionnées qu'une seule fois sur les quatre par les arbres CART et n'ont pas une très grande importance dans les forêts aléatoires. Cela permet de confirmer que le choix de les retirer pour les modèles linéaires généralisés en raison de leur corrélation à d'autres variables explicatives n'est pas aberrant. Il est à noter que les modèles linéaires généralisés des fréquences conservent plus les variables liées à une résiliation que les arbres CART. De plus, ces variables n'ont que peu d'importance d'après les résultats des forêts aléatoires. Cependant, comme les variables de résiliation ou de condamnation n'ont que deux modalités : *OUI* et *NON*, l'analyse peut être biaisée si les permutations aléatoires ne permettent pas de modifier suffisamment les données. Cependant, cette hypothèse ne peut pas être vérifiée. Par ailleurs, deux variables ne sont jamais sélectionnées par les modèles linéaires généralisés et les arbres CART : une résiliation pour fausse déclaration et le nombre de sinistres corporels. Les variables décrivant le véhicule ont une grande importance dans tous les modèles et pour toutes les garanties. Les variables décrivant le conducteur (hors résiliation, condamnation et antécédents) sont globalement conservées pour tous les modèles, et plus particulièrement pour ceux de fréquence.

Variable	Bris de glace						Dommages au véhicule					
	Fréquence			Coût			Fréquence			Coût		
	GLM	CART	RF	GLM	CART	RF	GLM	CART	RF	GLM	CART	RF
Classe SRA	**1**	**2**	**1**	**1**	**1**	**1**	**2**	**2**	**2**	**1**	**4**	**1**
Groupe SRA	11	**5**	7	**4**	-	**2**	7	7	6	**2**	**1**	6
Âge du véhicule	7	**4**	**5**	**2**	**2**	**3**	8	6	**5**	**4**	**3**	**5**
Zone	6	6	11	-	-	7	11	8	10	6	-	9
Âge du conducteur	**2**	**3**	**4**	-	**4**	**4**	**3**	**5**	**3**	**3**	**2**	**2**
Ancienneté du permis	**3**	**1**	**2**	-	-	**5**	**5**	10	**4**	-	8	**3**
CSP	10	-	6	**5**	**3**	9	**4**	**4**	7	-	9	7
Usage	-	-	12	-	-	11	-	-	14	-	6	12
CRM	**5**	7	**3**	-	6	6	**1**	**1**	**1**	**5**	**5**	**4**
Fractionnement	-	-	9	-	-	8	-	9	12	-	-	10
Résiliation non-paiement	8	-	13	6	-	14	9	3	11	-	-	17
Résiliation fausse déclaration	-	-	19	-	-	18	-	-	18	-	-	19
Résiliation fréquence de sinistres	14	-	16	**3**	-	16	-	-	16	-	-	16
Suspension de permis	-	8	15	-	-	13	13	-	17	-	-	15
Annulation de permis	**4**	-	18	-	-	15	14	-	15	7	-	14
Condamnation pour alcoolémie	12	-	17	-	-	19	12	-	13	-	-	18
Nombre de sinistres corporels	-	-	14	-	-	17	-	-	19	-	-	13
Nombre de sinistres matériels responsables	13	-	10	-	-	12	10	-	9	-	7	8
Nombre de sinistres matériels non responsables	9	-	8	7	**5**	10	6	-	8	-	-	11

Tableau 11 Récapitulatif de l'importance des variables dans les différents modèles

5.2 Comparaison de la prime pure modélisée

L'objectif de l'assureur est que la somme des primes pures demandées aux assurés corresponde à la charge totale des sinistres. Afin de vérifier que les modèles n'engendrent pas des primes qui pourraient conduire l'assureur à être déficitaire, la prime pure des assurés est calculée pour les deux garanties étudiées. Cette prime correspond au produit des résultats des deux modèles : celui de fréquence et celui de coût : $\mathbb{E}[S] = \mathbb{E}[N] \times \mathbb{E}[C]$. Comme dans le cadre de la modélisation, la prime n'est calculée que si l'assuré a souscrit la garantie étudiée.

Le tableau ci-dessous présente la part de l'ensemble des primes modélisées par rapport à la charge totale réellement payée.

Modèle / Garantie	Modèle linéaire généralisé	Arbre CART	Forêt aléatoire
Bris de glace	99.8%	100,1%	100,4%
Dommages	98,7%	101,3%	100,6%

Tableau 12 Pourcentage de primes modélisées par rapport à la charge réelle de sinistralité

Pour l'ensemble des modèles, les primes pures demandées aux assurés correspondent globalement à la charge réelle payée par l'assureur, ce qui signifie que la tarification n'est pas surévaluée ni sous-évaluée. Pour les deux garanties étudiées, la somme des primes modélisées grâce aux modèles linéaires généralisés est inférieure à la charge des sinistres. Cependant, ceci n'est pas problématique puisque le chargement de sécurité, qui n'est pas calculé dans le cadre de ce mémoire, sert à couvrir les erreurs de modélisation.

Le fait d'obtenir des résultats proches de 100% montre que la modélisation ne pose pas de problème de rentabilité, néanmoins, cela ne signifie pas que le modèle est correct. En modélisant la prime de manière triviale, c'est-à-dire en appliquant la fréquence moyenne et le coût moyen à l'ensemble des assurés, la charge de sinistres serait exactement couverte. Toutefois, cette modélisation ne serait pas pour autant satisfaisante puisque la prime demandée aux assurés serait unique, ce qui engendrerait une antisélection et la faillite de l'assureur. Comme évoqué dans le premier chapitre, il est indispensable que l'assureur segmente les tarifs qu'il propose.

Afin de vérifier la capacité des modèles à discriminer de façon correcte les données, les courbes de Lorenz sont tracées. Elles permettent de visualiser la capacité de segmentation des modèles.

Mathématiquement, cela revient à considérer un échantillon de données $\left(x_{1,i}; \dots ; x_{k,i}; y_i\right)_{i=1}^{n}$ et à estimer la réponse $\widehat{y_i}$. Pour l'ensemble des données, les valeurs y_i sont classées par ordre croissant, ce qui permet d'obtenir un échantillon trié $\left(x_{1,j}; \dots ; x_{k,j}; \widehat{y_j}\right)_{j=1}^{n}$ (l'ordre des indices j est défini par le classement croissant des estimations $\left(\widehat{y_j}\right)_{j=1}^{n}$). L'abscisse des courbes de Lorenz correspond au pourcentage d'assurés classés par ordre croissant selon la variable réponse. L'axe des ordonnées mesure le pourcentage cumulé des réponses prédites sur les données de la base. Les propriétés de cette courbe sont la croissance et la convexité. Plus le courbe est convexe, plus la capacité de segmentation du modèle est élevée. Deux cas limites se présentent :

- Modèle unique : pire segmentation : indépendamment des variables explicatives, le modèle prédit systématiquement la réponse moyenne. Dans ce cas, la courbe est une droite reliant les deux extrémités.
- Modèle parfait : segmentation optimale : la courbe est tracée en utilisant la variable y_j réellement observée.

Plus la courbe est proche de la courbe optimale, meilleure est la segmentation du modèle.

Les courbes de Lorenz sont tracées pour la fréquence et le coût des deux garanties étudiées. Le modèle unique sans segmentation correspond à la droite noire en pointillés. Dans ce cas, la fréquence est répartie équitablement entre tous les assurés. Le modèle parfait est représenté par la courbe verte, les prédictions correspondent exactement aux observations. Les trois autres courbes indiquent la segmentation des trois modèles : en orange les modèles linéaires généralisés, en violet les arbres CART et en rouge les forêts aléatoires.

Tout d'abord, les courbes sont tracées pour les modèles de fréquence de la garantie Bris de glace.

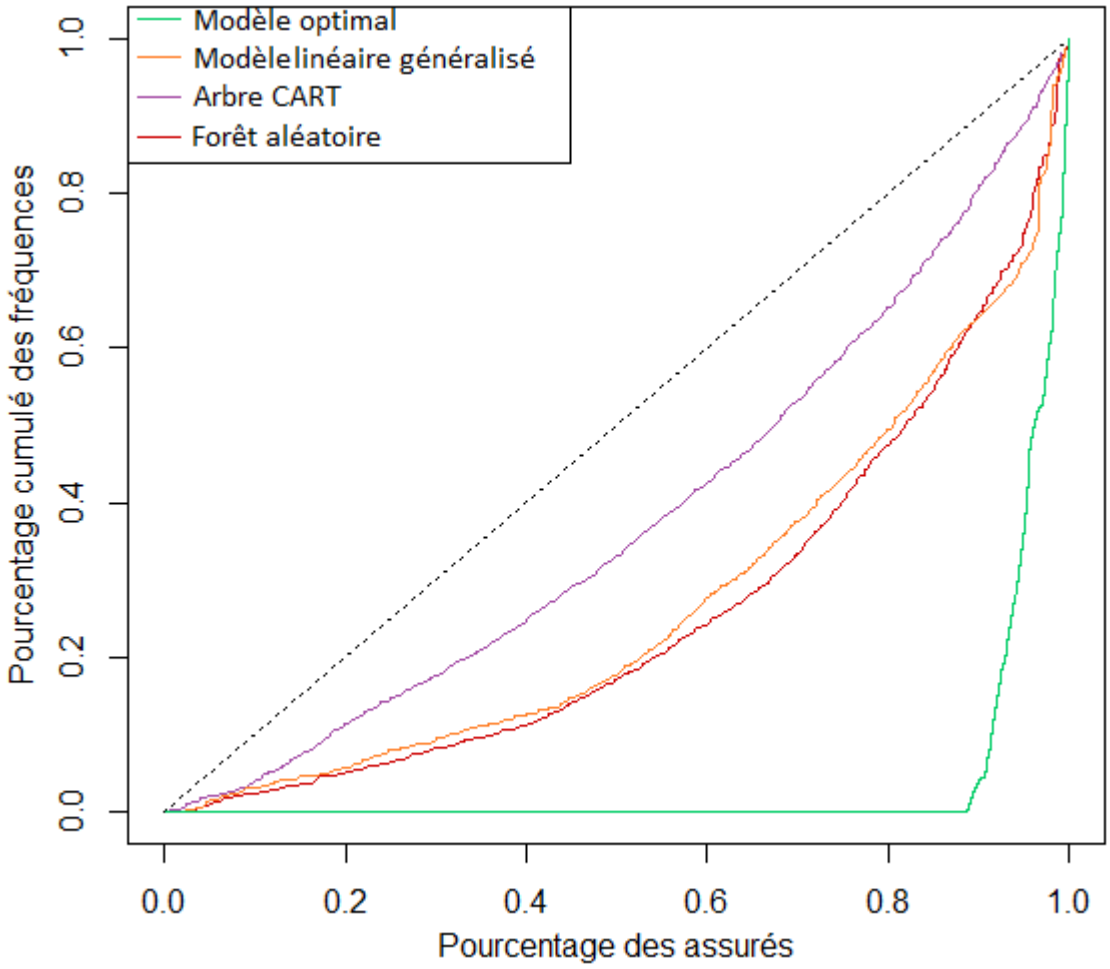

Figure 80 Courbe de Lorenz des modèles de fréquence de la garantie Bris de glace

La segmentation réalisée par le modèle linéaire généralisé et la forêt aléatoire modélisant la fréquence de Bris de glace sont globalement équivalentes. La courbe verte représente une segmentation parfaite, c'est-à-dire que la fréquence modélisée correspond exactement à la fréquence observée. Cela reviendrait à faire payer une prime aux assurés uniquement si ceux-ci ont un sinistre, or, cette situation va à l'encontre du principe de mutualisation des risques et l'assureur ne peut pas connaitre individuellement le risque de chaque assuré.

Après avoir étudié la segmentation de la fréquence de la garantie Bris de glace, celle du coût de cette même garantie est analysée

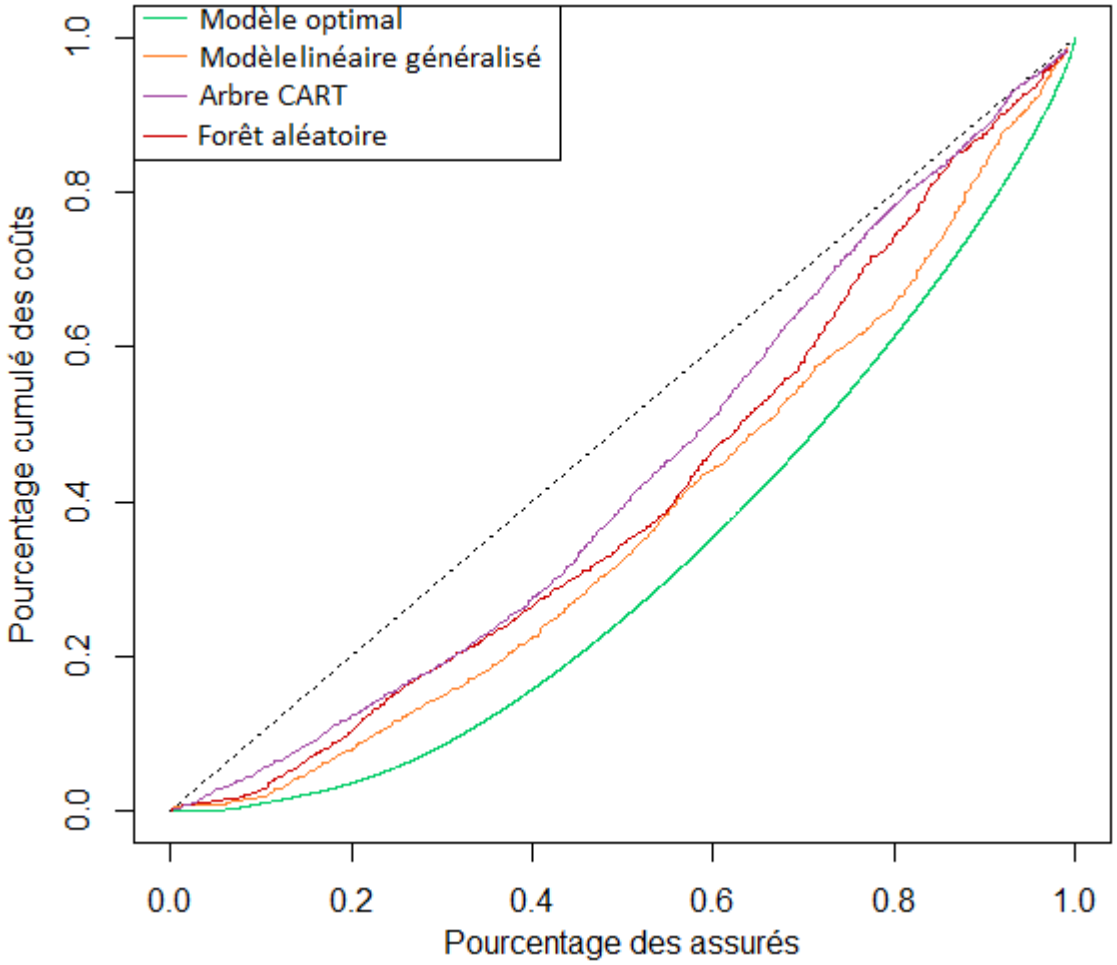

Figure 81 Courbe de Lorenz des modèles de coût de la garantie Bris de glace

Contrairement à la fréquence, en théorie, les coûts devraient être parfaitement identifiés individuellement puisqu'ils sont essentiellement liés au type de véhicule. Ceci est d'autant plus vrai que les coûts affectés à la garantie bris de glace sont essentiellement dus à des remplacements de pare-brise. Néanmoins, la facture d'un remplacement de pare-brise prend en compte le prix d'achat d'un nouveau pare-brise, mais aussi les frais de main-d'œuvre. Bien que l'assureur ait connaisse des caractéristiques du véhicule, le prix d'un pare-brise varie selon les fournisseurs et selon le type de pare-brise choisi par le propriétaire (pare-brise athermique, chauffant, teinté, acoustique, détecteur de pluie, etc…) et les frais de remplacement diffèrent selon les garages automobiles. Par ailleurs, une meilleure modélisation pourrait être effectuée si d'avantage de renseignements étaient présents dans les bases fournies par les courtiers, notamment des informations sur les pièces endommagées car le prix d'un pare-brise et d'une vitre ne sont pas similaires, par conséquent, ils ne devraient théoriquement pas être modélisés simultanément par la même loi. Bien que cette remarque ne soit pas applicable par manque d'informations dans les bases, les trois modèles proposent une segmentation correcte du coût des sinistres bris de glace. Le modèle linéaire généralisé obtient les meilleurs résultats.

Les courbes de Lorenz sont également tracées pour les modèles de la garantie Dommages.

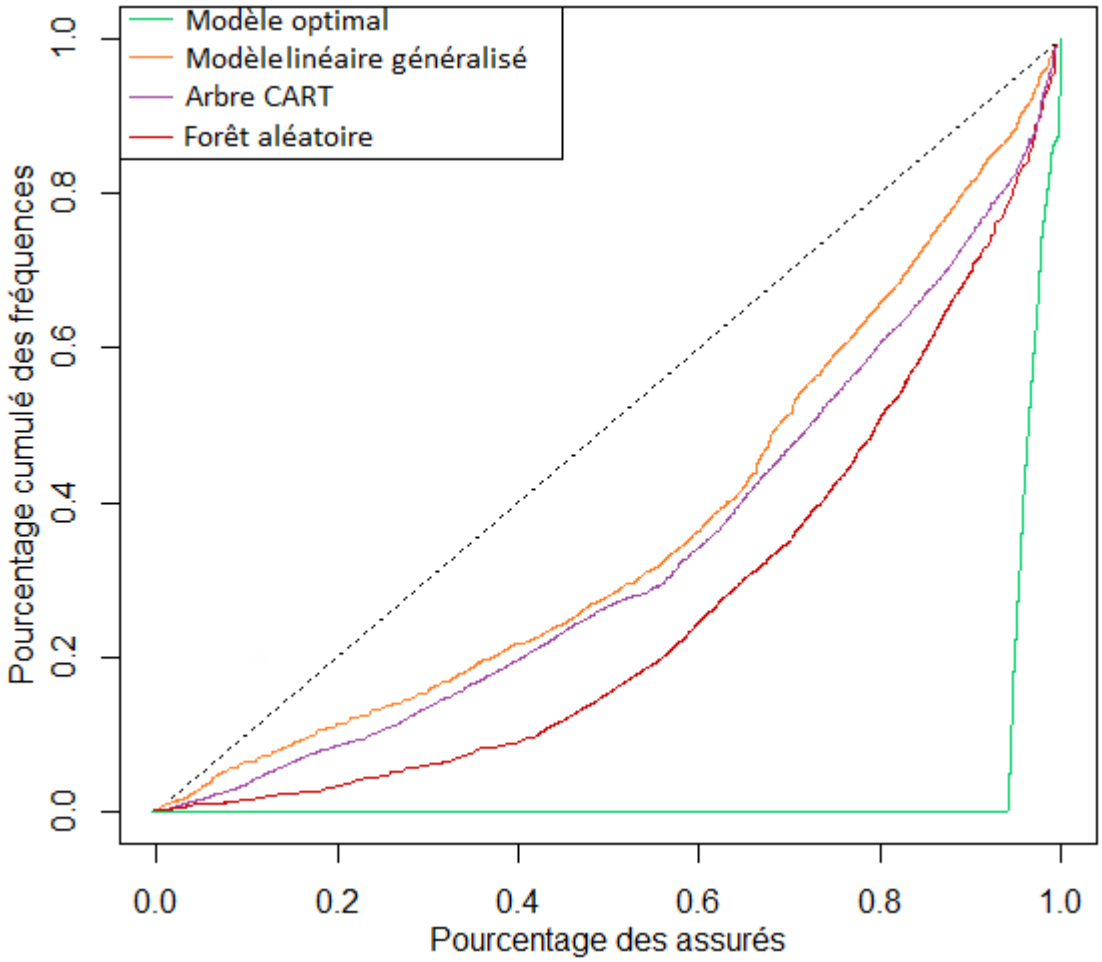

Figure 82 Courbe de Lorenz des modèles de fréquence de la garantie Dommages

Comme pour la modélisation de la fréquence des sinistres bris de glace, il est normal que les courbes de Lorenz des modèles s'éloignent de la courbe optimale car un assureur ne peut pas modéliser de fréquence nulle, même si, historiquement, un assuré n'a jamais eu de sinistre. La meilleure segmentation est celle de la forêt aléatoire, tandis que le modèle linéaire généralisé segmente assez mal les profils de risque des assurés.

Pour finir, les courbes de Lorenz sont tracées pour les modèles de coût des dommages.

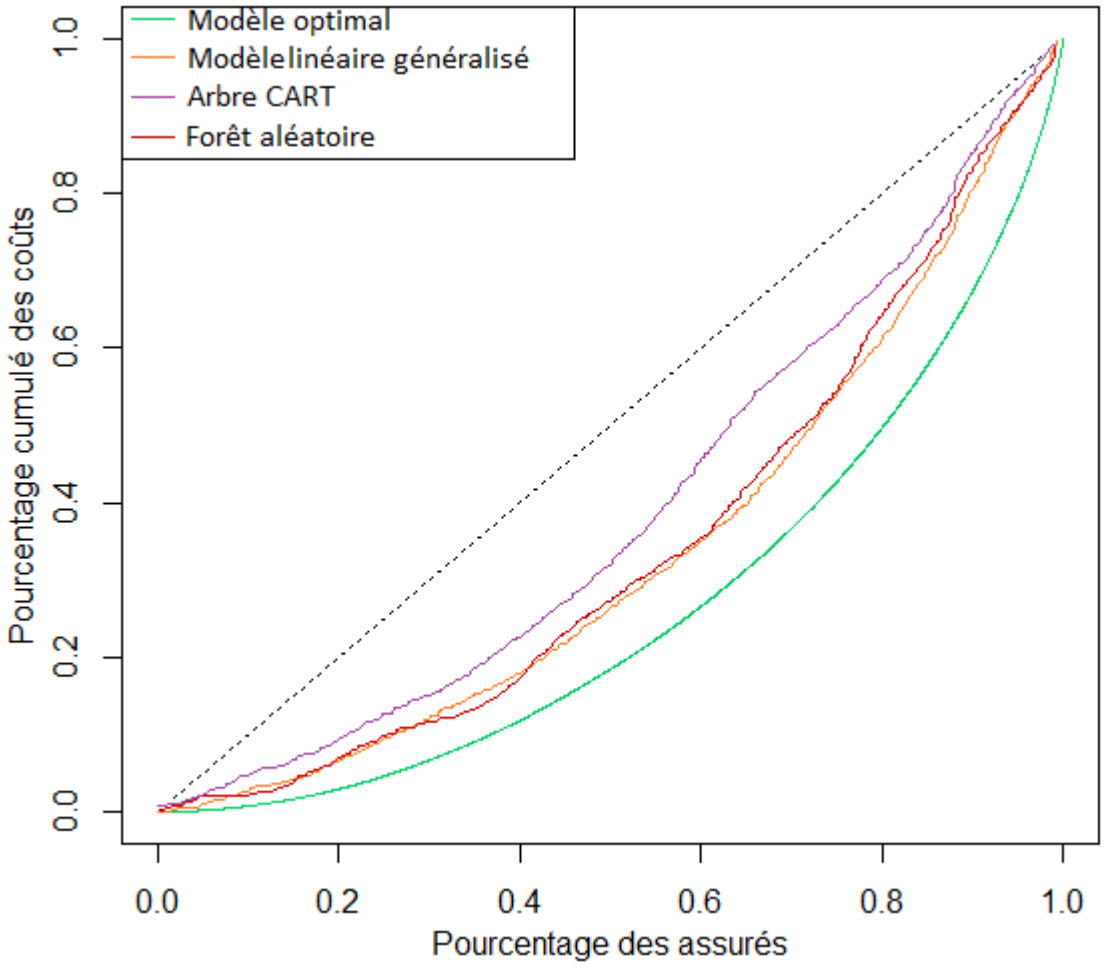

Figure 83 Courbe de Lorenz des modèles de coût de la garantie Dommages

La difficulté majeure concernant l'estimation du coût des sinistres dommages réside dans le fait que ces coûts dépendent du niveau de détérioration du véhicule. Ainsi, il est impossible pour l'assureur de connaitre exactement quelles pièces seront endommagées (portière, aile, moteur, roues, etc…) ni même si le véhicule sera considéré comme une épave par l'expert et devra donc être totalement remboursé. Malgré cette problématique et le manque d'informations supplémentaires dans les bases à disposition, les trois modèles proposent globalement une bonne segmentation du risque. Le modèle linéaire généralisé et la forêt aléatoire proposent même une segmentation assez similaire.

De manière générale, le modèle linéaire généralisé et la forêt aléatoire permettent la même forme de segmentation. Les arbres CART segmentent moins bien les risques, ce qui peut être expliqué par le peu de feuilles conservées après la phase d'élagage.

5.3 Évaluation du caractère prédictif des modèles

Afin de tester la qualité des modèles et dans le but de vérifier qu'ils se généralisent correctement, l'erreur est calculée sur la base test, c'est-à-dire sur la base totalement indépendante de la conception des modèles.

Pour juger la qualité globale des modèles, le critère RMSE (*Root mean square error* en anglais) est utilisé. Il est défini comme la racine de l'erreur quadratique moyenne :

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

Le meilleur modèle est celui minimisant le critère RMSE. En effet, le modèle minimisant le RMSE est celui dont les estimations sont les plus proches des valeurs observées. Il prend en compte à la fois le biais et la variance. Cependant, juger de la qualité du modèle est difficile en se basant seulement sur le RMSE. En effet, la valeur du RMSE varie fortement en fonction du type de données prédites (le RMSE des modèles de coût est nécessairement plus élevé que celui des modèles de fréquence car les fréquences correspondent à des petits nombres tandis que les coûts s'élèvent à plusieurs centaines voire milliers d'euros). C'est la raison pour laquelle il est intéressant d'étudier des indicateurs normalisés. Le critère RSR (*standard deviation ratio* en anglais) est utilisé. Il est défini comme la valeur du RMSE divisée par l'écart type des variables observées :

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

Plus la valeur de ce critère est proche de 0, meilleure est l'adéquation du modèle. En pratique, un modèle est jugé acceptable si la valeur du RSR est inférieur à 0,2. Il s'interprète comme le pourcentage de l'écart-type ne pouvant pas être expliqué par le modèle.

Pour chacune des garanties étudiées, les valeurs du RMSE et du RSR sont calculées et comparées pour les trois modèles étudiés dans ce mémoire : les modèles linéaires généralisés, les arbres CART et les forêts aléatoires. Le modèle trivial correspond à un modèle sans segmentation, c'est-à-dire à un modèle où la variable à expliquer est toujours associée à la moyenne.

Dans un premier temps, la comparaison est effectuée sur les modèles de fréquence de la garantie Bris de glace.

Modèle	RMSE	RSR	Pourcentage d'amélioration du RMSE par rapport au modèle trivial
Trivial	0,311	1	-
GLM	0,040	0,127	**87,14%**
CART	0,051	0,164	83,60%
RF	0,042	0,135	86,50%

Tableau 13 Comparaison des critères RMSE et RSR pour les modèles de fréquence de Bris de glace

Le modèle linéaire généralisé, l'arbre de régression CART et la forêt aléatoire sont jugés de bonne qualité car seulement 13 à 16% de l'écart-type n'est pas expliquée par les modèles. Pour la modélisation de la fréquence des bris de glace, le modèle linéaire généralisé obtient les meilleurs résultats. Ceci est en concordance avec l'analyse des résidus effectuée dans le troisième chapitre, qui présentait un modèle de bonne qualité. Par conséquent, pour modéliser la fréquence des bris de glace, la méthode « traditionnelle » obtient de meilleurs résultats que les méthodes de Machine Learning. De plus, le modèle se généralise correctement puisqu'il présente de bons résultats sur la base de test.

Après avoir conclu que le meilleur modèle de fréquence des bris de glace est le modèle linéaire généralisé, la même comparaison est effectuée sur les modèles de coût de cette garantie.

Modèle	RMSE	RSR	Pourcentage d'amélioration du RMSE par rapport au modèle trivial
Trivial	309,206	1	-
GLM	50.314	0,163	83,73%
CART	63,228	0,204	79,55%
RF	37.841	0,122	**87,76%**

Tableau 14 Comparaison des critères RMSE et RSR pour les modèles de coût de Bris de glace

Seuls le modèles linéaire généralisé et la forêt aléatoire permettent de modéliser convenablement le coût des sinistres et se généralisent correctement à d'autres données. En effet, le critère RSR de l'arbre CART est supérieur à 0,2, cela peut éventuellement être expliqué par le fait que l'arbre optimisé comporte assez peu de feuilles. La forêt aléatoire permet d'obtenir les meilleurs résultats.

Après avoir comparé les modèles pour la garantie bris de glace, les mêmes analyses sont effectuées pour la garantie dommages du véhicule.

Modèle	RMSE	RSR	Pourcentage d'amélioration du RMSE par rapport au modèle trivial
Trivial	0,218	1	-
GLM	0,044	0,201	79,91%
CART	0,029	0,133	86,70%
RF	0,021	0,097	**90,30%**

Tableau 15 Comparaison des critères RMSE et RSR pour les modèles de fréquence de Dommages

Comme l'a montré l'analyse des résidus de Pearson, le modèle linéaire généralisé présente quelques défauts de modélisation. Cela se constate également en analysant le critère RSR calculé sur la base test. Ainsi, le modèle linéaire généralisé présente les moins bons résultats. Les deux méthodes de Machine Learning diminuent le RMSE donc améliorent la prédiction. En particulier, la forêt aléatoire améliore le RMSE de 13% par rapport à celui obtenu à partir des données prédites grâce au modèle linéaire généralisé.

Enfin, la comparaison est effectuée par rapport aux modèles de coût de la garantie dommages.

Modèle	RMSE	RSR	Pourcentage d'amélioration du RMSE par rapport au modèle trivial
Trivial	2295,774	1	-
GLM	194,635	0,085	91,52%
CART	408,705	0,134	86,55%
RF	180,565	0,079	**92,13%**

Tableau 16 Comparaison des critères RMSE et RSR pour les modèles de coût de Dommages

Les trois modèles présentent des bons résultats généralisables à la base de test. La forêt aléatoire présente des résultats légèrement meilleurs que le modèle linéaire généralisé.

L'analyse effectuée dans cette dernière partie permet de conclure que les trois modèles modélisent correctement la fréquence et le coût des sinistres Bris de glace et Dommages au véhicule. Les modèles linéaires généralisés et les forêts aléatoires présentent les meilleurs résultats.

Ce mémoire a donc permis de modéliser la prime pure de deux garanties en utilisant plusieurs méthodes de tarification : les modèles linéaires généralisés et deux méthodes de Machine Learning, les arbres de régression CART et les forêts aléatoires. Ces modélisations permettent à la Direction des Partenariats de proposer des primes segmentées et adaptées aux risques des assurés. Les forêts aléatoires proposent des résultats qui se généralisent mieux aux données de la base test. Cependant, lorsqu'une loi s'adapte bien à la distribution de la variable réponse, les modèles linéaires généralisés proposent également une bonne modélisation des risques. Ils ont de plus l'avantage d'être facilement interprétables et de pouvoir être implémentés informatiquement facilement car ils fournissent une prime de référence et des coefficients multiplicatifs. Par ailleurs, le temps de calcul des modèles linéaires généralisés est inférieur à celui nécessaire aux méthodes de Machine Learning.

Conclusion

Ce mémoire a permis de comparer différentes méthodes de modélisation de la prime pure de deux garanties d'un produit « risque aggravé » en assurance automobile, une méthode classique : le modèle linéaire généralisé et deux méthodes de Machine Learning : les arbres de régression CART et les forêts aléatoires.

L'objectif de la tarification est de proposer des primes adaptées aux assurés en fonction de leur risque, tout en mutualisant l'ensemble des risques.

L'analyse des variables conservées par les différents modèles a permis de montrer que les modèles sélectionnent globalement les mêmes variables explicatives.

Les modèles linéaires généralisés, bien que largement utilisés en actuariat, présentent certaines limites, notamment lorsqu'aucune loi ne permet de modéliser correctement les données. Dans ce cas, les modèles de Machine Learning proposent une approche plus adaptée, ce qui est le cas pour la modélisation du coût des sinistres *Dommages tous accidents*. Lorsque la loi s'ajuste correctement à la distribution de la variable d'intérêt, les modèles linéaires généralisés et les forêts aléatoires permettent d'obtenir des résultats assez similaires, tant au niveau de la segmentation des assurés qu'au niveau des prédictions. Bien que les arbres CART présentent les moins bons résultats, leur facilité d'interprétation offre aux assureurs une meilleure compréhension des risques en proposant une vision synthétique des variables explicatives, ce qui enrichit leur stratégie tarifaire.

Les remarques quant au choix du meilleur modèle sont à nuancer car les résultats obtenus dépendent fortement de la base de données. Les estimations étant intrinsèquement liées aux données, le choix du meilleur modèle ne peut pas être généralisé à tous les produits d'assurance.

Les méthodes de Machine Learning sont assez peu utilisées par les actuaires en raison de leur difficulté d'interprétation par rapport aux modèles linéaires généralisés. Il pourrait être intéressant de combiner les différentes méthodes afin de viser une modélisation optimale, par exemple en déterminant la significativité des variables explicatives et les regroupements optimaux grâce à des arbres CART et en appliquant ces variables aux modèles linéaires généralisés. Le modèle linéaire généralisé reste toutefois un très bon estimateur des risques et est généralement largement utilisé par les actuaires à la fois pour sa facilité d'interprétation, mais aussi pour son développement informatique pratique. Il est en effet plus commode, notamment pour des courtiers, d'implémenter dans les logiciels des coefficients multiplicatifs que des méthodes de Machine Learning.

Des travaux ont démontré que d'autres méthodes de Machine Learning obtiennent des résultats satisfaisants. Parmi elles, les méthodes de Boosting, les réseaux de neurones ou les SVR (Support Vector Machine for Regression) peuvent être cités.

Bibliographie

BELLINA R. (2014). Méthodes d'apprentissage appliquées à la tarification non-vie. Mémoire ISFA.

BIERNAT E., LUTZ M. (2015, Librairie Eyrolles). Data science : fondamentaux et études de cas.

BOLLACHE V. (2015). Tarification automobile : des évolutions apportées dans le cadre de la loi Hamon. Mémoire ISFA.

BOUCHTA G. (2017). Mise en œuvre de méthodes innovantes de tarification. Mémoire Dauphine.

BOUTTIER F. (2015) Modélisation de la prime pure de la garantie RC Auto. Mémoire CEA.

DENUIT M., CHARPENTIER A. (2004, Economica). Mathématiques de l'assurance non-vie, Tome 1 et Tome 2

DISSARD G. (2013). Évolutions tarifaires des produits d'assurance automobile et IRD. Mémoire DUAS.

FAUGERE N. (2016). Tarification d'un portefeuille automobile sur le marché luxembourgeois. Mémoire CEA.

LUO Y. (2016). Amélioration de la modélisation de sinistres graves à l'aide d'une approche d'apprentissage. Mémoire ISFA

MAITI M. (2012). Modélisation de la prime pure en assurance automobile. Mémoire ISUP.

NGUYEN-HUU T. (2014). Barème d'évaluation des sinistres corporels en automobile. Mémoire CNAM.

PARTRAT C. (2004, Librairie Eyrolles). Assurance non-vie : modélisation, simulation.

PLANCHET F., MISERAY A. (2017). Tarification IARD, Introduction aux techniques avancées. Cours ISFA

RINDER P. (2016). Modélisation des sinistres corporels en automobile. Mémoire ENSAE.

SURU A. (2018). Principes de l'assurance dommages. Cours de Master 2 Actuariat à l'Université Paris Dauphine.

TINKEU NGATCHOU J. (2017). Modélisation de la garantie incendie multirisques habitation, méthode paramétrique vs méthodes non-paramétriques. Mémoire ISFA.

Annexes

Annexe 1 : Présentation de Generali **109**

Annexe 2 : Abréviations **111**

Annexe 3 : Démonstrations **112**

Annexe 1 : Présentation de Generali

A Le groupe Generali

Fondé le 26 décembre 1831 à Triste, en Italie, Generali se démarque dès son origine par sa vocation internationale. Dès 1832, Generali s'implante en France. Par ailleurs, la compagnie propose des garanties dans toutes les branches de l'assurance, ce qui est une pratique inhabituelle pour l'époque.

Aujourd'hui, le Groupe Generali est l'un des principaux assureurs au monde et se positionne au troisième rang des plus gros assureurs en Europe. En effet, en 2018, son chiffre d'affaires s'élève à 66,7 milliards d'euros et son résultat net à 2,3 milliards d'euros. Implanté dans plus de 60 pays, les 71 000 collaborateurs de Generali proposent des solutions d'assurance à plus de 61 millions de clients.

B Generali France

Generali est implanté en France depuis 1832. Au fil des années, la filiale française a acquis différentes sociétés, qui ont été regroupées progressivement pour aboutir en 2006 à la création d'une entreprise unique : Generali France. Il s'agit du troisième marché de Generali dans le monde, après l'Italie et l'Allemagne.

En 2018, le chiffre d'affaires de Generali France s'élève à 12,3 milliards d'euros (se décomposant en 9,5 milliards d'euros pour l'activité Vie et 2,7 milliards d'euros pour l'activité Biens et responsabilité), plaçant ainsi la compagnie au 7° rang sur le marché français, assurant 7 millions de clients particuliers et 800 000 entreprises.

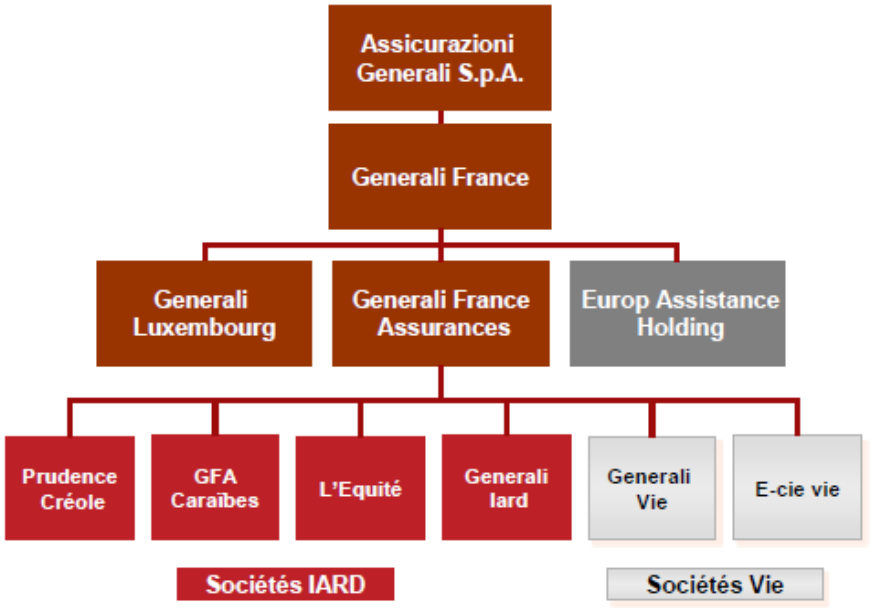

Figure 84 : Organisation des entités de Generali France

C L'Équité et la Direction des Partenariats

La société l'Afrique Française est fondée en 1887. Elle est renommée l'Équité en 1957 et rejoint le groupe Generali quelques années plus tard, en 1976.

Depuis 2011, grâce à la création de la Direction des Partenariats, l'Équité est devenue le spécialiste du groupe Generali en France des solutions sur mesure à destination des particuliers et conçues spécifiquement pour des partenaires, qui peuvent être des courtiers, des courtiers grossistes ou des courtiers spécialisés. L'Équité propose ainsi des solutions d'assurance de biens et responsabilité (auto, moto, habitation, assurance affinitaire (vol, casse et panne des téléphones portables et des vélos à assistance électrique, extension de garantie), loyers impayés, animaux de compagnie), d'assurance de personnes (santé, emprunteur, prévoyance, obsèques) et de protection juridique.

Les partenaires doivent remplir l'ensemble des conditions requises par le Code des assurances (Livre V : Intermédiaires d'assurance) et être inscrits au Registre unique des intermédiaires en assurance, banque et finance (ORIAS).

La spécificité de l'Équité repose sur le principe de délégation. Ainsi, les produits sont vendus sous la marque des courtiers et l'Équité a un rôle de porteur de risques. Cela signifie que la gestion des contrats, l'encaissement des primes, la gestion des sinistres et la communication d'informations aux assurés sont totalement laissés à la charge des partenaires. Pour sa part, la Direction des Partenariats étudie le développement de nouveaux partenariats et l'élaboration de solutions techniques, gère les documents contractuels et veille au pilotage et au suivi des partenariats. Ces différentes tâches sont détaillées dans le diagramme ci-dessous.

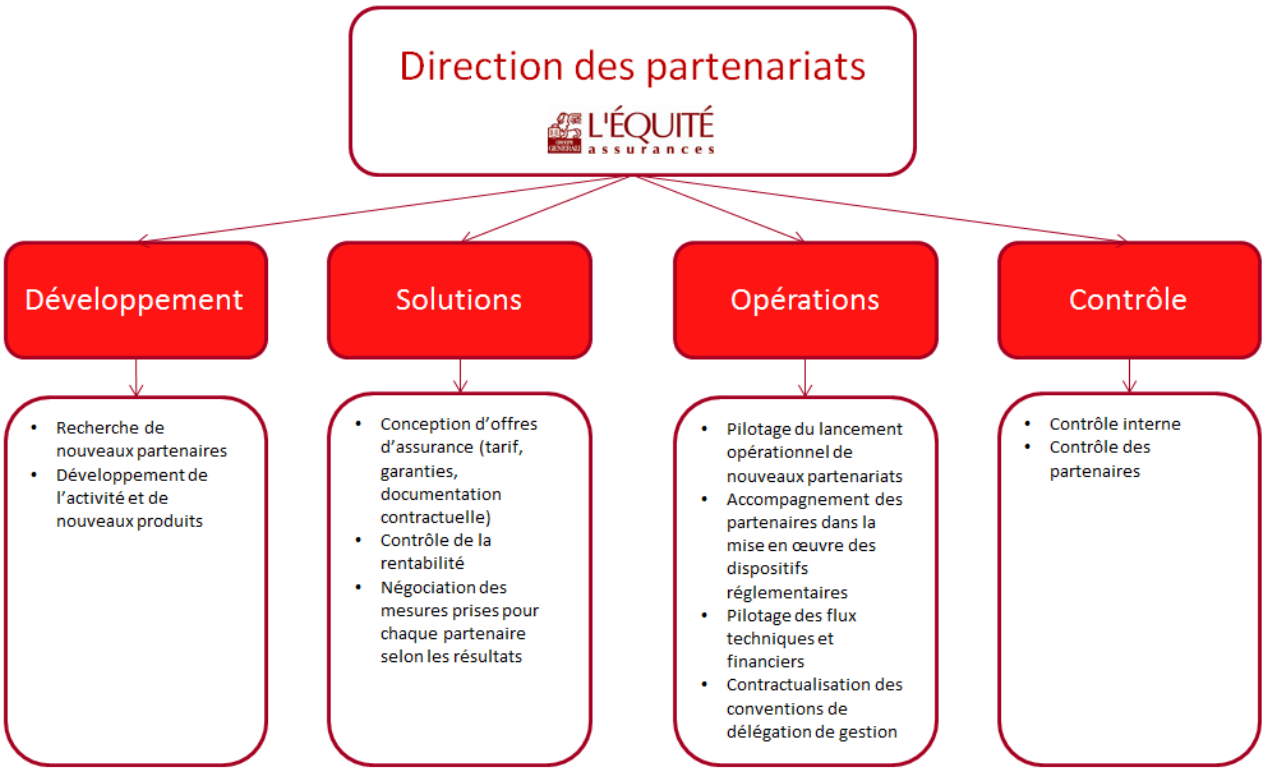

Figure 85 : Organisation de la Direction des Partenariats

Annexe 2 : Abréviations

AIC : Critère d'information d'Akaike

AIPP : Atteinte permanente à l'intégrité physique et psychique

BCT : Bureau central de tarification

BDG : Bris de glace

BIC : Critère d'information bayésien

CART : Arbres de classification et de régression

CRM : Coefficient de réduction-majoration (« bonus-malus »)

DDA : Directive sur la distribution d'assurance

DPRSA : Défense pénale et recours après un accident de la circulation

DTA : Dommage tous accidents

FFA : Fédération Française de l'Assurance

FGAO : Fonds de garantie des assurances obligatoires de dommages

FVA : Fichier des véhicules assurés

GLM : Modèle linéaire généralisé

GPC : Garantie personnelle du conducteur

IDA : Indemnisation directe de l'assuré

IRCA : Convention d'indemnisation et de recours corporel automobile

IRSA : Convention d'indemnisation directe de l'assuré et de recours entre sociétés d'assurance automobile

LOM : Projet de loi d'orientation des mobilités

OOB : Échantillon Out-of-Bag

ORIAS : Organisme pour le registre des intermédiaires en assurance

RC : Responsabilité civile

RF : Forêt aléatoire

RGPD : Règlement général sur la protection des données

RMSE : Racine de l'erreur quadratique moyenne

SRA : Sécurité et réparation automobiles

SIV : Système d'immatriculation des véhicules

VI : Vol-incendie

Annexe 3 : Démonstrations

A Loi des grands nombres

Théorème : Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires réelles indépendantes et identiquement distribuées et définies sur le même espace de probabilité. Soit $\mu = \mathbb{E}[X_i]$ et $\sigma = \mathbb{V}[X_i] < \infty$. Alors, la convergence suivante est vérifiée :

$$\frac{1}{n}\sum_{i=1}^{n} X_i \underset{n\to\infty}{\overset{\mathbb{P}}{\to}} \mu.$$

Démonstration* :

La démonstration est donnée pour des variables aléatoires X_i positives. Cependant, elle est valable dans le cas général en décomposant $X_i = {X_i}^+ - {X_i}^-$ avec ${X_i}^+ \geq 0$ et ${X_i}^- \geq 0$.

Soient :

- $S_n = \sum_{i=1}^{n} X_i$
- $Y_i = X_i \mathbb{1}_{X_i \leq i}$
- $S_n^* = \sum_{i=1}^{n} Y_i$
- $\varepsilon > 0$
- $\alpha > 1$
- $k_n = \lfloor \alpha^n \rfloor$: partie entière de α^n
- C : constante positive variant d'étape en étape
- $\varphi = \mathbb{P}_{X_i}$: loi de X_i

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\frac{S_{k_n}^* - \mathbb{E}[S_{k_n}^*]}{k_n}\right) \leq C \sum_{n=1}^{\infty} \frac{\mathbb{V}[S_{k_n}^*]}{k_n^2} = C \sum_{n=1}^{\infty} \frac{1}{k_n^2} \sum_{i=1}^{k_n} \mathbb{V}[Y_i] \leq C \sum_{i=1}^{\infty} \frac{\mathbb{E}[Y_i^2]}{i^2} = C \sum_{i=1}^{\infty} \frac{1}{i^2} \int_0^i x^2 \varphi(dx)$$

$$= C \sum_{i=1}^{\infty} \frac{1}{i^2} \sum_{k=0}^{i-1} \int_k^{k+1} x^2 \varphi(dx) \leq C \sum_{k=0}^{\infty} \frac{1}{k+1} \int_k^{k+1} x^2 \varphi(dx) \leq C \sum_{k=0}^{\infty} \int_k^{k+1} x\, \varphi(dx) = C\mathbb{E}[X_1] < \infty$$

Or, l'espérance de $[X_1]$ est définie par :

$$\mathbb{E}[X_1] = \lim_{n\to\infty} \int_0^n x\, \varphi(dx) = \lim_{n\to\infty} \mathbb{E}[Y_n] = \lim_{n\to\infty} \frac{\mathbb{E}[S_{k_n}^*]}{k_n}$$

D'après le lemme de Borel-Cantelli : $\frac{S_{k_n}^*}{k_n} \to \mathbb{E}[X_i]\ ps$

Or,

$$\sum_{n=1}^{\infty} \mathbb{P}(Y_n \neq X_n) = \sum_{n=1}^{\infty} \mathbb{P}(X_n > n) = \sum_{n=1}^{\infty} \int_n^{\infty} \varphi(dx) = \sum_{n=1}^{\infty} \sum_{i=n}^{\infty} \int_i^{i+1} \varphi(dx)$$

$$= \sum_{i=1}^{\infty} i \int_i^{i+1} \varphi(dx) \leq \mathbb{E}[X_1] < \infty$$

D'après Borel-Cantelli, $\mathbb{P}(Y_n \neq X_n) = 0$ donc $\lim_{n\to\infty} \frac{S_n}{k_n} = \mathbb{E}[X_1]\ ps$

---

* Cette démonstration a été faite par N. Etemadi en 1981.

Pour $n \geq 1$, soit $m(n) \geq 0$ tel que $k_{m(n)} \leq n \leq k_{m(n)+1}$. La suite S_n est croissante donc :

$$\liminf_{n\to\infty} \frac{S_n}{n} \geq \liminf_{n\to\infty} \frac{S_{k_{m(n)}}}{k_{m(n)}} \frac{k_{m(n)}}{k_{m(n)}+1} \geq \frac{1}{\alpha} \lim_{n\to\infty} \frac{S_{k_{m(n)}}}{k_{m(n)}} = \frac{1}{\alpha}\ \mathbb{E}[X_1] < \infty$$

De même : $\limsup_{n\to\infty} \frac{S_n}{n} \leq \alpha \mathbb{E}[X_1]$

Comme $\alpha > 1$ est arbitraire, la démonstration est conclue en prenant $\alpha \to 1$.

B Calcul de la prime pure

Propriété : Soit S la charge totale des sinistres et π la somme des primes versées par chaque assuré, ayant eu un sinistre ou non. La prime pure permettant de minimiser l'erreur quadratique $\mathbb{E}[(S-\pi)^2]$ est donnée par $\pi = \mathbb{E}[S]$.

Démonstration :

Par définition, $\mathbb{E}[(S-\pi)^2] = \mathbb{V}[S-\pi] + (\mathbb{E}[S-\pi])^2$.

Or, π est déterministe donc $\mathbb{V}[S-\pi] = \mathbb{V}[S]$.

Par conséquent, minimiser $\mathbb{E}[(S-\pi)^2]$ revient à minimiser $(\mathbb{E}[S-\pi])^2 > 0$, c'est-à-dire à minimiser $\mathbb{E}[S-\pi]$.

$\mathbb{E}[S-\pi] = 0 \Leftrightarrow \mathbb{E}[S] - \mathbb{E}[\pi] = 0 \Leftrightarrow \mathbb{E}[S] = \pi$

D'où $\pi = \mathbb{E}[S]$

C. Espérance et variance de la charge des sinistres

Propriété :

Dans le modèle collectif, la charge de sinistres est définie par :

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

avec :

- S : charge totale des sinsitres.
- N : nombre totale de sinsitres survenus.
- C_i : montant du i^{ème} sinistre.
- $(C_i)_{1\leq i\leq N}$ forment un échantillon de variables aléatoires indépendantes et identiquement distribuées de loi parente C.
- $(C_i)_{1\leq i\leq N}$ et N sont des variables aléatoires indépendantes.

L'espérance et la variance de cette charge S valent :

$$\mathbb{E}[S] = \mathbb{E}[N] \times \mathbb{E}[C]$$

$$\mathbb{V}[S] = \mathbb{V}[C] \times \mathbb{E}[N] + \mathbb{V}[N] \times \mathbb{E}[C]^2$$

Démonstration :

$$\mathbb{E}[S] = \mathbb{E}\left[\sum_{i=1}^{N} C_i\right] = \mathbb{E}\left[\mathbb{E}\left[\sum_{i=1}^{N} C_i \middle| N\right]\right] = \mathbb{E}\left[\mathbb{E}\left[\sum_{i\geq 1} C_i \mathbb{1}_{i\leq N} \middle| N\right]\right] = \mathbb{E}\left[\sum_{i\geq 1} \mathbb{1}_{i\leq N}\, \mathbb{E}[C_i|N]\right] = \mathbb{E}\left[\sum_{i\geq 1} \mathbb{1}_{i\leq N}\, \mathbb{E}[C]\right]$$

$$= \mathbb{E}[C] \times \mathbb{E}\left[\sum_{i\geq 1} \mathbb{1}_{i\leq N}\right] = \mathbb{E}[C] \times \mathbb{E}[N] = \mathbb{E}[N] \times \mathbb{E}[C]$$

Cela démontre donc que l'espérance de la charge des sinistres est égale au produit des espérances du nombre de sinistres et du coût d'un sinistre.

D'après la formule de la variance conditionnelle, $\mathbb{V}[S] = \mathbb{E}\big[\mathbb{V}[S|N]\big] + \mathbb{V}\big[\mathbb{E}[S|N]\big]$

Ces deux termes sont calculés séparément :

$$\mathbb{E}\big[\mathbb{V}[S|N]\big] = \sum_{n\geq 0} \mathbb{P}(N=n)\, \mathbb{V}[S|N=n] = \sum_{n\geq 0} \mathbb{P}(N=n)\, \mathbb{V}\left[\sum_{i=1}^{n} C_i\right] = \sum_{n\geq 0} \mathbb{P}(N=n) \times n \times \mathbb{V}[C]$$

$$= \mathbb{V}[C] \times \sum_{n\geq 0} \mathbb{P}(N=n) \times n = \mathbb{V}[C] \times \mathbb{E}[N]$$

$$\mathbb{V}\big[\mathbb{E}[S|N]\big] = \mathbb{V}\left[\mathbb{E}\left[\sum_{i\geq 1} C_i \mathbb{1}_{i\leq N} \middle| N\right]\right] = \mathbb{V}\left[\sum_{i\geq 1} \mathbb{E}[C_i \mathbb{1}_{i\leq N}|N]\right] = \mathbb{V}\left[\sum_{i\geq 1} \mathbb{1}_{i\leq N}\, \mathbb{E}[C_i|N]\right] = \mathbb{V}\left[\sum_{i\geq 1} \mathbb{1}_{i\leq N}\, \mathbb{E}[C]\right]$$

$$= (\mathbb{E}[C])^2 \times \mathbb{V}\left[\sum_{i\geq 1} \mathbb{1}_{i\leq N}\right] = (\mathbb{E}[C])^2 \times \mathbb{V}[N]$$

La démonstration est achevée en additionnant les deux résultats.

D Espérance et variance d'une variable aléatoire appartenant à la famille exponentielle

Définition : Famille exponentielle :

La loi d'une variable aléatoire Y appartient à la famille exponentielle si sa fonction de densité (par rapport à la mesure de comptage sur $\mathbb{N}$ ou à la mesure de Lebesgue sur $\mathbb{R}$) s'écrit sous la forme :

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

avec :

- $a(\phi)$: fonction non nulle
- $b(\theta)$: fonction de classe $\mathcal{C}^2$ et connexe
- $c(y,\phi)$: fonction
- θ : paramètre naturel car il est lié aux deux premiers moments de la loi
- ϕ : paramètre d'échelle

Propriété :

L'espérance et la variance de Y vérifient les relations suivantes :

$$\begin{cases} \mathbb{E}[Y] = \mu = b'(\theta) \\ \mathbb{V}[Y] = \sigma^2 = b''(\theta) \times a(\phi) \end{cases}$$

Démonstration :

f est une densité, donc par définition, $\forall\, \theta, \phi$: $\int_A f_{\theta,\phi}\,(y)dy = 1 = u(\theta;\phi)$

Or, d'après le lemme de Fatou :

$\frac{\partial u}{\partial \theta}(\theta;\phi) = 0 = \lim\limits_{h\to 0} \frac{u(\theta+h;\phi)-u(\theta;\phi)}{h}$

$= \lim\limits_{h\to 0} \frac{\int_A f_{\theta+h,\phi}(y)dy - \int_A f_{\theta,\phi}(y)dy}{h} = \liminf\limits_{h\to 0} \frac{\int_A f_{\theta+h,\phi}(y)dy - \int_A f_{\theta,\phi}(y)dy}{h} = \limsup\limits_{h\to 0} \frac{\int_A f_{\theta+h,\phi}(y)dy - \int_A f_{\theta,\phi}(y)dy}{h}$

Par conséquent, la dérivée partielle de $f_{\theta,\phi}$ par rapport à θ existe.

L'espérance de Y est obtenue en utilisant la dérivée partielle de $f_{\theta,\phi}$ et la variance en utilisant la dérivée seconde.

En repartant de la définition d'une famille exponentielle, il est possible de calculer cette dérivée partielle : $\frac{\partial f_{\theta,\phi}}{\partial \theta}(y) = \frac{y-b'(\theta)}{a(\phi)} f_{\theta,\phi}(y)$

Donc $\int_A \frac{\partial f_{\theta,\phi}}{\partial \theta}(y)dy = 0 = \frac{1}{a(\phi)}\left[\int_A y f_{\theta,\phi}(y)dy - b'(\theta)\int_A f_{\theta,\phi}(y)dy\right]$

Donc $0 = \int_A y f_{\theta,\phi}(y)dy - b'(\theta)$

Donc par définition de l'espérance, $\mathbb{E}[Y] = b'(\theta)$

La dérivée seconde de $f_{\theta,\phi}$ s'écrit : $\frac{\partial^2 f_{\theta,\phi}}{\partial \theta^2}(y) = \frac{-b''(\theta)}{a(\phi)} f_{\theta,\phi}(y) + \left(\frac{y-b'(\theta)}{a(\phi)}\right)^2 f_{\theta,\phi}(y)$

Donc $\int_A \frac{\partial^2 f_{\theta,\phi}}{\partial \theta^2}(y)\,dy = 0 = \frac{-b''(\theta)}{a(\phi)}\int_A f_{\theta,\phi}(y)dy + \frac{1}{(a(\phi))^2}\int_A \left(y - b'(\theta)\right)^2 f_{\theta,\phi}(y)dy$

Donc $0 = \frac{1}{a(\phi)}\left[-b''(\theta) + \frac{1}{a(\phi)}\int_A (y-\mathbb{E}[Y])^2 f_{\theta,\phi}(y)dy\right]$

Donc $0 = -b''(\theta) \times a(\phi) + \int_A (y-\mathbb{E}[Y])^2 f_{\theta,\phi}(y)dy$

D'où, par définition de la variance, $\mathbb{V}[Y] = b''(\theta) \times a(\phi)$

E Tests asymptotiques

Propriété :

L'estimateur du maximum de vraisemblance du θ est noté $\widehat{\theta_n}$. $\mathbf{I}(\theta) < \infty$ désigne l'information de Fisher. Les hypothèses variables sont supposées indépendantes et le modèle régulier.

L'estimateur $\widehat{\theta_n}$ vérifie la propriété asymptotique suivante :

$$\sqrt{n}\left(\widehat{\theta_n} - \theta\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0;\mathbf{I}^{-1}(\theta))$$

Démonstration :

Soit $h_n(X_1;\dots;X_n;\ \theta) = \frac{1}{n}\sum_{j=1}^{n} \frac{\partial \ln f(X_j;\theta)}{\partial \theta}$

$\left(\frac{\partial \ln f(X_j;\theta)}{\partial \theta}\right)_n$ est une suite indépendante et identiquement distribuée de variables aléatoires centrées et de variance $\mathbf{I}(\theta)$. Par application du théorème central limite, la convergence suivante est obtenue :

$$\sqrt{n}h_n(X_1;\dots;X_n;\ \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0;\mathbf{I}(\theta))$$

Puis, en développant h_n en utilisant le théorème des accroissements finis :

$$h_n(X_1; \dots; X_n;\ t) = h_n(X_1; \dots; X_n;\ \theta) + (t-\theta)K_n(X_1; \dots; X_n;\ \theta_n^*)$$

Avec :

- $\theta_n^* \in]t; \theta[$
- $D_2(X_j; u) = \frac{\partial^2 \ln f(X_j;\theta)}{\partial \theta^2}$
- $K_n(X_1; \dots; X_n;\ u) = \frac{1}{n}\sum_{j=1}^{n} D_2(X_j; u)$

En remplaçant t par $\widehat{\theta_n}$, les résultats précédents deviennent :

$0 = h_n(X_1; \dots; X_n;\ \theta) + (\widehat{\theta_n} - \theta)K_n(X_1; \dots; X_n;\ \theta_n^*) \quad (a.)$

Par application de la loi des grands nombres à la suite i.i.d. $\frac{\partial^2 \ln f(X_j;\theta)}{\partial \theta^2}$, la convergence suivante est vérifiée presque-sûrement :

$$K_n(X_1; \dots; X_n;\ \theta) = \frac{1}{n}\sum_{j=1}^{n} D_2(X_j; \theta) \to -\mathbf{I}(\theta)$$

Par ailleurs,

$$|K_n(X_1; \dots; X_n;\ \theta_n^*) - K_n(X_1; \dots; X_n;\ \theta)| \leq \frac{1}{n}\sum_{j=1}^{n} |D_2(X_j; \theta_n^*) - D_2(X_j; \theta)| \leq \sup_x |D_2(x; \theta_n^*) - D_2(x; \theta)|$$

Comme $\widehat{\theta_n}$ converge presque sûrement vers θ, il en est de même pour θ_n^*. En supposant la continuité uniforme en x de la dérivée seconde $D_2(x; \theta)$, il en résulte : $K_n(X_1; \dots; X_n;\ \theta_n^*)) \to -\mathbf{I}(\theta) \quad (b.)$

En utilisant les résultats $(a.)$ et $(b.)$ et le lemme de Slutsky, la propriété est démontrée.

Propriété : Soit Y_n une suite de vecteurs aléatoires telle que$Y_n \xrightarrow{\mathcal{L}} Y$ et $Y \sim \mathcal{N}(0; \Sigma)$.

Alors $Y_n' \Sigma^{-1} Y_n \xrightarrow{\mathcal{L}} Z \sim \mathcal{X}^2(k)$ où $k = \dim(\Sigma)$.

Démonstration :

Cette démonstration est disponible dans le cours de statistique mathématique d'Alain Monfort.

Propriété : L'estimateur du maximum de vraisemblance vérifie la convergence :

$$n(\widehat{\theta_n} - \theta)' \mathbf{I}(\theta)(\widehat{\theta_n} - \theta) \xrightarrow{\mathcal{L}} \mathcal{X}^2(k)$$

Démonstration :

Cette démonstration est directe en utilisant les deux théorèmes précédents, et en posant $Y_n = \sqrt{n}(\widehat{\theta_n} - \theta)$.

Illustration graphique des trois tests :

Test de Wald : $T_W = n(\widehat{\theta_n} - \theta)' \mathbf{I}(\theta)(\widehat{\theta_n} - \theta)$.

Le test de Wald permet d'étudier la différence entre $\widehat{\theta_n}$ et θ. Asymptotiquement, en utilisant la propriété précédente, cette statistique tend vers une loi du $\mathcal{X}^2$ si l'estimation est correcte.

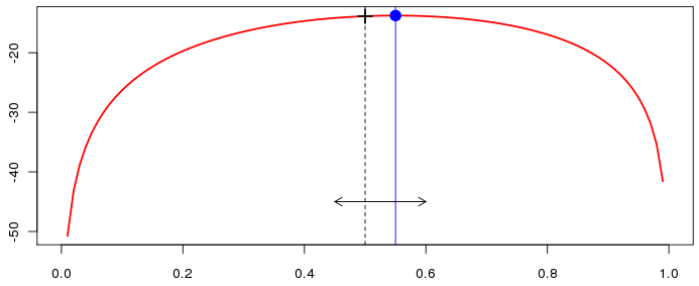

Figure 86 Exemple graphique de log-vraisemblance pour illustrer le test de Wald

Test de rapport de vraisemblance : $T_R = 2\left(\mathcal{LL}(\widehat{\theta_n}) - \mathcal{LL}(\theta)\right)$.

Ce test a pour but d'étudier les valeurs de la log-vraisemblance. Si l'estimation est correcte, la statistique de test converge vers une loi du $\mathcal{X}^2$.

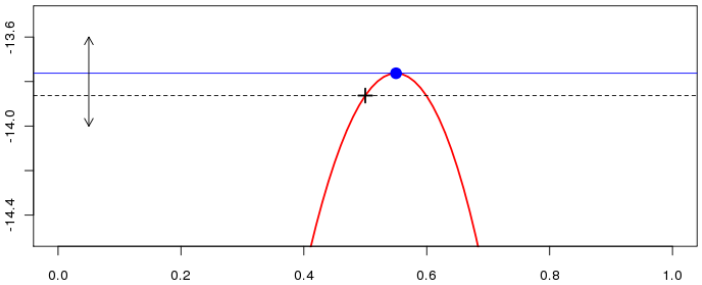

Figure 87 Exemple graphique de log-vraisemblance pour illustrer le test du rapport de vraisemblance

Test du score : Soit $S_\theta(X) = \frac{\partial}{\partial\theta}\mathcal{LL}(\theta)$

$$T_S = n\left(\frac{1}{n}\sum_{i=1}^{n} S'_\theta(X_i)\right)' \mathbf{I}^{-1}(\widehat{\theta_n})\left(\frac{1}{n}\sum_{i=1}^{n} S'_\theta(X_i)\right)$$

Le test du score permet d'étudier la pente de la log-vraisemblance en θ. Celle-ci doit être « petite » si l'estimation est correcte. Asymptotiquement, la statistique de test du score tend vers une loi du $\mathcal{X}^2$.

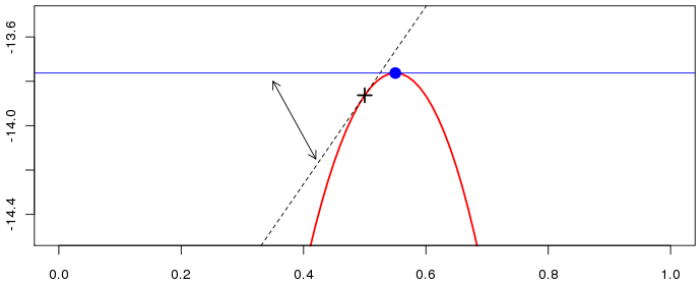

Figure 88 Exemple graphique de log-vraisemblance pour illustrer le test du score

F Maximisation de la log-vraisemblance pour déterminer les coefficients du modèle linéaire généralisé

Théorème : La solution du maximum de vraisemblance se calcule par approximations successives par la formule :

$$\beta^{k+1} = \left(X'W^kX\right)^{-1}X'W^kz^k$$

avec :

- X *:matrice des données explicatives à* n *lignes et* k *colonnes*
- $W^k = \mathrm{diag}\left(\frac{\left((g^{-1})'\right)^2(X_i^T\beta^k)}{b''(\theta_i)}\right)$ *: matrice des poids*
- $z^k = X\beta^k + \mathrm{diag}\left(\frac{1}{(g^{-1})'(X_i^T\beta^k)}\right)\left(y - b'(\theta)\right)$

Démonstration :

Avant tout, il est nécessaire de calculer la log-vraisemblance d'un individu i de l'échantillon.

Soit $l(y_i;\theta;\phi) = \ln\left(f_{\theta,\phi}(y_i)\right) = \frac{y_i\theta - b(\theta)}{a(\phi)} + c(y_i,\phi)$ et soit $\mu_i = \mathbb{E}[Y_i] = b'(\theta_i) = g^{-1}(X_i^T\beta)$

De plus, $\frac{\partial l}{\partial \beta_j} = \frac{\partial l}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j}$.

Or, $\frac{\partial l}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)}$.

De plus, $\mathbb{V}[Y] = b''(\theta) \times a(\phi)$ donc $\mathbb{V}[Y|X = X_i] = b''(\theta_i) \times a(\phi) = \frac{\partial\left(b'(\theta_i)\right)}{\partial \theta_i} \times a(\phi) = \frac{\partial \mu_i}{\partial \theta_i} \times a(\phi)$.

Donc $\frac{\partial \theta_i}{\partial \mu_i} = \frac{a(\phi)}{\mathbb{V}[Y|X=X_i]}$

De plus, $\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial\left(g^{-1}(X_i^T\beta)\right)}{\partial \beta_j} = X_{i,j}(g^{-1})'(X_i^T\beta)$

Par conséquent, $\frac{\partial l}{\partial \beta_j} = \frac{y_i - b'(\theta_i)}{a(\phi)} \times \frac{a(\phi)}{\mathbb{V}[Y|X=X_i]} \times X_{i,j}(g^{-1})'(X_i^T\beta) = \frac{y_i - b'(\theta_i)}{\mathbb{V}[Y|X=X_i]} \times X_{i,j}(g^{-1})'(X_i^T\beta)$.

Donc $\frac{\partial l}{\partial \beta_j} = \frac{y_i - \mu_i}{b''(\theta_i)\times a(\phi)} \times X_{i,j}\frac{\left((g^{-1})'\right)^2(X_i^T\beta)}{(g^{-1})'(X_i^T\beta)} = \frac{y_i - \mu_i}{a(\phi)} \times \frac{\left((g^{-1})'\right)^2(X_i^T\beta)}{b''(\theta_i)} \times \frac{X_{i,j}}{(g^{-1})'(X_i^T\beta)}$

En notant $W_i = \frac{\left((g^{-1})'\right)^2(X_i^T\beta)}{b''(\theta_i)}$ le poids associé à l'observation i, le résultat précédent s'écrit :

$$\frac{\partial l}{\partial \beta_j} = \frac{y_i - \mu_i}{a(\phi)} \times \frac{W_i}{(g^{-1})'(X_i^T\beta)} \times X_{i,j}$$

Cependant, μ_i dépend de β et il n'existe pas de solution explicite permettant d'annuler la dérivée. Il faut donc utiliser une méthode numérique : l'algorithme de Newton-Raphson ou celui de Fisher.

La démonstration suivante est effectuée en utilisant la méthode de Fisher.

Un point β^0 est initialisé. Puis, l'étape $k+1$ se construit à partir de l'étape k en cherchant à augmenter la vraisemblance : $\mathcal{L}(\beta^k) < \mathcal{L}(\beta^{k+1})$.

La formule itérative de l'algorithme de Fisher est : $\beta^{k+1} = \beta^k - \mathbb{E}[\nabla^2\mathcal{L}_k]^{-1}\left(\nabla\mathcal{L}(\beta^k)\right)$.

$(\nabla\mathcal{L}^2)_{j,k} = \frac{\partial^2\mathcal{L}}{\partial\beta_j\partial\beta_k}$ et $\mathcal{L} = \sum_{i=1}^n l(y_i;\theta;\phi)$.

D'après les calculs précédents, $\frac{\partial^2 l}{\partial\beta_j\partial\beta_k} = \frac{\partial}{\partial\beta_k}\left(\frac{y_i - \mu_i}{a(\phi)}\frac{W_i}{(g^{-1})'(X_i^T\beta)}X_{i,j}\right)$

Donc $\frac{\partial^2 l}{\partial\beta_j\partial\beta_k} = -\frac{1}{a(\phi)}\frac{W_i}{(g^{-1})'(X_i^T\beta)}X_{i,j}\frac{\partial\mu_i}{\partial\beta_k} + \frac{y_i - \mu_i}{a(\phi)}X_{i,j}\frac{\partial}{\partial\beta_k}\left(\frac{W_i}{(g^{-1})'(X_i^T\beta)}\right)$

Le second terme est noté $K = \frac{y_i - \mu_i}{a(\phi)} X_{i,j} \frac{\partial}{\partial \beta_k}\left(\frac{W_i}{(g^{-1})'(X_i^T\beta)}\right)$

Puis, en remplaçant $\frac{\partial \mu_i}{\partial \beta_k}$ par sa valeur calculée précédemment :

$$\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} = -\frac{1}{a(\phi)} \frac{W_i}{(g^{-1})'(X_i^T\beta)} X_{i,j} X_{i,k} (g^{-1})'\left(X_i^T\beta\right) + K = -\frac{1}{a(\phi)} W_i X_{i,j} X_{i,k} + K$$

Or, $\mu_i = \mathbb{E}[Y_i]$ donc $\mathbb{E}[K] = 0$ donc $-\mathbb{E}\left[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}\right] = \frac{1}{a(\phi)} W_i X_{i,j} X_{i,k}$.

Par conséquent, $-\mathbb{E}[\nabla^2 \mathcal{L}_k] = \sum_{i=1}^n -\mathbb{E}\left[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}\right] = \sum_{i=1}^n \begin{pmatrix} \ddots & \vdots & \iddots \\ \cdots & \frac{1}{a(\phi)} W_i X_{i,j} X_{i,k} & \cdots \\ \iddots & \vdots & \ddots \end{pmatrix} = \frac{1}{a(\phi)} X' W^k X$ où $W^k = \mathrm{diag}(W_i)$ évalué à l'étape k.

Par conséquent, $-\mathbb{E}[\nabla^2 \mathcal{L}_k]^{-1} = a(\phi)\left(X' W^k X\right)^{-1}$

Donc $\beta^{k+1} = \beta^k + a(\phi)\left(X' W^k X\right)^{-1}\left(\nabla \mathcal{L}\left(\beta^k\right)\right)$

Or, $a(\phi)\nabla \mathcal{L}\left(\beta^k\right) = \begin{pmatrix} \vdots \\ \sum_{i=1}^n \frac{\partial l}{\partial \beta_k} \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ \sum_{i=1}^n (y_i - \mu_i) X_{i,k} \frac{W_i}{(g^{-1})'(X_i^T\beta)} \\ \vdots \end{pmatrix} = X' \mathrm{diag}\left(\frac{1}{(g^{-1})'(X_i^T\beta^k)}\right) W^k (y - \mu)$

Donc finalement,

$$\beta^{k+1} = \beta^k - \mathbb{E}[\nabla^2 \mathcal{L}_k]^{-1}\left(\nabla \mathcal{L}\left(\beta^k\right)\right)$$

$$\beta^{k+1} = \beta^k + \left(X' W^k X\right)^{-1} X' \mathrm{diag}\left(\frac{1}{(g^{-1})'(X_i^T\beta^k)}\right) W^k (y - \mu)$$

$$\beta^{k+1} = \left(X' W^k X\right)^{-1}\left(\left(X' W^k X\right)\beta^k + X' \mathrm{diag}\left(\frac{1}{(g^{-1})'(X_i^T\beta^k)}\right) W^k (y - \mu)\right)$$

$$\beta^{k+1} = \left(X' W^k X\right)^{-1} X' W^k \left(X\beta^k + \mathrm{diag}\left(\frac{1}{(g^{-1})'(X_i^T\beta^k)}\right)(y - \mu)\right)$$

En reprenant la notation donnée dans l'énoncé du théorème : $z^k = X\beta^k + \mathrm{diag}\left(\frac{1}{(g^{-1})'(X_i^T\beta^k)}\right)(y - \mu)$, le résultat est démontré :

$$\beta^{k+1} = \left(X' W^k X\right)^{-1} X' W^k z^k$$