Mémoire présenté le :

pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA et l'admission à l'Institut des Actuaires

Par : Lucas Mietton

Titre : Analyse et optimisation d'un produit automobile dans un contexte économique concurrentiel

Confidentialité : ☒ NON ☐ (Durée : ☐ 1 an ☐ 2 ans)
Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de l'Institut des Actuaires — *Signature*

L. Laurent .....

M. Louchard .....

C. Pigeon

A. Bontoux .....

Membres présents du jury de l'ISFA

.....

.....

.....

Entreprise : ACTUELIA
Nom : BOUKOBZA Frank
Signature :

Directeur de mémoire en entreprise :
Nom : BOUKOBZA Frank
Signature :

Invité :
Nom :
Signature :

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

Signature du responsable entreprise

Signature du candidat

Mietton

Résumé

Mots clés : ***Assurance automobile, réassurance, rentabilité, tarification, clustering, GLM, CART, Random Forest, Gradient Boosting, Solvabilité 2.***

L'objectif de ce mémoire consiste à comprendre les difficultés auxquelles font face les compagnies d'assurance de petite et de moyenne taille, au sein d'un environnement hautement concurrentiel qu'est le domaine de l'assurance automobile. Prenant comme cas pratique un groupe d'assurance, les travaux sont de ce fait axés autour de la problématique de la rentabilité pour le réassureur du contrat lié aux garanties automobiles. L'idée est qu'en optimisant le portefeuille et en identifiant les mauvais profils du portefeuille dans un premier temps, et dans un second temps en revoyant la méthode de tarification, la rentabilité amélioré du assureur permettra de transmettre une part de résultat et de solutionner le problème.

Il a d'abord été convenu de préparer les données pour leur utilisation dans le cadre d'analyses bivariées. En traitant les problèmes concernant la qualité des données, des premiers résultats sont obtenus, permettant notamment de préciser les profils mal tarifés. Ceux-ci ont permis de tirer les premières conclusions quant aux variables apparemment intéressantes à traiter par la suite lors de la mise en place de nouvelles méthodes de tarification pour le groupe.

Une phase de retraitement des données a ensuite été effectuée. Des clustering suivant la méthode de Ward ont permis d'obtenir des informations sur des regroupements par la suite retraités au cas par cas. L'étude des corrélations préparatoire à l'application d'un Generalized Linear Model (GLM) a également été effectuée. Enfin, une des variables n'étant pas totalement complétée, des méthodes de Data Science telles que les Classification And Regression Trees (CART) et les Random Forest ont été déployées afin de résoudre le problème.

Les travaux de tarification ont par la suite été abordés dans l'optique d'utiliser un GLM. Après la visualisation des résidus fournis par les sorties du modèle pour le modèle de fréquence, il a été question de faire appel aux mêmes méthodes de Data Science que pour la complétion de la variable afin d'obtenir un modèle de coût adapté. Dans ce contexte, un algorithme de Gradient Boosting Model (GBM) est utilisé pour modéliser les coûts de chaque garantie.

Ces travaux ont mené à l'analyse des résultats après l'application du modèle de coût-fréquence obtenu. Faisant écho aux résultats discutables obtenus par l'analyse des résidus, le GLM semblait non-adapté pour la modélisation de la fréquence, en raison du peu de données mises à disposition. Une méthode de Random Forest est donc finalement appliquée afin d'obtenir une modélisation plus réaliste.

Par la suite, les résultats sont intégrés au programme de réassurance, et les rentabilités du réassureur et de l'assureur sont discutées. Une méthode est mise en place afin de constater si une meilleure répartition de la prime totale entre les garanties est possible pour améliorer le résultat global pour le réassureur.

Enfin, les résultats de la modélisation sont intégrés au Business Plan du groupe dans le cadre du pilier 2 de la Directive Solvabilité 2 afin d'en observer les impacts éventuels sur le résultat et le ratio de solvabilité de l'entreprise.

Abstract

***Keywords :** Automobile insurance, reinsurance, profitability, pricing, clustering, GLM, CART, Random Forest, Gradient Boosting, Solvency 2.*

This paper is aimed at understanding the issues the little to medium sized insurance companies are facing within the automobile insurance field, a highly concurrential market. Therefore, as the practical case of an insurance group is taken, the work is focused around the reinsurer profitability problematic for the contract based on the automobile warranties. By optimizing the portfolio and by identifying its bad profiles in a first part, and by reconsidering the pricing method in a second part, the increased profitability for the insurer will allow a transfer of a part of the result, which will theoretically solve the issue.

First, the data needed to be prepared to be used in bivariate analysis. By solving the issues about data quality, some results are obtained, notably allowing to precise the characteristics of the badly priced profiles. These allowed a draw of the first conclusions concerning the interesting variables to be treated during the implementation of the new pricing method for the group.

Then, the data were reprocessed. Some clustering based on the Ward method were obtained, and then were studied case by case. The correlations study preceding the Generalized Linear Model (GLM) application was also made. Finally, as a variable was not fully completed, some Data Science methods such as CART (Classification And Regression Trees) and Random Forest have been deployed in order to solve the issue.

As a GLM was used, some pricing work were tackled. Following the visualization of the residuals obtained with the frequency model method, the same Data Science methods as the one used for the variable completion were called to obtain an adapted cost model. In this context, a Gradient Boosting Model (GBM) algorithm is used to model each warranty's cost.

This work leads to the results analysis after the cost-frequency model was applied. Echoing the questionable results obtained with the residuals analysis, the GLM seemed also unadapted for the frequency modelling, due to the lack of data at disposal. A Random Forest method is finally applied in order to get a more realistic modelling.

Then, the results are incorporated within the reinsurance program, and the reinsurer's and insurer's profitabilities are discussed. A method is set up to see if a better premium distribution among the warranties is possible to improve the reinsurer global result.

Finally, the modelling results are implemented in the group's Business Plan as part of the pillar 2 from the Solvency 2 directive, to observe the potential effects on the group's result and solvency ratio.

Remerciements

Dans un premier temps, je souhaite adresser mes remerciements à David Fitouchi et Frank Boukobza pour la confiance qu'il m'ont témoignée, et pour m'avoir permis de prendre part à l'aventure du Conseil au sein d'Actuelia.

Un mémoire est un travail principalement personnel, mais il est certain que sans appui, sa bonne réalisation est compromise. C'est pour cela que je tiens à adresser des remerciements tout particuliers à Victoire Piat, Consultante Senior du cabine Actuelia, qui m'a aidé dans mes recherches et m'a aiguillé lorsque j'avais des interrogations sur certains sujets. Ses conseils, ainsi que ceux de Frank Boukobza, m'ont permis de bien tenir le cap durant l'intégralité de la réalisation de ce mémoire.

Si l'aide apportée a son importance, l'ambiance de travail est essentielle. À cet effet, j'aimerais remercier l'entièreté de l'équipe d'Actuelia. Pour tout ce que j'ai appris et partagé avec vous, merci.

J'adresse également mes remerciements à l'ISFA, à ma tutrice, Ying Jiao, ainsi qu'à Xavier Milhaud et à Esterina Masiello, qui ont tous les trois su m'éclairer quant aux problèmes d'ordre technique que j'ai rencontrés tout au long de ce mémoire.

Enfin, merci à ma famille et notamment à mes parents, qui m'ont soutenu durant toutes ces années d'étude et tout mon parcours actuariel.

Sommaire

Introduction

Le secteur automobile, en matière d'assurance, présente des caractéristiques qui en font un pilier dans le paysage assurantiel français. En effet, l'obligation pour les particuliers de souscrire à une police d'assurance automobile donne lieu à un marché où les sociétés se font une concurrence extrêmement marquée, faisant naître un ratio combiné fortement dégradé comparé aux autres secteurs d'assurance.

Afin de bien comprendre et de pouvoir proposer des résultats destinés à résoudre ces problématiques, un mode d'action est proposé. Dans un premier temps, il s'agira de comprendre les problématiques générales liées à l'assurance automobile, et d'en tirer les enseignements nécessaires pour définir une procédure adaptée destinée à répondre à l'ensemble des besoins du groupe étudié.

Dans un deuxième temps, il sera question de nettoyer la base de données mise à disposition et de l'utiliser une première fois pour analyser certaines variables jugées pertinentes pour l'explication de la mauvaise rentabilité du portefeuille. Il s'agit d'analyses statistiques descriptives qui s'avèrent primordiales lorsque l'objectif est de déterminer avec précisions les problèmes auxquels est confronté le groupe.

Enfin, alors que des problèmes liés à la qualité des données seront discutés, les travaux concernant la tarification en elle-même seront présentés. La méthode classique GLM est étudiée dans un premier temps pour le modèle de fréquence, mais s'avèrera peu efficace et fragile au niveau des hypothèses, et sera ainsi supplantée par une modélisation par random forest. Le modèle de coût a fait l'objet de travaux supplémentaires via la mise en place de plusieurs techniques de Data Science, et une méthode de gradient boosting sera finalement retenue. Une fois le modèle correctement calibré, une simulation a posteriori de la réassurance sera appliquée afin d'obtenir les résultats qui permettront d'amener à la conclusion.

Cette dernière fera état des différences observées avec la précédente méthode de tarification, et mettra en lumière les impacts d'une redistribution du total de primes sur la réassurance. Enfin, une contextualisation des résultats dans le contexte règlementaire actuel sera effectuée afin d'en discerner les aboutissants.

De ce fait, ce mémoire a pour horizon principal de fournir des solutions dans un cadre pratique par la mise en application de méthode actuarielles usuelles. Par ces procédés, il conviendra d'exprimer une réponse claire à la problématique posée via les résultats obtenus, et de s'intéresser aux élargissements pouvant découler de la conclusion ainsi exprimée.

1 Partie I : Paysage du secteur automobile et horizons de l'étude

1.1 Présentation de l'environnement de l'assurance automobile

1.1.1 Quelques généralités

Le marché de l'assurance automobile représente une partie très importante du marché assurantiel français puisque le chiffre d'affaires de l'assurance automobile représente 10% de l'ensemble des cotisations de l'assurance française en 2017 [1]. Le marché des particuliers représente un chiffre d'affaire de 19,2 milliards d'euros en 2015. Historiquement, et ce indépendamment de l'année, les portefeuilles d'assurance automobile n'ont jamais montré une rentabilité exacerbée. En effet les ratios combinés de ce domaine dépassent généralement les 100%. Les marges proviennent généralement des autres produits d'assurance vendus en parallèles qui eux présenteront davantage de rentabilité pour la société.

Cette rentabilité peu présente s'explique par l'obligation d'assurance automobile, ce qui induit l'existence d'un marché dit hyperconcurrentiel. Les prix vont donc s'aligner au minimum, et par une logique de micro-économie vont s'approcher du coût réel du sinistre, ne laissant ainsi pas de place à une marge commerciale. Cela implique pour les entreprises d'assurance une nécessité à innover et, notamment, à proposer des garanties annexes. En raison de ce milieu de concurrence, la tarification des produits doit être la plus précise, et représentative de la réalité du risque porté par la société. La segmentation résultante, puisque les profils assurés ne sont pas homogènes, doit également être suffisamment travaillée pour ne pas donner lieu à des disparités en terme d'efficacité du tarif. Il est par ailleurs à noter que la segmentation du profil assuré permet de déterminer la prime adaptée, processus primordial puisque cela amène à minimiser les risques d'anti-sélection et d'aléa-moral.

L'anti-sélection, ou sélection adverse, est en effet un problème inhérent au domaine de l'assurance. Dans le contexte automobile, il est d'autant plus important de se prémunir contre ce genre de phénomène que le moindre écart de tarification peut donner lieu à une rentabilité particulièrement dégradée. L'anti-sélection consiste en la souscription, d'un assuré, à un contrat qui ne lui correspond pas puisqu'il est sous-tarifé par rapport à son profil de risque. Celui-ci a en effet intérêt à chercher une sous-tarification n'entrant pas en adéquation avec son profil réel, ce qui induira une prime insuffisante pour l'entreprise. L'aléa moral, d'autre part, est la propension de l'assuré à modifier son comportement en raison de l'assurance à laquelle il a souscrit. Là encore, il s'agit de faire en sorte de limiter au maximum cette adaptation de comportement en précisant la grille tarifaire au maximum. Ces deux phénomènes, bien connus dans le milieu assurantiel, sont donc les principaux écueils qu'il faut chercher à éviter par des études de tarifications.

Les assurances automobiles se décomposent généralement en plusieurs garanties, l'objectif est donc de visualiser quelles garanties présentent usuellement plus de sinistres que les autres. La répartition des sinistres par garantie, en 2015, la charge des prestations est concentrée dans les garanties Responsabilité Civile Corporelles et Dommages Toute Automobile. Le graphique suivant présente plus en détail la répartition de cette charge :

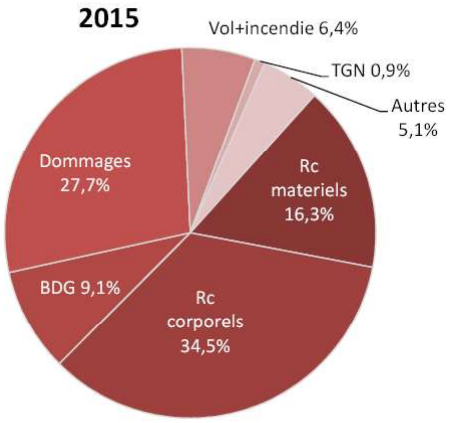

FIGURE 1 – Répartition de la charge des sinistres (hors assistance automobile)

1. https ://www.ffa-assurance.fr/etudes-et-chiffres-cles/assurances-de-biens-et-de-responsabilite-donnees-cles-par-annee

En 2017, il a été observé que les primes moyennes d'assurances automobile se sont élevées à 595 euros par an [2]. La hausse des primes n'est néanmoins pas homogène sur toute la France, mais cela montre que les assureurs souhaitent actuellement pallier à la hausse des coûts des sinistres. L'augmentation de la vente des voitures neuves est d'ailleurs un facteur explicatif de cette hausse des coûts de sinistres, engendrant de ce fait une hausse des primes. Le marché de l'assurance automobile ne cesse de faire face à des coûts en constante augmentation et n'a donc d'autre choix que d'utiliser le levier de la prime demandée au client.

1.1.2 Caractérisation du marché

Par ces premières informations, il devient clair que la problématique se pose d'autant plus pour les sociétés d'assurance de moyenne ou petite taille. Les coûts pour les leaders du marché notamment sont plus réduits : ayant plus de contrats à honorer, les accords avec les professionnels chargés des réparations, avec les experts pour estimer le coût du sinistre et tous les acteurs de l'automobile sont bien plus avantageux. Les moyennes et petites cédantes se retrouvent donc avec des coûts plus importants, sans opportunité toutefois d'augmenter les primes en conséquences, ce qui crée des difficultés grandissantes quant à la rentabilité obtenue.

Concernant les acteurs principaux du marché, la répartition est présentée dans le tableau suivant :

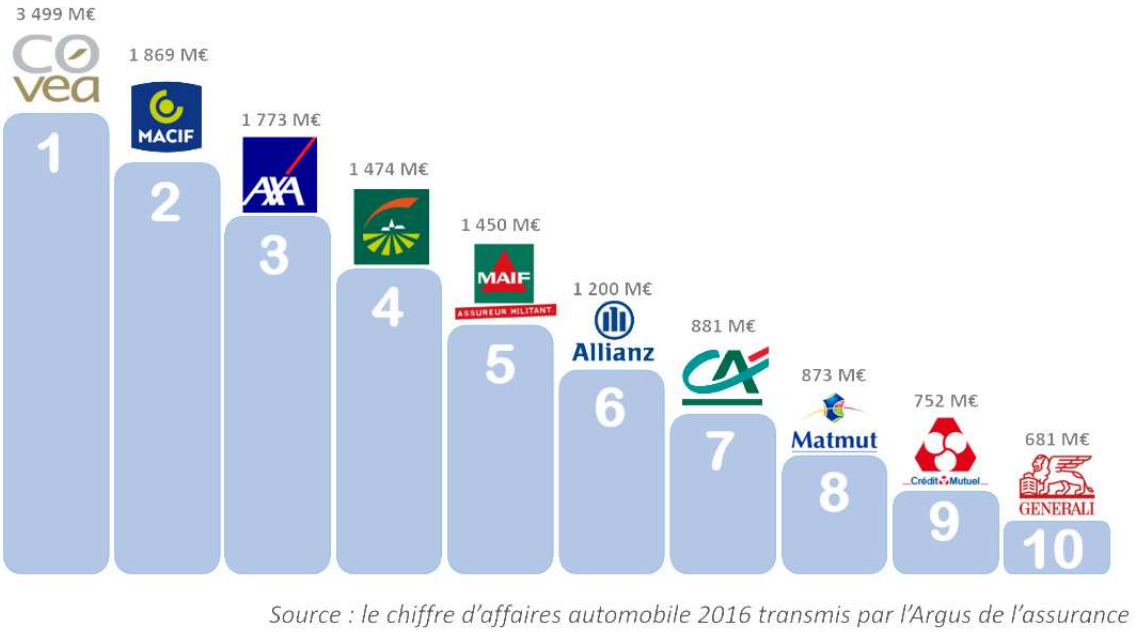

FIGURE 2 – Les dix premiers acteurs du marché de l'assurance automobile

Pour la majorité des leaders du marché de l'assurance automobile, dans le secteur des deux-roues, le chiffre d'affaire est à la hausse, excepté pour le groupes AXA. Bien que Covéa produise deux fois le chiffre d'affaires du deuxième assureur, il apparaît que la concurrence semble ouverte puisque personne ne dispose de plus de 20% des parts du marché. Toutefois, ce résultat est à modérer, puisqu'en rapportant ces chiffres au total de 2015, et en prenant en compte les chiffres d'affaires des secteurs deux-roues et flottes, le cumul des chiffres d'affaires des dix premières sociétés d'assurance couvre près de 90% du marché.

La tarification automobile fait l'objet de beaucoup d'études chaque année en raison du chiffre d'affaires en jeu, et de l'obligation à s'assurer pour ce qui concerne la responsabilité civile. Si le marché est relativement stable, le défi des assureurs est bien de faire face aux coûts des prestations dans un environnement concurrentiel ne permettant pas un ajustement aisé des primes. Afin de mener à bien les travaux de tarifications pour les entreprises, plusieurs méthodes sont mises en application de manière récurrente. La principale reste l'application d'un GLM afin d'avoir un modèle relativement simple d'utilisation pour les entreprises. Toutefois, la digitalisation donne lieu à l'exploration de nouvelles méthodes notamment liées au machine learning qui pourraient supplanter les

2. https ://www.lelynx.fr/assurance-auto/comparaison/meilleure-offre/etude-prix-2018/

méthodes classiques à terme. Ces méthodes seront développées et mises en application dans le cadre de ce mémoire pour apporter des solutions envisageable à des problèmes autrement difficilement solvables.

1.1.3 L'intérêt des conventions IRSA et IRCA

Ces deux conventions IRSA (Indemnisation Règlement des Sinistres Automobiles) et IRCA (Indemnisation et Recours Corporel Automobile) ont été créées afin de raccourcir les délais et de faciliter l'indemnisation à la suite d'un sinistre automobile.

Les conventions IRSA et IDA

L'IRSA et l'IDA (Indemnisation Directe de l'Assuré) ont été mises en place par plusieurs assureurs auto/moto dans l'optique de faciliter les indemnisations des assurés. Les deux conventions fonctionnent par paire. A la suite du dépôt du constat de l'assuré, l'assurance va déterminer, par le biais du barème forfaitaire IDA, la part de responsabilité de l'assuré dans la survenance du sinistre (0/100, 100/0 ou 50/50). De cette part de responsabilité découlera le droit ou non à indemnisation, selon les garanties prévues par le contrat. C'est bien l'assureur qui indemnise son assuré, et non pas l'assureur de la partie adverse, même en cas de sinistre non-responsable. L'assureur est par la suite libre d'effectuer un recours contre la compagnie d'assurance adverse.

Ce recours est forfaitaire si le montant de dommages est inférieur à un plafond fixé par la convention (6 500 €). Il sera proportionnel au niveau de responsabilité de l'auteur du dommage. Le recours est réel si le montant des dommages est supérieur au plafond. Cela implique notamment que dans les cas des sinistres à faibles montants, l'assureur aura un coût à payer même si l'assuré n'est pas responsable.

La convention IRCA

Concernant l'IRCA, les sinistres corporels attritionnels représentent 90% de la masse des dossiers traités quand aux sinistres entraînant des conséquences corporelles. L'application d'une convention traitant ce genre de cas permet donc aux entreprises d'assurance notamment de réduire leurs coûts de gestion, ainsi que le poids financier des sinistres. Du côté de l'assuré, cela a pour avantage de raccourcir les délais de règlement des dossiers en accélérant le processus d'indemnisation aux victimes. Contrairement à l'IRSA, l'IRCA découle des articles 12 et suivants de la loi dite "Badinter" du 5 juillet 1985, cette dernière ayant été incluse au sein du Code des assurances aux articles L. 221-9 et suivants. Son origine est donc légale. Des limites sont toutefois observées pour les dossiers entrant dans le domaine d'application de l'IRCA :

- Les dossiers à faibles conséquences corporelles : le taux d'IPP causé doit être inférieur ou égal à 5%.
- La zone géographique : seuls les sinistres survenus en France, métropolitaine et outre-mer, ainsi qu'à Monaco sont pris en compte.
- Un minimum de deux véhicules terrestres à moteur doivent être impliqués. Ces derniers doivent être rattachés à des sociétés d'assurance signataires de la convention.

Si le sinistre entre dans ce cadre précis, l'indemnisation sera alors effectuée par l'assureur direct "responsabilité civile".

Ces deux conventions peuvent poser des problèmes, le plus évident étant lié au fait que l'assureur ne sera pas forcément indemnisé en totalité par l'assureur adverse en cas de non-responsabilité de l'assuré. Cela peut conduire l'assureur à rembourser au minimum l'assuré. Toutefois, le fait d'éviter le processus du cas par cas, et l'échange parfois lent entre les deux assureurs consiste en un avantage non-négligeable en la faveur de ces conventions. Il est important, dans la suite de ce mémoire, d'être informé de ces procédures puisqu'elles forment la majeure partie des affaires de l'assureur. Il s'agit d'un sujet majeur dans le domaine de la tarification.

L'importance du recours en assurance automobile

L'introduction de ces deux conventions fait entrevoir un paramètre non-négligeable rentrant dans le cadre de la tarification automobile. En effet, les différents procédés de recours donnent lieu à des modifications importantes du coût final imputé à la cédante. Cela peut engendrer des coûts de sinistres négatifs, dans le cas d'un sinistre lié à un assuré mais où sa responsabilité n'est pas mise en jeu, et mène à des difficultés supplémentaires dans la tarification des garanties. Ce phénomène est donc à garder à l'esprit puisque des modifications de coûts potentiels donnent lieu à la modification de la prime à demander.

Il sera précisé à nouveau dans les parties suivantes lorsqu'il en sera fait question que le coût envisagé comportera également les éventuels recours liés aux sinistres. Ces données ne seront pas traitées de la même manière pour le modèle de coût qui sera mis en place, et leur traitement sera exposé dans la suite de ce mémoire.

1.2 Présentation de l'entité et des problématiques associées

1.2.1 Image globale du groupe et mode d'action retenu

La société étudiée, ci-après nommée le groupe, est un regroupement de plusieurs entités assurantielles indépendantes. Bien que chacune de ces sociétés disposent de leurs propres systèmes de gouvernance, méthodes de tarification, organisation commerciale, la réassurance et la réponse aux exigences règlementaires sont assurées par le groupe en question. Elles sont chacune régies par le Code des Assurances.

Le groupe ne réalise aucune opération d'assurance directe. Son but est uniquement de réassurer les sociétés d'assurance membres dans leur intégralité, et de leur garantir la suffisance de leurs provisions ainsi que leur solvabilité. Il veille à la réponse aux normes règlementaires prudentielles, telles celles énoncées par la Directive Solvabilité II. Le groupe porte qui plus est l'entièreté du risque, et est donc à l'origine de la politique de souscription de chacune des cédantes. Le fonctionnement est décrit par le schéma suivant :

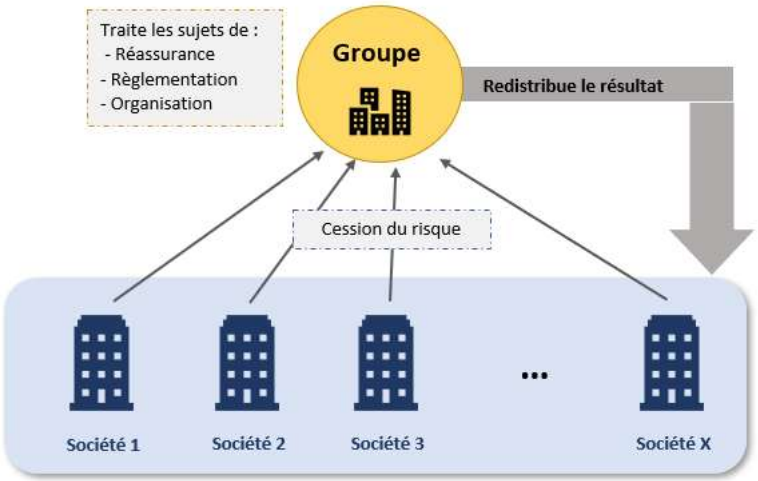

FIGURE 3 – Organisation et fonctionnement du groupe

En raison d'une telle disposition, observer les impacts de la tarification automobile sur la réassurance soulève plusieurs questionnements. En effet, la tarification est réalisée sous un modèle homogène par toutes les mutuelles mais la segmentation y est faite différemment, et les processus de tarifications font apparaître des différences majeures au sein du groupe. Il faudra donc définir un périmètre précis qui permettra d'uniformiser la tarification de chacune des cédantes. Si les études de tarification pourraient s'effectuer individuellement, il est primordial de travailler sur le portefeuille du groupe sans faire une distinction de ces dernières puisque ce ne sont pas chaque mutuelles qui sont réassurées mais bien le groupe qui a souscrit un seul contrat de réassurance.

Il a été observé une rentabilité peu satisfaisante concernant les garanties automobiles. Le but de l'étude est donc dans un premier temps de s'intéresser à un éventuel phénomène d'anti-sélection au sein du portefeuille global. Les résultats permettront alors d'effectuer en parallèle une tarification, en utilisation une méthode par GLM, et une surveillance de la rentabilité du contrat de réassurance pour le réassureur. En effet, cette rentabilité n'apparaît pas suffisante actuellement, et faire davantage les liens entre la qualité du portefeuille et la fréquence des sinistres graves pourrait renseigner le groupe sur la stratégie de réassurance à adopter, ou rassurer les réassureurs sur la rentabilité du contrat.

1.2.2 Périmètre global du portefeuille

Dans le présent mémoire, l'accent sera porté sur les garanties automobile de l'ensemble du groupe afin d'avoir un montant de données minimal pour mener à bien les travaux. En s'intéressant aux montants de sinistres de

ces dernières il apparaît en effet que l'information perdue n'influerait pas énormément sur les conclusions qui pourraient être tirées concernant la réassurance ou l'étude de tarification.

Le périmètre concernant les types de véhicules étudiés a rapidement été réduit aux 4 roues (excluant ainsi les deux roues, les engins agricoles, véhicules collectifs, etc.), afin de faire coïncider les produits assurés de l'ensemble des cédantes. Il est à noter que les garanties dont le groupe ne porte pas le risque n'ont pas été conservées (telles que l'assistance automobile par exemple). Les garanties suivantes seront donc étudiées :

— Accessoires auto
— Adhésion à l'association du groupe
— **Bris de glace** (BDG)
— Catastrophe technologique
— Catastrophe naturelle
— **Dommage tout accident** (DTA)
— Garantie du conducteur
— Incendie
— Panne moteur
— **Responsabilité civile** (RC)
— Tempête/Grêle
— Vol

Les garanties en gras sont les plus importantes du groupe, et l'intérêt sera principalement porté sur celles-ci par la suite. En effet, il convient usuellement de tarifer chaque garantie individuellement. La répartition des primes par garanties est donné par le graphique suivant dans le cadre choisi de l'étude (avant traitement de la base pour les études de tarification) :

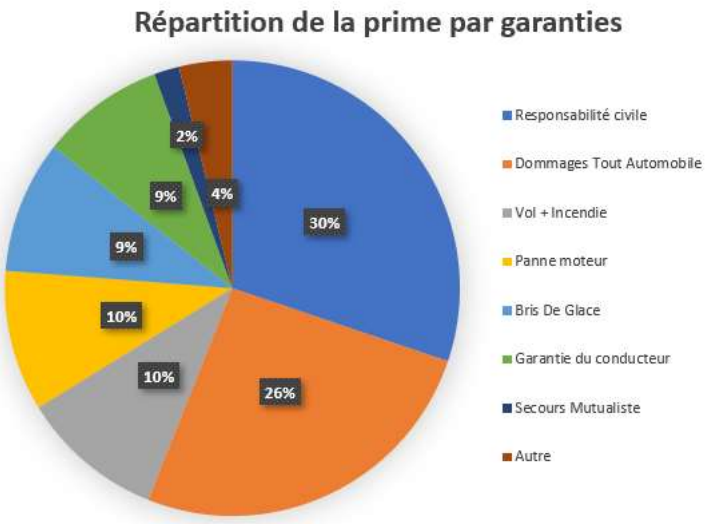

FIGURE 4 – Répartition des primes du groupe

Les garanties majeures sont ici les mêmes que celles observées globalement sur le marché français (cf figure 1). Il apparaîtra clairement dans les résultats présentés tout au long du mémoire que certaines garanties ne présenteront pas d'intérêt réels à être étudiées individuellement, mais la prime correspondante pourra servir à contrebalancer la sinistralité d'autres garanties plus sinistrées. Globalement, il sera observé que les garanties Bris de Glace, RC, Dommages tout accident seront les plus intéressantes, et fourniront suffisamment de données pour donner lieu à des conclusions utilisables pour obtenir la tarification finale. Cette recrudescence de données concernant ces garanties est commune à l'ensemble des cédantes du groupe.

1.3 Introduction de la réassurance

1.3.1 La réassurance automobile en France

La réassurance est un processus répondant à deux besoins des assureurs dans le paysage assurantiel actuel. Dans un premier temps, il est évident que les assureurs ne peuvent assumer la couverture des sinistres graves seuls. Le transfert de risque permet en effet une deuxième échelle de mutualisation du risque, faisant jouer la loi des grands nombres de sorte à éviter une ruine potentielle. Les couvertures d'assurance peuvent ainsi prendre plusieurs formes qui répondent à des besoins différents, qui seront décrits dans la suite de cette partie. La deuxième raison concerne la Directive Solvabilité II, où le transfert de risque à une compagnie de réassurance permet d'optimiser la couverture du ratio SCR afin de répondre plus facilement aux exigences règlementaires. Bien sûr, cela induit un risque de contrepartie, mais dont l'impact sera moins important généralement que si la compagnie ne se réassure pas.

Le domaine de l'assurance automobile n'échappe pas à cette règle, et si toutes les garanties ne nécessitent pas forcément une réassurance, certaines peuvent mener à des montants de sinistres trop élevés pour la capacité de l'entreprise. Le groupe réassurant plusieurs portefeuilles, les contrats passés correspondent à des garanties précises. Deux types de contrats sont alors envisagés par le groupe : un contrat excédent de sinistre et un contrat quote-part.

1.3.2 L'organisation de la réassurance au sein du groupe

Le contrat quote-part

Les contrats quote-part sont parmi les plus simples en matière de réassurance. Ils mettent en jeu un taux de cession de $x\%$, soit la part du portefeuille prise en charge par le réassureur. Cela signifie qu'il s'engage à couvrir $x\%$ des risques moyennant $x\%$ de la prime perçue pour ce portefeuille. Ce type de contrat est utilisé principalement par les assureurs ne disposant pas du capital suffisant pour gérer tous les risques résultants du portefeuille pris en compte. En effet, le quote-part permet à l'assureur d'agrandir son portefeuille sans pour autant détériorer son ratio de solvabilité. Toutefois, ce type de contrat ne couvre pas les risques extrêmes.

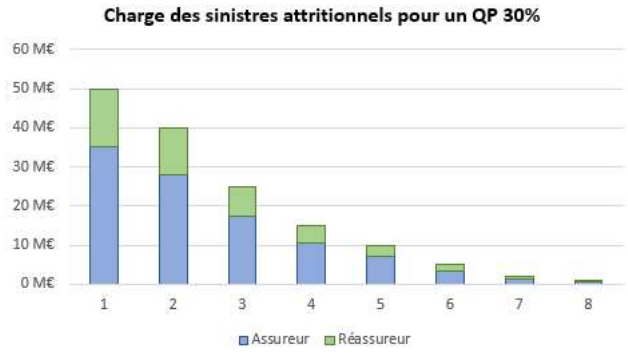

FIGURE 5 – Illustration d'un contrat quote-part

Le groupe utilise deux traités de quote-part concernant respectivement les garanties RC automobile et Dommages Automobiles. Les contrats concernent également d'autres branches d'assurance non-vie mais ce sont ces deux garanties qui seront étudiées dans la suite de ce mémoire.

Le contrat excédent de sinistre

Les contrats excédent de sinistre (aussi appelés contrats XS) sont souscrits dans l'optique de se protéger des sinistres donnant lieu à des indemnisations très élevées. Ils sont généralement découpés en plusieurs tranches, chaque tranche étant définie par une priorité et une portée. Ainsi, la charge C du réassureur concernant un sinistre de montant X couvert par un traité XS à tranche unique de priorité (deductible) D et de portée (limit) L s'écrit :

$$C = min(max(X - D, 0), L)$$

Ce contrat peut également être noté L XS D. En raison de la possibilité de sinistres pouvant toucher plusieurs polices pour le même évènement, mais sans dépasser pour chacune de ces polices la portée indiquée dans le contrat, une distinction a été faite entre deux types de contrats XS. Le contrat par risque s'intéresse aux sinistres de manière individuelle, tandis que le contrat par évènement va s'appliquer au montant cumulé de chaque sinistres qui seraient liés au même évènement. Ce type de contrat implique donc une définition précise de l'évènement, généralement une catastrophe naturelle ou technologique, afin d'éviter tout potentiel litige. La figure suivante illustre la différence qui réside entre ces deux types de contrat :

FIGURE 6 – Comparaison des deux types de contrats XS

L'exemple est pris pour des contrats à tranche unique, l'un par risque 11000 XS 4000 et l'autre par évènement 11 000 XS 4000. Un évènement fait survenir cinq sinistres dont les montants varient de 3000 à 8000 euros. Les graphiques illustrent la différence de coût pour l'assureur et pour le réassureur, en montrant l'intérêt du contrat par évènement. D'autre part, cela montre que les contrats par évènement peuvent avoir des sinistres dont le montant cumulé va excéder la limite fixée par le contrat. Dans ce cas là, il est utile d'avoir recours à des tranches supérieures afin de couvrir le risque dans son entièreté.

Le groupe fait appel à trois contrats distincts de réassurance. Les deux premiers sont des XS par risque, l'un destiné à la garantie RC automobile et l'autre aux garanties DA. Le troisième est un contrat par évènement, renforçant le premier contrat par risque pour les garanties DA. En effet, une catastrophe à grande échelle (qui pourrait être considérée comme un évènement) pourrait déclencher plusieurs garantie et ainsi causer des montants de sinistres, une fois additionnés, difficile à gérer pour l'entreprise d'assurance si cette dernière est seulement protégée par un contrat par risque. D'un autre côté, le contrat par risque permet de cibler davantage les accidents à moins grande ampleur, mais toujours à coûts importants. Le problème d'évènement ne se pose pas vraiment pour ce qui est de la RC puisqu'un évènement de grande ampleur causé par une personne ne concernera que sa propre police, et ainsi il s'agit d'un seul risque à assurer. Il est proposé un exemple en annexe afin de mieux cerner l'utilité de cumuler les deux contrats par risque et par évènements pour couvrir la garantie DA.

1.4 Contextualisation de la réassurance au sein de la problématique

Les limites des deux contrats quote-part souscrits par le groupe auprès des réassureurs coïncident avec les priorités des premières tranches des contrats XS par risque. L'organisation de la réassurance du groupe permet donc de couvrir l'ensemble des risques de sinistralité, qu'ils soient extrêmes ou fréquents. En conséquent, la rentabilité du contrat de réassurance est directement liée à celle du contrat d'assurance. Pendant les 4 années d'observation, il est noté que les contrats XS n'ont pas été déclenchés pour ce qui est des garanties automobiles

(cela s'explique par le faible nombre de donnée, induisant une probabilité assez faible d'observer des sinistres extrêmes au sein du portefeuille).

Si l'intérêt est principalement porté sur la réassurance, résoudre les problèmes liés à l'assurance directe aura un effet directement bénéfique pour la réassurance. Si le ratio sinistre sur primes global baisse, la rentabilité de la réassurance augmentera, et cette réassurance aura alors l'intérêt principal pour l'assureur de nettement améliorer son ratio de Solvabilité 2. L'objectif est donc ici de cibler les problématiques de rentabilité du contrat d'assurance, et de s'intéresser aux manières de les résoudre avant de procéder à une méthodologie classique de tarification. L'idée est que si ces sources de déficit sont correctement identifiées, il faudra ensuite faire le lien avec les contrats de réassurance, puisque chaque garantie n'est pas réassurée de la même manière. Le contrat de quote-part est ici celui qui donnera lieu aux observations les plus intéressantes, et l'étude des sinistres éventuellement déclencheurs du contrat XS ne sera pas menée ici en raison de l'absence de données. Ces contrats pour la garantie auto ont davantage pour but d'améliorer le ratio de solvabilité que de limiter les pertes.

Là où la question sort des problématiques habituelles de tarification, c'est que la réassurance est appliquée à l'ensemble du groupe et non pas à une seule cédante au portefeuille homogène. La suite du mémoire montrera notamment qu'en fusionnant les différents portefeuilles, le caractère hétérogène des données empêchera parfois de tirer des conclusions précises concernant la problématique, causant des soucis de qualité des données notamment.

2 Partie II : Initialisation de l'étude et visualisation du portefeuille

2.1 Appropriation des données

2.1.1 Éléments mis à disposition et discussion sur la méthode employée par le groupe

Afin de bien visualiser les opérations nécessaires à l'obtention finale des données de l'étude, il est intéressant de se pencher sur la manière dont le groupe tarife ses contrats. Cela se réalise en deux étapes :

— Dans un premier temps, l'outil tarifaire s'intéresse aux caractéristiques de base de l'assuré. Cela concerne notamment son âge, son ancienneté de permis, les caractéristiques du véhicule assuré (Groupe, Classe et Puissance) ainsi que le CRM.
— Par la suite, des correctifs sont ajoutés. La logique tarifaire derrière y est toutefois moins précise. Si la base de l'outil apparaît pertinente, l'ajout de correctifs résultant d'observations ou d'idées préconçues peut s'avérer néfaste pour la significativité globale du tarif

Les données concernant la police regroupe de ce fait les caractéristiques même de l'assuré, la décomposition tarifaire par garantie, les correctifs appliqués ainsi que la liste des conducteurs déclarés pour le véhicule. Pour ce qui est de la base sinistre, cela fait simplement état des caractéristiques du sinistré, et de la décomposition du coût du sinistre par garantie. Ces données sont décomposées par années et sont disponibles de manière uniforme pour chaque cédante du groupe.

La problématique à ce fonctionnement peut ne pas apparaître de façon immédiate, mais il s'avère que procéder étape par étape en ajoutant les correctifs tarifaires pour régler les problèmes observées lors d'un exercice en particulier pose certains problèmes d'ordre statistique. En effet, s'il peut paraître judicieux de récompenser ou de pénaliser certains assurés en fonction d'une certaine caractéristique, il est dangereux d'agir ainsi puisque cela revient à ignorer les potentielles corrélations de cette variable avec d'autres déjà prises en compte dans la formule tarifaire.

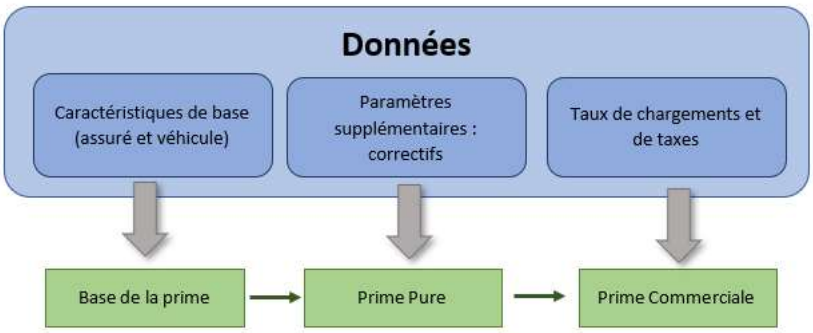

FIGURE 7 – Obtention de la prime commerciale

Là où habituellement, la prime pure ne dépend que d'une étude précise portée sur les probabilités et coût d'un sinitre, qu'il soit attritionnel ou grave, elle est ici modifiée par les correctifs tarifaires. Il convient donc de discuter de la pertinence de ces correctifs appliqués a posteriori, et de les intégrer si possible à la base de l'outil de calcul tarifaire.

S'il n'est pas certain que le fonctionnement de l'outil actuellement utilisé soit la source des problèmes de rentabilité, puisqu'il peut également s'agir d'une attraction accrue de mauvais profils ou même d'une conséquence globale d'une détérioration du secteur assurantiel automobile, il faut garder ce fonctionnement à l'esprit durant la suite des travaux. Étant dans le cas d'un groupe de petite taille, un des intérêt majeur de ce mémoire sera également de s'intéresser à l'application d'une méthode classique de tarification afin de visualiser l'impact que cela pourrait avoir.

2.1.2 Mise en place des bases d'études

La base de données mise à disposition nécessite une série de traitement pour être utilisable. En effet, les fichiers au format csv fournis n'ont jamais fait l'objet de traitements par le groupe, et il s'agissait de la première étude détaillée réalisée sur ces derniers. Ainsi, par cédante et par année, plusieurs fichiers sont destinés à être utilisés. L'objectif était de travailler sur ces bases de sorte à converger vers des bases de données exploitables dans le cadre d'un GLM. Si à l'origine, les données n'étaient pas directement mises sous la forme de deux bases, l'une renseignant les polices (destinée à devenir une base de fréquence) et l'autre les sinistres (destinée à devenir une base de coût), il a fallu réaliser des opérations pour avancer dans cette direction.

Les premiers travaux ont donc concerné la fusion des bases brutes fournies par le groupe. Il s'agissait à l'origine de six fichiers CSV par année, selon le nombre d'années de données disponibles. Les fichiers de 2015 et postérieurs ont donc été repris sous format Excel, puis fusionnés de sorte à n'avoir plus que deux bases par an par cédante. Les fichiers faisant état des conducteurs et des correctifs ont été retraités afin de mettre en ligne un numéro de police et en colonne les variables correspondants aux conducteurs pour le premier, et pour mettre en variable chaque correctif dans le second afin de mettre de la même manière le numéro de police unique en ligne. Des fusions ont ensuite été effectuées sur Access afin de récupérer les deux bases Police et Sinistre.

Les garanties pré-sélectionnées comme faisant partie du périmètre ont été conservées tandis que celles hors scope ont été abandonnées, de même pour les véhicules autres que les 4-roues. La question s'est posée concernant les sinistres à conserver selon leur état. Finalement, les sinistres ouverts et soldés ont tous deux été conservés, tandis que les sinistres sans suite ont été évincés de la base.

Un autre traitement a eu lieu pour la création de variables à partir de celles déjà présentes dans la base. L'âge, l'ancienneté du permis ont dont été rajoutés, puis plusieurs variables groupant des correctifs similaires ont pu être créées. En raison du grand nombre de correctifs utilisés par les cédantes du groupe, le nombre de variable était assez important, mais il faut garder à l'esprit que toutes ne sont pas ensuite retenues pour la tarification du produit. La première étape étant d'identifier les mauvais profils du portefeuille, et les éventuelles raisons de baisse de rentabilité pour le groupe, conserver le maximum de variables explicatives consistait en une bonne approche.

A partir des chiffres contenus dans la base sinistre, les coûts totaux des sinistres ont été ajoutés pour chaque ligne. Le coût total du sinistre correspond au coût réel du sinistre. Il s'agit de la somme des règlements des frais, des honoraires et des indemnités ainsi que des provisions restantes correspondantes avec le montant de franchise appliqué. Le coût total pour le groupe est la somme des règlements des frais, honoraires et indemnités ainsi que des provisions restantes correspondantes, à laquelle sont soustraits les recours perçus, dont l'origine et le fonctionnement ont été évoqués précédemment, ainsi que les provisions correspondantes.

Ces variables étant créées, la dernière étape consistait à obtenir les bases qui seraient par la suite traitées sur R pour obtenir les premières statistiques permettant une meilleure vision globale du portefeuille. En effet, les deux bases présentent des doublons au niveau des identifiants de polices et de sinistres puisqu'elles sont classées par garanties. Ainsi, pour les études portées sur le nombre de police et de sinistres correspondant à une certaine catégorie du portefeuille, les montants ne sont pas important et cela implique d'avoir une base de données sans doublons sur les polices. Cependant, des doublons peuvent apparaître d'une année sur l'autre, cela veut simplement dire que la police a été renouvelée. La suppression des doublons doit se faire année par année puis les données peuvent alors être agrégées à nouveau pour être traitées.

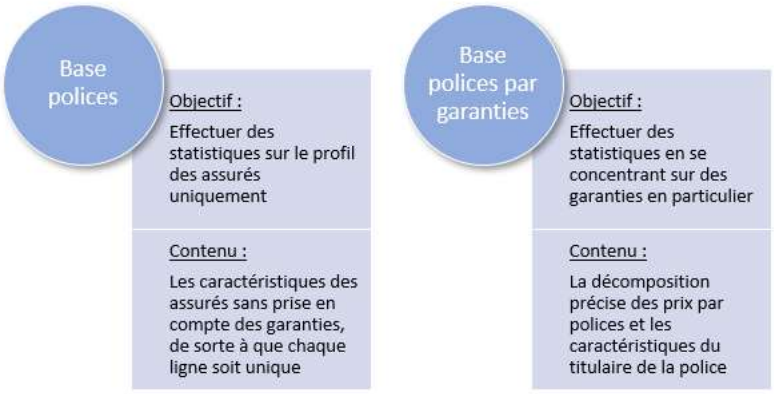

FIGURE 8 – Les deux bases pour les polices

De ce fait, ce seront bien quatre bases qui seront disponibles par cédantes, permettant à la fois des études sur les garanties, et sur le portefeuille global. Deux bases regrouperont les polices uniques par années, et donc n'auront pas l'information des garanties précises, ceci afin d'affiner les informations résultants du portefeuille. Les bases par garantie seront à l'origine des bases fréquence et coût qui seront mises en place pour les travaux concernant la tarification en elle-même.

Avant de s'intéresser aux statistiques résultants d'une première étude de la composition du portefeuille, il est important de rappeler que l'étude concerne un groupe de plusieurs sociétés, où l'homogénéisation n'est donc pas parfaite pour ce qui concerne certaines variables. En effet, les correctifs n'étant pas les mêmes d'une cédante à l'autre, cela s'ajoutant aux problèmes de manque de donnée ou de présence de données aberrantes, il convient de réaliser des traitements supplémentaires durant le processus de mise en place des bases. Cela rejoint par ailleurs le problème de la qualité des données qu'il convient d'aborder dans le cadre d'une étude de tarification.

2.1.3 Points d'attention concernant la qualité des données

Avec l'arrivée de Solvabilité II, et l'avènement des études liées au Big Data, il est devenu primordial de s'intéresser plus en détail à la qualité des données utilisées dans le cadre d'études statistiques ou probabilistes. Le domaine de la tarification automobile ne fait pas exception à la règle puisque le tarif récupéré en sortie dépend directement des données utilisées. Il est donc important de procéder avec précaution pour affiner la base brute fournie, et en tirer une source utilisable d'informations, qui n'aura pas été dénaturée par les retraitements.

Quelques précisions

Il est judicieux, avant de poursuivre l'analyse, de définir dans un premier temps comment évaluer la qualité d'une donnée ou d'une base de donnée. Trois conditions se posent alors :

- La donnée doit présenter un caractère unique. En effet, un effet de duplication nuira aux statistiques globales effectuées sur la base.
- Elle se doit également d'être intelligible. Une logique est à conserver, et elle se doit de présenter les caractéristiques qui sont attendues.
- Enfin, elle doit être correcte. Ses caractéristiques renseignées doivent être exactes.

La qualité d'une donnée est directement liée à l'utilisation à laquelle elle est destinée, ainsi qu'au contexte dans lequel elle s'inscrit. Ainsi, la qualité de la donnée est subjective puisque selon le but de cette dernière (faire l'objet d'un audit ou être utilisée dans le cas d'une étude interne par exemple), cela ne sera pas jugé de la même façon. Ce n'est pas un concept absolu et en conséquence donner une définition constante n'est pas réaliste. Assurer la qualité de la donnée, c'est attester de la qualité dans le temps de la donnée, et donc de surveiller les évolutions qui peuvent s'y appliquer.

En se focalisant notamment sur le domaine de l'assurance automobile, s'assurer de la qualité des données c'est mettre en place un contrôle régulier des variables n'évoluant pas linéairement avec le temps (caractéristiques de la formule choisie, nombre d'enfants, détention d'un garage, etc.) qui peuvent être renseignés. Dans un premier temps, un contrôle lors de la déclaration du souscripteur est à effectuer, ainsi qu'une surveillance de

la saisie informatique.

À noter également, dans le cadre de ce mémoire, la base de données n'a ici pu être analysée qu'à posteriori, mais mettre en place un contrôle dès les premières saisies s'avère bien plus efficace pour s'assurer de la bonne qualité de la base. Un suivi régulier des données (correspondant à un suivi des caractéristiques des souscripteurs) peut être également mis en place. Par exemple, un assuré peut avoir souscrit à l'origine pour une certaine formule kilométrique. Idéalement, il serait nécessaire de mettre en place des contrôles afin de s'assurer de la véracité de la déclaration, et du respect de la formule souscrite. Un mauvais suivi peu donner lieu à des données erronées car non exacte, et faussera les conclusions tirées de l'étude d'un correctif en particulier.

Les contrôles devraient ainsi être effectués à l'origine de la souscription et être régulièrement effectués tout au long du contrat. Cela permet par la suite de rendre plus juste les études et de pouvoir exposer les résultats en étant convaincu de leur pertinence et de leur bonne transcription de la réalité. Il s'agit par ailleurs d'un premier moyen de s'assurer de la véracité des caractéristiques des polices souscrites par les assurés, et d'éviter une perte de profit liée à une éventuelle déclaration frauduleuse. Avant même de s'intéresser aux profils sinistrés du portefeuille, le premier axe de travail et d'amélioration concerne la bonne saisie des informations des assurés qui le composent.

Traitement de la base étudiée et analyse de la qualité des données

Dans un premier temps, il a fallu délimiter une plage temporelle pour effectuer l'étude. Les données de certaines cédantes remontaient à plus d'une dizaine d'années, mais ce n'était pas le cas pour toutes. De plus, en raison de changement de produits au cours du temps, les données d'une année à l'autre pouvaient renvoyer à des tarifications différentes, et des changements avaient pu survenir entretemps. Ainsi, et en se basant sur des travaux semblables en tarification, il a été décidé de ne s'intéresser qu'aux quatre dernières années, de 2015 à 2018. La masse de données ainsi obtenue serait suffisante, et ne laisserait pas apparaître trop de disparités concernant sa qualité.

La problématique de la qualité de la donnée est exacerbée dans le cas du groupe étudié, puisque si les données des cédantes ont été transmises sous le même modèle, les inputs pour certaines catégories diffèrent et peuvent créer des différences non désirées et surtout, non significatives. Cela a notamment concerné le nom des garanties et des correctifs, qui ont nécessité une uniformisation au sein du groupe. De plus, un nombre non négligeable de lignes faisaient apparaitre des données manquantes, principalement au regard des correctifs. Il faut savoir pour ces derniers que certain ont été adoptés durant la période étudiée, et d'autres ont été abandonnés. Visualiser rapidement les correctifs dans ce cas permet de les exclure prématurément de l'étude. Des erreurs de saisies ont également pu arriver, et certains correctifs ont été abandonnés quand il semblait que les résultats obtenus étaient trop éloignées des attentes éventuelles : lorsque la qualité semblait insuffisante, les études concernant la variable ne pouvaient donner lieu à des conclusions cohérentes avec la réalité.

Un problème s'est également posé concernant les dates de naissance de la base listant les différents conducteurs d'une même police (que ce soit pour le premier conducteur ou les suivants), qui présentait un écart irréaliste pour l'âge correspondant, mais surtout pour la date de permis associée. Cependant, les lignes où ce genre de problème est apparu semblait principalement résulter d'une erreur de saisie, qui a été corrigé en se basant sur la date de naissance provenant de la base police, et parfois inversement. Ces données ont été conservées car elles semblaient toutefois apporter de l'information, bien qu'elles ne pourront être utilisées pour la tarification en elle-même.

Les problèmes de qualité de données ne seront pas tous énoncés ici, mais de nombreux retraitements ont été nécessaire afin d'obtenir une base qui soit la plus propre possible. Certaines suppressions ont été nécessaires, notamment lorsqu'une variable clé nécessaire au GLM n'était pas renseignée. Des soucis de types de variables ont été traités, qui gênaient par la suite la fusion des bases par Access. Il est à noter que les montants renseignés ont l'air de bonne qualité, et que les fusions font bien voir que toutes les polices sont présentes dans chacune des bases, signe que la saisie a été bien effectuée.

Cependant, il faut bien insister sur le fait que la qualité globale des données laisse la place à une amélioration non négligeable. Par ailleurs, plus la qualité est bonne, plus les travaux concernant les données seront rapides, ce qui joue finalement en la faveur de la société concernée. Qui plus est, les résultats seront représentatifs de la réalité de façon parfaite puisqu'il n'y aura pas eu de nécessité de retraitement. Dans le contexte actuel, où ce genre d'étude est plus important que jamais, veiller à la qualité des données dans la mesure où elles feront ensuite l'objet d'étude est devenu trop important pour être négligé.

Problématique liée à une variable : la formule kilométrique

Illustrant le problème de la qualité des données au sein du groupe, une variable en particulier mérite une attention accrue. Il est habituel dans les études de tarification d'utiliser un système de kilométrage afin de préciser la prime demandée à l'assuré par la suite. Le raisonnement sous-jacent est assez simple : un assuré roulant plus de vingt-mille kilomètres par an sera plus à même de causer un sinistre qu'un assuré roulant moins de cinq-mille kilomètres par an. La variable est bien renseignée dans la majeure partie des cas, mais certaines polices n'entrent dans aucune catégorie (formule "4" dans la base). Cela s'explique par la présence de différents types de produits au sein de la base. La volonté ici est d'avoir un produit unique pour découper la formule.

L'objectif serait alors de trouver un moyen d'associer ces polices à une formule correspondantes. La répartition des polices suivant la formule kilométrique est la suivante :

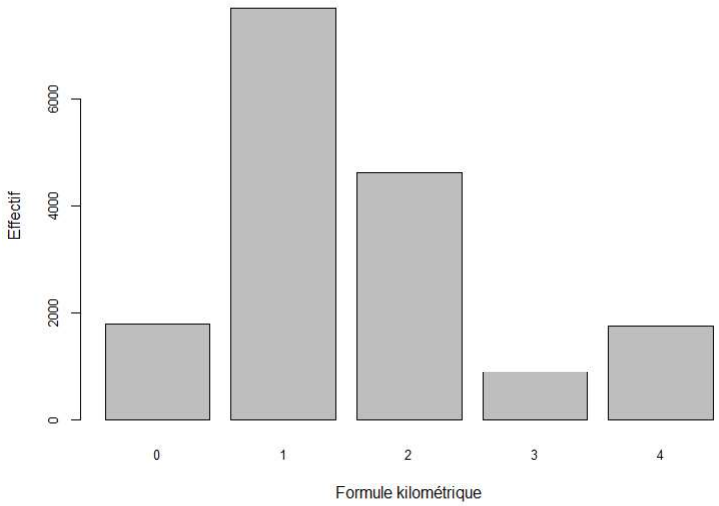

FIGURE 9 – La répartition des différentes formules au sein de la base

— La formule 0 correspond aux assurés ayant souscrit à une formule de moins de 5 000 km par an.
— La formule 1 correspond aux assurés ayant souscrit à une formule entre 5 000 et 15 000 km par an.
— La formule 2 correspond aux assurés ayant souscrit à une formule entre 15 000 et 25 000 km par an.
— La formule 3 correspond aux assurés ayant souscrit à une formule de plus de 25 000 km par an.

La part de polices où la formule kilométrique souscrite n'est pas renseignée s'élève à plus de 10,5% du portefeuille total. Une approche simplifiée serait de distribuer la part de polices posant problème selon la répartition observée pour chacune des formules. Toutefois, cette approche pourrait être erronée puisque les profils n'ayant pas de formule kilométrique présentent des caractéristiques potentiellement différente du profil type de la base de données. Ce type de problème justifie également l'intérêt des statistiques descriptives appliquées sur le portefeuille, puisque ces dernières permettront une modélisation plus exacte des formules kilométriques pour ces polices. L'intérêt sera de nouveau porté sur la gestion de cette variable à la suite de la partie faisant état de la préparation de la base au GLM.

2.2 Statistiques descriptives

Il a donc été proposé de réaliser plusieurs études pour les variables, concernant essentiellement les statistiques univariées et bivariées. L'intérêt est ici de tirer les premières conclusions sur la composition du portefeuille, et de regrouper plusieurs pistes d'information sur les profils à ratios dégradés du groupe. Cette phase est probablement la plus parlante pour les personnes non initiées à l'actuariat car elle permet de visualiser directement, par le biais de graphiques ou de tableaux, les profils qui sont plus sinistrés que les autres. Cela remet en question des décisions tarifaires qui ont pu être prises par instinct, en montrant des tendances qui vont parfois à l'encontre de la logique. Les études réalisées concernent ainsi (par variables) :

— L'effectif (le nombre de polices)
— La police moyenne payée
— L'effectif concernant les sinistres
— Le coût moyen d'un sinistre
— La fréquence de sinistre
— Les ratios S/P [3]

Chaque année est traitée individuellement, et les études concernent chaque garantie pour les S/P et les effectifs, afin de voir si certains pics observés au global sont causés par des garanties en particulier.

Afin d'obtenir les graphiques des différentes études selon une ou deux variables par année, l'outil R a été utilisé. Le fonctionnement est assez simple, puisqu'il consiste en l'exécution d'un fichier faisant appel à plusieurs scripts pour créer et enregistrer les graphiques destinés à être étudiés.

Chaque code, propre à une variable, permet donc l'export automatique des graphiques correspondants, rangés dans un fichier permettant leur traitement immédiat. Il est à noter que lorsque le graphique n'était pas suffisamment visuel, un traitement a été effectué en aval afin d'obtenir les informations voulues de façon plus claire sur excel. Dans un premier temps, avant les études de tarification, l'ensemble des garanties sont prises en compte pour les statistiques bivariées.

2.2.1 Quelques précisions

Les statistiques bivariées mesurent l'impact d'une variable, au cours des 4 années d'étude, sur les domaines précités. Elles permettent de mettre en lumière les différences de rentabilité selon la variable, mais ne font pas voir les éventuelles corrélations entre deux variables. En conséquence, il faut prendre avec précautions les résultats obtenus, et plutôt que de tirer une conclusion immédiate qui pourrait s'avérer hâtive, prendre note des résultats et creuser l'étude en croisant la variable avec d'autres afin d'en préciser les tendances tirées.

La majorité des variables ont fait l'objet de plusieurs analyses afin de visualiser les éventuelles singularités du portefeuille du groupe comparé à ce qui est généralement observé. Différents types d'analyses peuvent donc être réalisées, que ce soit l'évolution d'une variable sur les quatre années, la visualisation du ratio sur 2018, ou même des explications de variation en s'intéressant à des garanties précises. La base contient ainsi 10 363 polices uniques en 2018, qui ont donné lieu à 1450 sinistres, soit une fréquence de 14%. Le ratio S/P global, en prenant en compte les montants sans frais, est estimé à 95%, soit un montant nettement supérieur à celui observé habituellement dans le secteur automobile.

Il faut toutefois modérer ce résultat puisqu'il est assez changeant selon l'année. En effet, l'étude se basant sur un groupe de plusieurs cédantes, la volatilité du résultat en devient non-négligeable. La masse de données n'est pas suffisamment importantes pour supporter des épisodes de sur-sinistralité, ce qui cause des variations importantes du ratio S/P selon l'exercice. Ce dernier est d'ailleurs bien plus faible en 2016 et 2017, mais approche le même ordre de grandeur en 2015 qu'en 2018. Les résultats présentés dans les parties suivantes, se focalisant sur des variables bien précises, sont donc comparés avec les autres années afin notamment de discerner les tendances des anomalies singulières.

Au total, plus de vingt études ont été réalisées. Seront présentés ici les résultats porteurs du plus d'intérêt puisque le résultat de certaines variables n'apporte pas d'information pertinente dans l'optique de déterminer un profil particulièrement sinistré.

3. La définition donnée au ratio S/P durant la totalité du mémoire, sauf mention contraire, sera celle du montant total des sinistres rapporté au volume total des primes, tous deux bruts de frais et de réassurance

2.2.2 Résultats généraux par garantie

Sont résumés dans un premier temps les statistiques par garanties, permettant de voir la répartition du portefeuille selon ces dernières.

	2018			TOTAL (2015 + 2016 + 2017 + 2018)		
	Prime	Sinistre	S / P	Prime	Sinistre	S / P
ACCESSOIRES AUTO	1%	0%	0	1%	0%	0
ADHESION ASSOCIATION	0%	0%	0%	0%	0%	0%
BDG AUTO + AUTO AGRIC	9%	12%	122%	9%	16%	129%
CATAST. TECHNOLOGIQUES AUTO	1%	0%	0%	1%	0%	0%
CATAST.NAT AUTO + AUTO AGRICOL	1%	0%	0%	1%	0%	0%
DTA AUTO + AUTO AGRICOLE	26%	29%	105%	26%	33%	98%
GARANTIE DU CONDUCTEUR	9%	6%	65%	9%	5%	40%
INCENDIE AUTO	2%	0%	23%	2%	1%	35%
PANNE MOTEUR	10%	0%	0%	10%	0%	3%
RC AUTO + AUTO AGRICOLE	31%	50%	153%	30%	43%	107%
SECOURS MUTUALISTE	2%	0%	0%	2%	0%	0%
TEMPETE GRELE AUTO	1%	0%	0%	1%	0%	0%
VOL AUTO + AUTO AGRICOLE	8%	3%	33%	8%	2%	18%
Total général	**100%**	**100%**	**95%**	**100%**	**100%**	**76%**

FIGURE 10 – Répartition par garantie

D'importantes disparités s'observent parmi les garanties. Les plus importantes font voir des déficit tandis que les moins importantes sont en très large situation de bénéfices. L'intérêt peut être porté sur l'impact d'une meilleure répartition des primes selon les garanties, afin de voir notamment si le lissage des ratios S/P impacte la rentabilité de la réassurance. Le traité quote-part étant différent selon la nature de la garantie, des différences pourront être observées.

Il faut retenir ici que les trois garanties principales sont en déficit de manière générale. Si par la suite l'attention sera davantage portée sur ces garanties, il faut bien garder en tête que d'autres garanties tarifées par le groupe sont sources d'importants profits, et répartir les primes de façon plus adaptée pourrait résoudre une partie du problème pour ces trois garanties. Il faut cependant garder à l'esprit que la rentabilité globale serait impactée négativement puisque cela transfèrerait une partie des primes au réassureur.

2.2.3 Étude de la variable "Âge"

La variable Âge peut être considérée comme une des plus fournie en terme d'informations dans le cadre de la tarification automobile. En effet, les statistiques descriptives permettent de bien cibler la situation du portefeuille, de visualiser assez facilement si ce dernier est trop âgé ou trop jeune. Toutefois, l'âge n'est pas toujours retenu comme variable dans le cadre de la tarification en elle-même, notamment concernant l'utilisation d'un GLM. En général, la distinction est simplement faite entre les jeunes conducteurs et les autres, puisque la sinistralité chez les jeunes est plus importante.

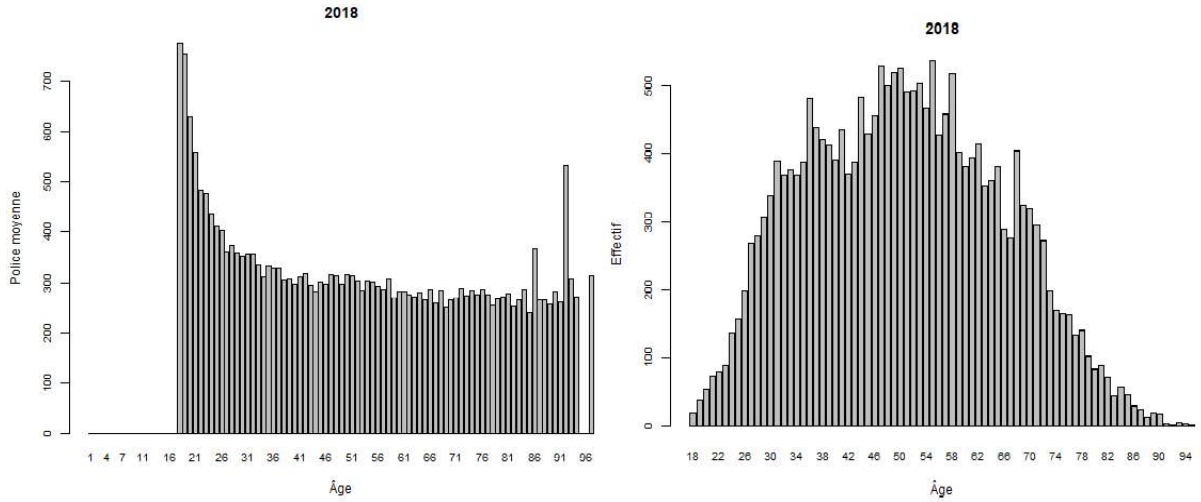

FIGURE 11 – Nombre de polices et police moyenne par âge

Les graphiques permettent ainsi de se faire une première idée de la composition du portefeuille du groupe. L'âge moyen est de 48,4 ans, ce qui correspond à un portefeuille relativement âgé. La répartition ressemble à ce qui est communément observé dans les études de tarification. Il est également à noter une décroissance du montant de la police demandée avec l'âge, à l'exception des assurés de plus de 90 ans. Cela s'explique par un faible effectif dans ces tranches d'âge et serait un moyen de lutter contre la hausse de la sinistralité résultant de ces âges avancés. Le but est donc ici de déterminer si cette prime plus élevée est justifiée au vu de la sinistralité résultant de cette tranche d'âge.

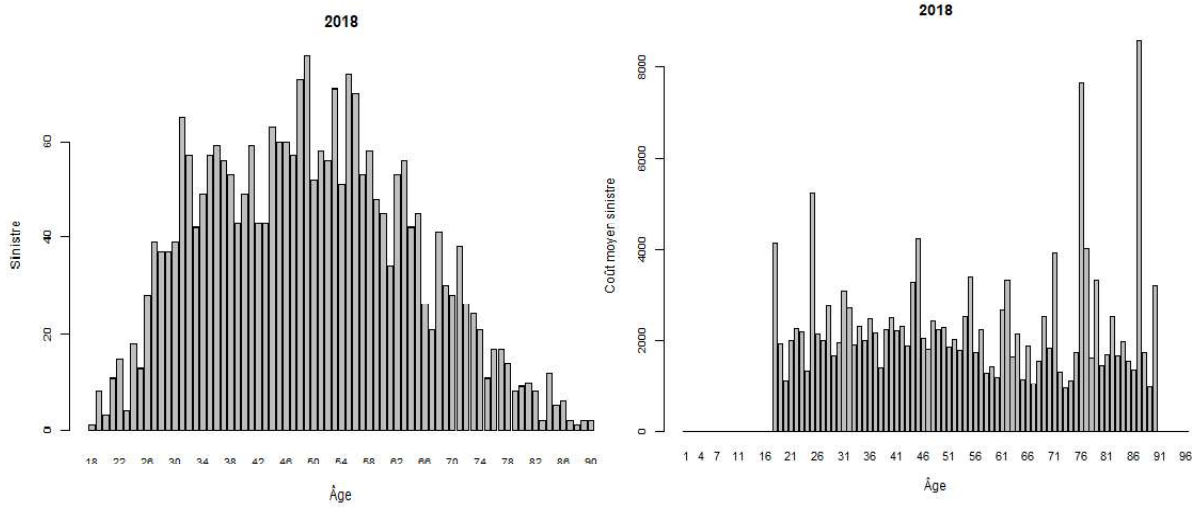

FIGURE 12 – Nombre de sinistres et coût moyen par âge

Les sinistres suivent une répartition semblable à celle des polices, quoique plus erratique car moins lissée par le grand nombre de données. La moyenne d'âge observée est de 46,8 ans, ce qui laisse apparaître un faible décalage avec celle de la base police. Les informations résident dans l'étude du coût moyen, dont le graphique met l'accent sur les sinistres importants qui par la suite dégraderont le ratio S/P résultant. En prenant un comparatif de ce graphique sur les quatre années d'exercices, ces pics n'apparaissent pas au même endroit selon l'année. Toutefois, ces pics apparaissent généralement pour les individus de plus de 70 ans. Là encore, puisque l'entreprise a peu d'assurés figurant parmi ces tranches d'âges, chaque sinistre a un impact certain sur les ratios. Il semblerait donc qu'éviter ce profil d'assuré puisse améliorer le résultat du groupe.

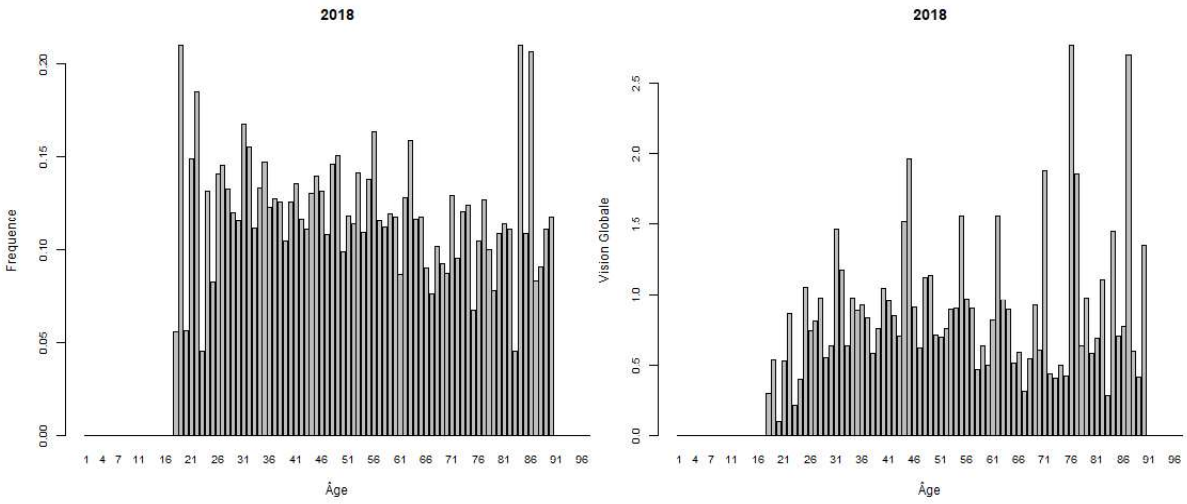

FIGURE 13 – Fréquence et S/P par âge

La fréquence permet de voir la combinaison des effectifs en terme de police et de sinistre. L'étude de cette dernière laisse apparaître une baisse avec l'âge. Des fréquences assez élevées sont observées pour les moins de 30 ans, tandis que la fréquence est nettement revue à la baisse pour les plus de 70 ans, bien qu'elle présente des hausses erratiques au sein de cette tranche d'âge. Concernant la visualisation des ratios S/P, elle permet de voir que les réels problèmes de tarification sont concentrés dans la classe d'âge des plus de 60 ans. D'un autre côté, l'augmentation des polices pour les jeunes assurés semble correctement contrecarrer la fréquence élevée des sinistres qu'ils peuvent causer. Toutefois, il semblerait que ne pas augmenter les polices pour les assurés aux âges avancés soit source de déficit pour le groupe. En effet, bien qu'il y ait peu d'effectif dans ces tranches d'âges, la probabilité d'avoir des sinistres aux montants élevés y est plus importante. De ce fait, il pourrait être demandé une augmentation de la prime pour pallier à ces cas relativement isolés. Une autre solution pourrait être d'éviter cette classe d'âge afin d'améliorer la rentabilité globale du portefeuille.

2.2.4 Étude de la variable "Ancienneté du véhicule"

La variable âge étant étudiée, il convient de s'intéresser à d'autres caractéristiques de la police pour en tirer des informations supplémentaires. Il ressort de l'ensemble des études que l'âge du véhicule semble porter un intérêt certain pour expliquer la sinistralité du portefeuille.

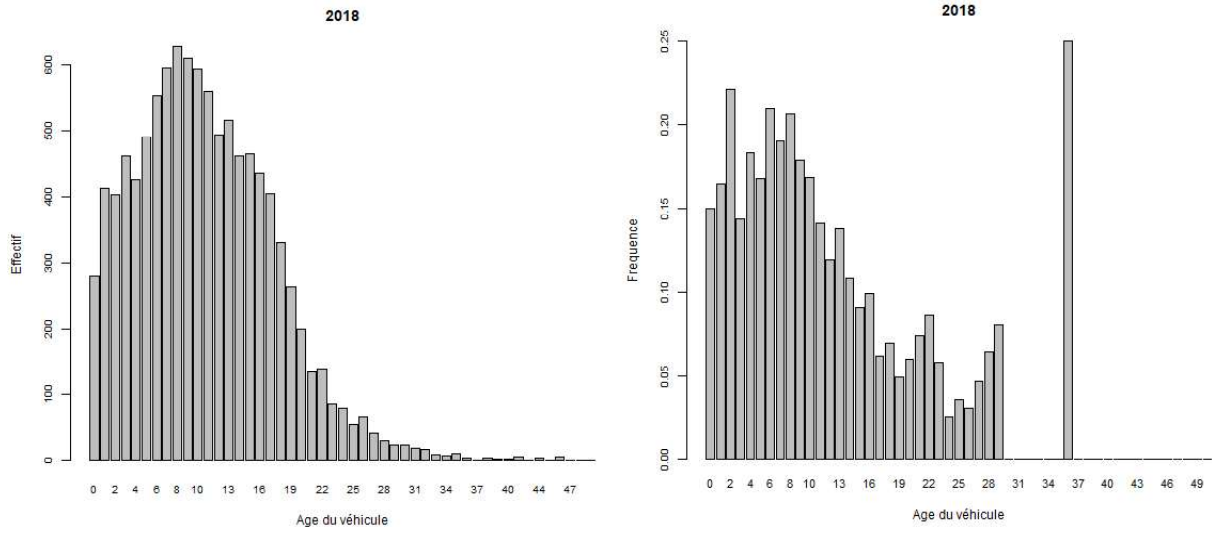

FIGURE 14 – Nombre de polices et fréquence par ancienneté du véhicule

L'ancienneté du véhicule est répartie, en termes d'effectifs, de manière assez classique. La moyenne d'âge des véhicules est de 10,4 ans, tandis que celle du parc automobile français est estimée à 9 ans en 2017. La répartition des polices est telle que l'effectif est croissant jusqu'aux véhicules de 10 ans, où il décroît alors fortement. La fréquence suit une tendance assez similaire, quoique les tendances ne se correspondent plus à partir du moment où le véhicule a plus de 20 ans.

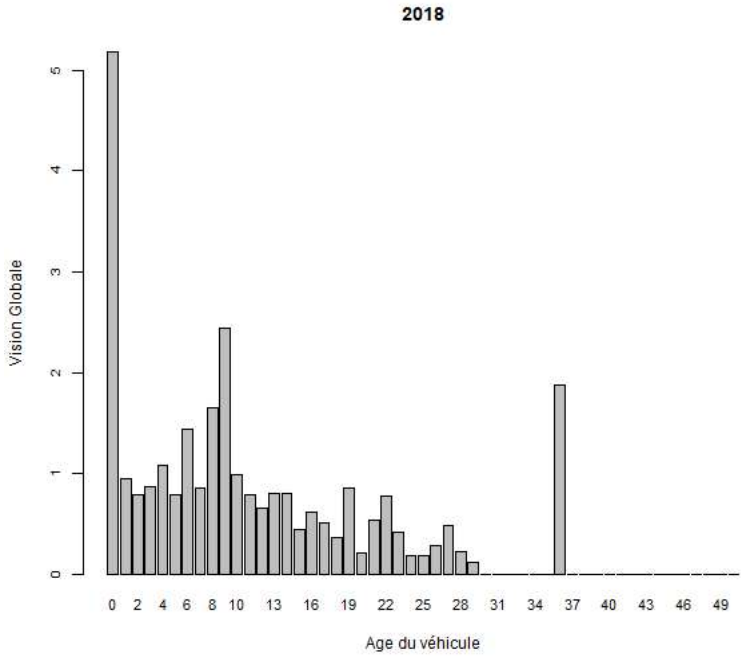

FIGURE 15 – Ratios S/P par ancienneté du véhicule

Sur le graphique représentant le ratio S/P, il apparaît que plusieurs catégories présentent des ratios particulièrement dégradés. En effet, les véhicules neufs voient leur ratio dépasser les 500%, chiffre à relativiser car ce pic n'apparaît pas sur les trois autres années, tandis que les véhicules de 6,8, et 9 ans dépassent eux-aussi les 100%. Enfin, un pic isolé situé au niveau des véhicules de 36 ans semblent davantage est le résultat d'un sinistre isolé que révélateur d'une réelle tendance.

Le but à partir de ces premières observation est notamment de creuser l'origine des pics de S/P. Un croisement de la variable a donc dans un premier temps été effectué avec l'âge de l'assuré. Cela montre notamment que les ratios élevés pour les véhicules ayant entre 7 et 10 ans sont causés par des jeunes de 21 ans et moins, ce qui s'explique par le fait que les jeunes conducteurs n'ont généralement pas les moyens de faire l'achat d'un véhicule neuf. Le ratio élevé pour les véhicules récents s'explique par deux tranches d'âge : les 30-37 ans et les 51-55 ans.

2.2.5 Étude de la variable "Département"

La variable département a également fait l'objet d'études plus poussées afin de visualiser la répartition des assurés en France. L'objet de cette étude est de montrer notamment à quel point le portefeuille du groupe est centralisé au sein d'une zone géographique précise.

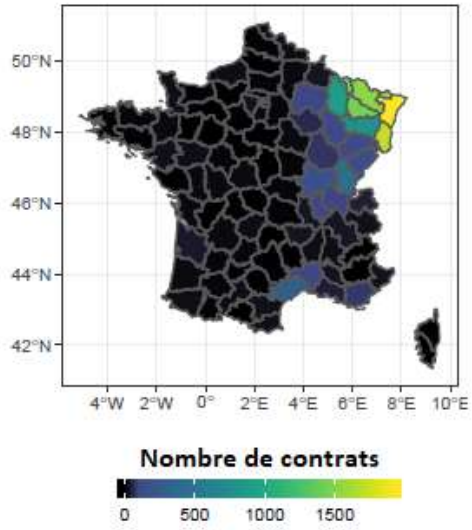

FIGURE 16 – Nombre de polices par département

Les polices souscrites sont principalement regroupées dans le Grand-Est. Toutefois, une diversification se fait voir puisqu'un certain nombre de polices sont situées en Occitanie. La présence d'assurés dans plusieurs régions s'explique par les distributeurs qui ne sont pas forcément situés dans le Grand-Est. Il est important de noter que le nombre de polices pour ces départements ne faisant pas partie des six principaux composant le portefeuille du groupe (54, 55, 57, 67, 68 et 88) est assez faible, ce qui peut mener à un risque supplémentaire important à encaisser pour le groupe. Cela est d'autant plus préoccupant que l'outil tarifaire utilisé à l'origine va être adapté aux profils les plus représentés, et que la police ne sera peut-être pas en mesure d'endiguer la sinistralité correspondantes pour ces zones géographiques mineures.

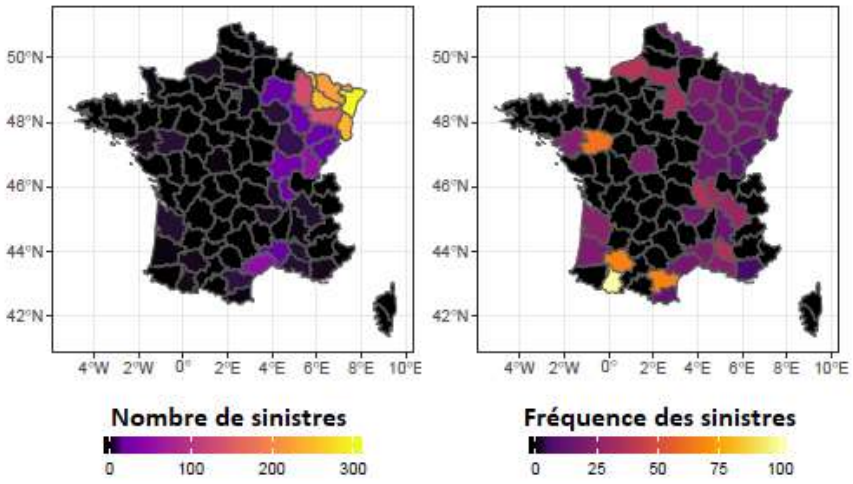

FIGURE 17 – Nombre de sinistres et fréquence par département

Le graphique le plus intéressant, puisqu'il confirme ce qui a été dit précédemment, comme quoi la sinistralité n'est pas suffisamment prise en compte lors de la tarification de contrat dans des départements autres que ceux du Grand-Est. Le faible nombre de polices ne permet pas de couvrir ne serait-ce que le coût d'un sinistre, et cela se fait ressentir par le biais des couleurs qui font apparaître des départements qui semblaient négligeables sur la carte précédente.

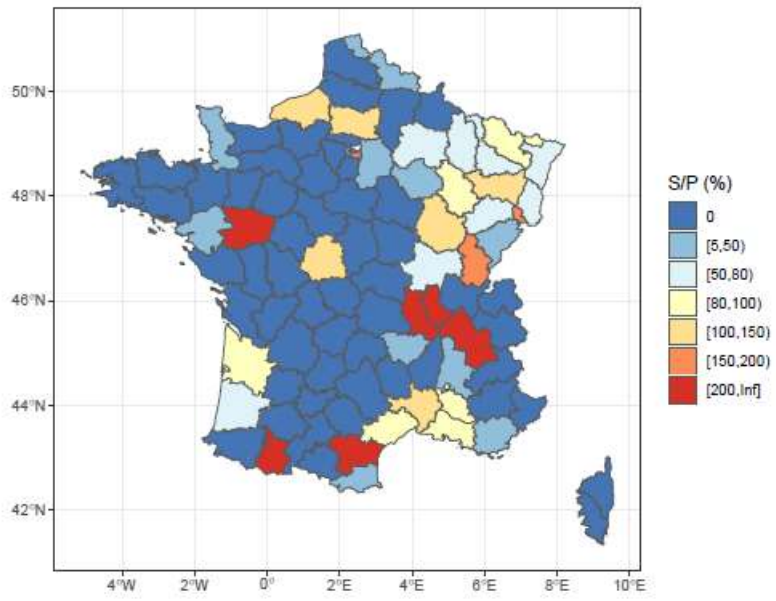

FIGURE 38 – Carte des S/P (%) par département

FIGURE 18 – Ratios S/P par département

Les chiffres finaux sont ainsi obtenus et consistent davantage en une confirmation de ce qui a été dit plutôt qu'en un apport d'information supplémentaire. Bien qu'il ressorte également de ce graphique que les départements principaux du groupe puissent avoir un ratio assez important, notamment le 88 où le S/P atteint les 120%, l'information réside dans les couleurs vives observées hors du Grand-Est. La sinistralité dégradée de ces départements dont le nombre de police est assez faible fait en effet bien voir que la stratégie de diversifier l'origine départementale du portefeuille n'apporte pas de profit au groupe, ceci principalement pour la moitié sud de la France.

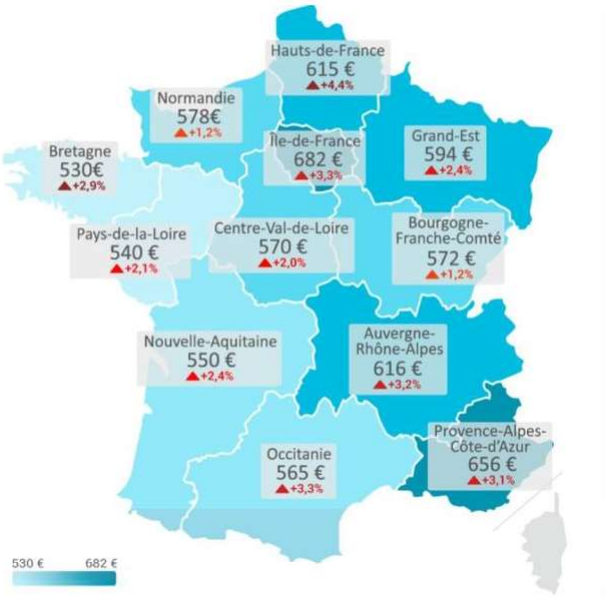

FIGURE 19 – Prime moyenne observée par région en 2018 et augmentation par rapport à 2017

En s'intéressant aux chiffres disponibles donnant la prime moyenne par régions [4] (et en supposant que ces primes proviennent de contrats correctement tarifés), les différences de tarification ne font pas apparaître la nécessité de sur-tarifer les polices provenant d'Auvergne Rhône-Alpes et d'Occitanie par rapport à celles du Grand Est. La conclusion découlant de cette information supplémentaire serait que les courtiers capteraient des mauvais

4. Source : http ://leparticulier.lefigaro.fr/article/assurance-auto-les-regions-les-plus-cheres/

profils dans ces zones. La logique expliquant ce phénomène est simplement que les assurés intéressés par la souscription d'une assurance automobile auprès d'une cédante basée dans une région différente résultent d'une forme de sélection. N'ayant potentiellement pas trouvé d'assureur acceptant leur profil de risque près de leur lieu de résidence, il peut être supposé qu'ils aient cherché qui les accepterait.

2.3 Enseignements à retenir

Jusqu'à présent, deux axes d'amélioration ressortent de cette première série d'études. Dans un premier temps, les études bivariées permettent de mieux visualiser les profils porteurs d'une sinistralité importante, et l'analyse descriptive globale des ratios S/P des garanties fait comprendre que des travaux concernant la répartition des primes pour chaque garantie pourraient être menés, ou du moins, que l'impact de ce manque d'uniformisation pourrait être étudié. La problématique étant cependant principalement liée à la politique du groupe en question, cette piste ne sera pas développée ici puisqu'elle n'entre pas dans le cadre de ce mémoire.

La conclusion globale de cette partie est caractéristique du problème majeur directement lié à la taille du groupe : le manque de données disponibles ne permet pas de dégager un profil type d'un assuré qui causerait une sinistralité particulièrement forte, et donc portant un impact négatif à la rentabilité globale du portefeuille. Certes, les études par département et par âge font voir des pistes quant à certains assurés peu rentables pour le groupe, notamment les personnes âgés et celles résidant ailleurs que dans les six principaux départements du groupe. Toutefois, cela ne suffit pas à expliquer tous les problèmes relatifs au portefeuille. Une part de l'excès de sinistralité peut-être imputable à ces profils pré-cités, mais il convient de poursuivre les travaux afin de voir si la tarification par rapport à son fonctionnement même peut être améliorée.

C'est pour cette raison que les travaux sont poursuivis, pour voir si un écart pourrait être constaté entre un processus de tarification usuel résultant d'une modélisation GLM et l'outil actuellement employé par le groupe qui a été assujetti à de nombreuses modifications. L'objet de la partie suivante est donc de mettre en place les bases de données à utiliser pour appliquer des méthodes de GLM, afin d'en obtenir un outil tarifaire utilisable, et de mesurer les écarts avec les données disponibles.

3 Partie III : Approches techniques inclinées à l'estimation de la prime

3.1 Préparation au GLM

Les premières études des données étant disponibles, et les premières conclusions ayant été présentées dans la partie précédente, il convient désormais de s'intéresser à la partie qui permettra d'avoir un premier aperçu la prime à demander pour les trois garantie étudiées. Les études de tarifications s'axent généralement autour de deux parties : les créations de bases fréquence et coût, incluant la création de classes au sein des variables disponibles ainsi que l'observation des éventuelles corrélations, forment la première partie. La seconde partie consiste principalement en l'exécution d'un GLM, dont les hypothèses et le fonctionnement seront décrits dans la suite de la partie.

La tarification sera basée sur un modèle de coût/fréquence. La variable fréquence correspond au nombre de sinistres total pondéré par la durée d'exposition, tandis que la variable coût correspond au coût total pour le groupe. Les deux variables ont été rajoutées à la base globale. Ces bases de données, pour chacune des trois garanties, sont alors soumises à divers traitement, induisant notamment l'étude des valeurs extrêmes, l'application de méthodes de clustering ayant pour objet la catégorisation de chaque variables en des classes bien distinctes, et enfin, la visualisation des corrélations.

Concernant l'étape des clusterings, celle-ci a une importance capitale. En effet, en supposant une prise en compte de l'ensemble des variables non modifiées, le GLM porterait sur plus d'une centaine de modalités, avec en sus des coefficients à affecter aux variables quantitatives. Outre le temps de calcul extrêmement conséquent qui en résulterait, le manque de données concernant certaines modalités causerait une perte de significativité.

Les travaux concernant les trois points précités seront développés plus particulièrement dans le cadre du modèle de fréquence. En effet, les opérations pour la mise en place de la base du modèle de coût sont assez semblables, et exposer les deux processus dans ce mémoire ferait apparaître plusieurs répétitions porteuses de peu d'informations complémentaires. L'exception sera faite pour l'étude du traitement des valeurs extrêmes de la base coût qui sera également développé en raison de choix importants à mettre en valeur. Les retraitements effectués pour les bases Bris de Glace et Dommages toute Automobile ne seront pas non plus développés ici. Seuls les résultats seront tous présentés dans le cadre de la quatrième partie de ce mémoire.

3.1.1 Traitement des valeurs extrêmes

Base fréquence

Il convient dans un premier temps d'éliminer les valeurs extrémales qui ne présentent guère d'intérêt pour la suite de l'étude. Deux variables sont ici concernées : l'exposition et la fréquence. D'un côté, une exposition trop faible est inintéressante dans la mesure où les statistiques en résultant serait quelque peu biaisées par exemple par la prise en compte d'une police qui ne serait restée que quelques jours au sein du portefeuille. De façon similaire, une fréquence trop importante pourrait avoir un poids conséquent pour la suite des travaux. Ainsi, dans un premier temps, il convient de s'intéresser à la répartition des deux variables au sein de la base étudiée. L'intérêt est donc porté sur les quantiles de fréquence (soit 200 points répartis entre 0 et 1).

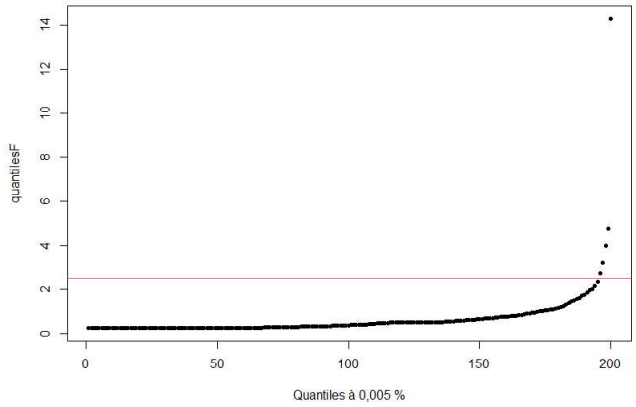

FIGURE 20 – Distribution des fréquences strictement positives de sinistres

Il apparaît pour la répartition de la fréquence que le décrochage, soit la limite où l'espacement des fréquences est clairement visible, se fait aux alentours de 2.5. En conséquence, il a été décidé de tronquer les valeurs supérieures à ce seuil observé graphiquement [5]. Concernant la répartition des expositions, il faut également supprimer les contrats donnant lieu à une trop faible exposition puisque cela pourrait fausser par la suite les statistiques réalisées. Les quantiles sont alors représentés sur le graphique suivant :

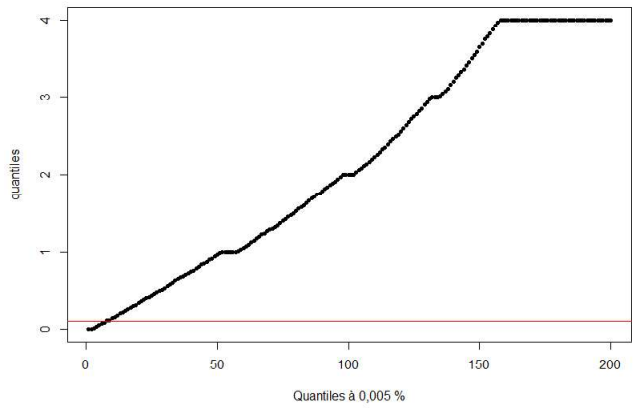

FIGURE 21 – Distribution des expositions

Les valeurs trop faibles (certains contrats souscrits et immédiatement résiliés notamment) sont donc ôtées de l'étude. Cela correspond aux expositions inférieure à un mois, comme il est d'usage de procéder puisque les rétractations quasi-immédiates n'ont pas à être étudiés parmi les autres polices de la base en raison de la durée d'observation jugée trop courte, voire inexistante. Le graphique permet également de voir un effet de "plateau" concernant les valeurs d'exposition entière. Toutefois, il apparaît suffisamment faible pour ne pas donner lieu à de plus amples questionnements.

Base coût

Le traitement pour la base coût concerne uniquement le retranchement des coûts trop important dans la base qui pourraient nuire aux statistiques par la suite. Dans un premier temps, il apparaît que certains sinistres, en prenant leur coût total pour le groupe, présentent des montants négatifs dûs aux éventuels recours déduits du coût du sinistre. La part de coûts négatifs s'élève à plus de 6% de la taille totale de la base coût. Le poids de ce type de donnée reste trop peu important pour faire l'objet d'une étude spécifique. De ce fait, la solution

5. Les quelques polices correspondantes n'ont pas été traitée ici. La théorie des valeurs extrêmes, permettant le traitement de ce type de problème est considéré comme hors du périmètre de ce mémoire. En raison du faible nombre de polices tronquées, la problématique n'est pas traitée davantage ici.

adoptée a été d'affecter un coefficient de redressement aux sinistres positifs pour simuler une répartition de ces coûts négatifs, qui ne sont par la suite pas considéré dans la base de données des sinistres.

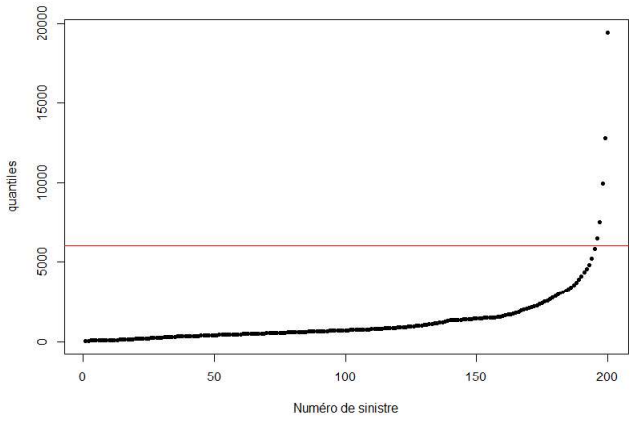

FIGURE 22 – Distribution des expositions

La représentation graphique des quantiles de coûts ne prend en compte que les coûts inférieurs à 20 000 en raison d'une visibilité réduite en cas de prise en compte des sinistres trop élevés (ce qui concerne 26 sinistres), ces sinistres auraient de toute manière été tronqués par la suite. La présence de ces coûts empêche en effet une bonne observation du seuil de décrochage à déterminer.

Le quantile à 95% nous renseigne d'une valeur de sinistre de 4895,21 euros pour le groupe. La limite finalement graphique finalement retenue est de 6000 euros, commune aux trois garanties. 165 sinistres sont de coût supérieur à ce seuil, ce qui représente près de 2,5% de la base coût totale[6]. Ils concernent en grande majorité la garantie RC, les deux autres garanties font voir une distribution des coûts beaucoup plus lissée.

Afin de prendre en compte le montant de ces sinistres puisqu'ils concernent en majorité une seule garantie, le total excédent le seuil déterminé est divisé, puis reporté sur chaque sinistre de la garantie RC non-excédentaire. Cette garantie se comportant différemment par rapport aux deux autres, il convient de ne pas mettre de côté ces sinistres plus onéreux.

Les polices et sinistres pouvant potentiellement poser problème étant écartées, il convient de s'intéresser au groupement des modalités de variables. Le clustering devant être effectué par la suite peut se faire par deux méthodes. La première consiste en la mise en place d'une classification ascendante hiérarchique, tandis que la deuxième se fait par des choix motivés par ceux historiquement réalisés par la profession.

3.1.2 Quelques mots sur la classification ascendante hiérarchique

Concernant la classification ascendante hiérarchique, il s'agit d'une méthode classique, utilisée dans le cadre d'analyses de bases de données, qui permet un apprentissage automatique non supervisé effectuant étape par étape des regroupements de modes d'une variable catégorielle. La méthode prend d'abord comme nombre de groupe le nombre de classes de la variable, puis les groupes englobent peu à peu davantage de classe, jusqu'à regrouper tous les modes de la variable. Cela explique le caractère ascendant (les groupes étant de plus en plus important) et hiérarchique (puisqu'à chaque nouveau groupe plus grand, un numéro d'étape est associé).

Chaque étape voit la fusion de deux groupes déjà formés, dont la désignation se fait par un processus d'optimisation dépendant de la méthode utilisée pour l'algorithme. Cette méthode correspond à une distance, à comprendre par sa définition mathématique, qui sera utilisée pour évaluer les écarts entre chaque groupe potentiel. Le regroupement finalement déterminé sera celui minimisant la distance.

Ici, il sera question de la distance euclidienne. Cette dernière est adaptée pour les variables quantitatives et permet de ne pas mettre excessivement l'accent les écarts pour les objets atypiques. La CAH sera effectuée

6. De la même manière que pour le modèle de fréquence, ces sinistres sont abandonnés puisque les techniques destinées à leur traitement sortent du cadre de ce mémoire. Qui plus est, le coût de ces sinistres ne déclenche pas le contrat XS souscrit auprès du réassureur, justifiant la décision de ne pas les retenir pour la base utilisée pour la tarification.

sur les fréquences des modalités de variables. Chaque modalité se verra associer une fréquence de sinistre total pondérée par l'exposition (sans prise en compte des garanties), puis la méthode de Ward sera alors appliquée pour obtenir les regroupements successifs à effectuer.

Soit $M = \{e_i \in [\![1, n]\!]\}$ un regroupement de n modalités, de fréquence moyenne pondérée par la distance f, et composé de k sous regroupements de modalités $M_1, ..., M_k$ d'effectifs $n_1, ..., n_k$. Soient maintenant $f_1, ..., f_k$ les fréquences moyennes pondérées par la distance de chaque sous-groupe. L'inertie interclasse I_e est alors définie par :

$$I_e = \sum_{i=1}^{k} n_i \times d(f_i, f)^2$$

La méthode de Ward consiste ainsi à déterminer le regroupement maximisant l'inertie interclasse pour chaque étape et regroupement possible. Les problèmes parfois posés concernent un sur-apprentissage qui peut apparaître durant la classification en cas d'effectifs trop faibles. En effet, cela induirait un ajustement trop poussé sur les données d'apprentissage, empêchant le modèle de fournir des prédictions efficaces en cas de confrontation à de nouvelles données. Elle sera ici testée sur deux variables différentes : les variables puissance et groupe. Ce sont des variables dont la valeur croît avec le risque associé. Naturellement, il serait attendu un regroupement par modalités croissantes. Les graphiques suivants montrent dans un premier temps le résultat du processus appliqué aux puissances et aux groupes de véhicules :

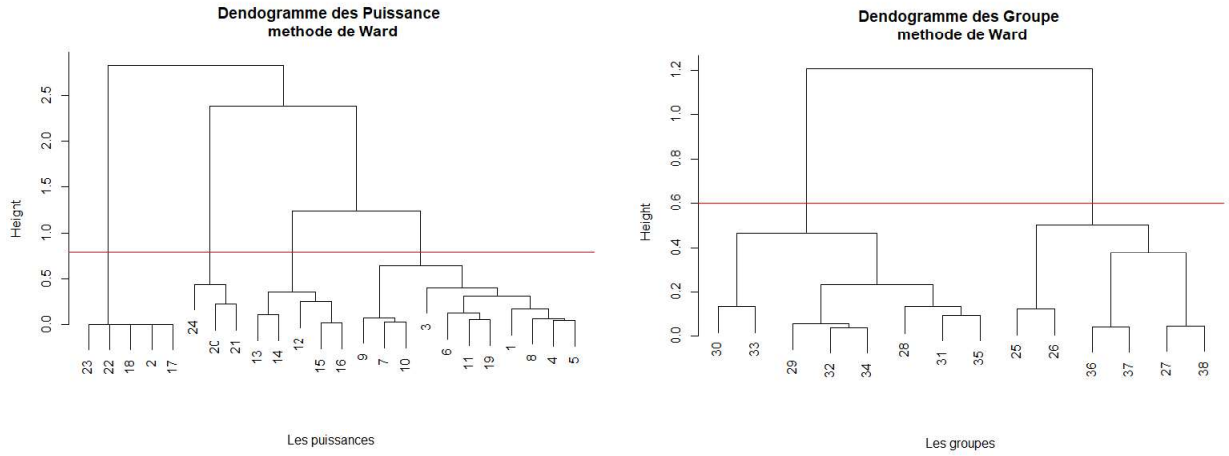

FIGURE 23 – Dendogrammes des puissances et des groupes

Concernant le dendogramme des puissances, il apparaît un regroupement effectué en premier lieu pour les puissances n'ayant pas fait apparaître de sinistre. Ces puissances correspondent en effet à très peu de polices, et ont été conservées dans l'étude afin d'appliquer la méthode de classification à un intervalle fermé. En observant les dendogrammes, il semble convenu de conserver deux groupes pour la variable groupe, et 4 pour la variable puissance. Cela nous donnerait donc la répartition suivante :

Puissances			
Groupe 1	Groupe 2	Groupe 3	Groupe 4
1 - 3 - 4 - 5 - 6 - 7 - 8 - 9 - 10 -11 - 19	2 - 17 - 18 - 22 - 23	12 - 13 - 14 - 15 -16	20 - 21 - 24

Groupes	
Groupe 1	Groupe 2
25 - 26 - 27 - 36 - 37 - 38	28 - 29 - 30 - 31 - 32 - 33 -

FIGURE 24 – Regroupement des modalités

Le groupement des modalités revient presque à une répartition croissante. En mettant de côté le groupement correspondant aux puissances où aucun sinistre n'a été constaté (groupe 2), seul le groupe 19 n'est pas dans la classe où il devrait être. Concernant les groupes, le découpage fait apparaître deux cluster : un avec les groupes les plus faibles et les plus élevés, et l'autre avec les groupes moyens. Les écarts constatés par rapport au résultat idéalement attendu s'expliquent par la masse de données trop peu importante. La conclusion résultant de ces observations est de ne pas retenir les résultats de ce type de clustering puisque les résultats peuvent être trop

éloignés de la réalité. Cependant, ils peuvent potentiellement aider à préciser le nombre de groupes à réaliser pour certaines variables.

3.1.3 Clustering par observations

Généralement, les clusters de modalités sont réalisés par expérience. En effet, les groupes sont souvent semblables entre les différents portefeuilles, et pour des variables communes Le processus de clustering ne sera pas décrit pour chaque variable, puisque pour certaines le mode d'action sera le même. La plupart des variables ont fait l'objet de l'observation de la fréquence de sinistre par modalité, et des regroupements correspondants ont été effectués en recoupant avec les résultats des dendogrammes obtenus précédemment afin de confronter les résultats. L'exemple est ici pris pour la mise en place des classes d'âge de véhicule.

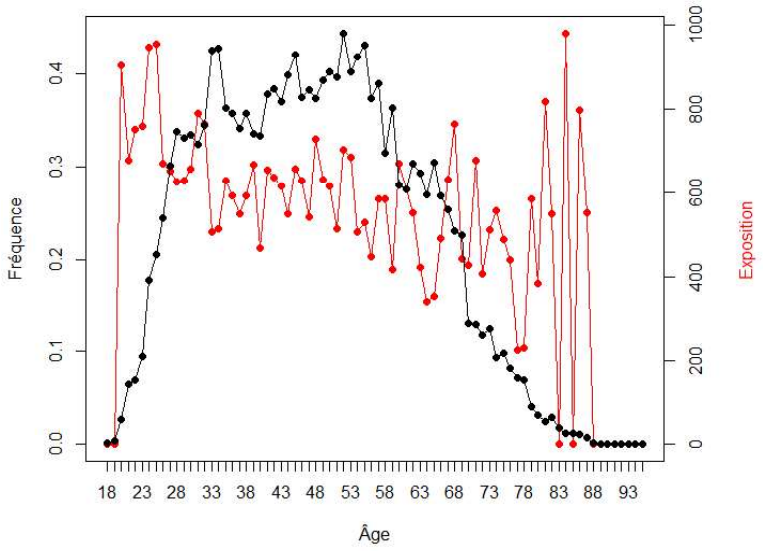

FIGURE 25 – Répartition des expositions et fréquences par âge

En raison de la forte volatilité de l'exposition (dont la trajectoire saccadée, usuellement plus lissée, s'explique par le manque de données), une segmentation par petit groupe d'âge est ici préférée. Cela se voit également sur la fréquence. Le résultat correspondant de l'application de la méthode de classification ascendante hiérarchique semble confirmer cette décision de grouper les âges par petits paquets, bien que les regroupements proposés ne soient pas en accord avec une classification par intervalles "continus". Ainsi, un regroupement par tranche d'âge de 5 ans est réalisé pour la suite de l'étude à partir de 25 ans et ce jusqu'à 75 ans, avec deux catégories regroupant les moins de 25 ans et plus de 75 ans. Les variables Âge du véhicule, Ancienneté du permis, Groupe, Puissance, Classe sont regroupées de la même manière.

Marques

De son côté, la variable marque est différente des autres puisqu'il n'y a pas réellement de caractère croissant à y retenir (contrairement à la classe, pour qui une lettre positionnée en aval de l'alphabet traduit un prix plus élevé). Un nombre non négligeable de marque se voyant associer une fréquence de sinistre nulle en raison du manque de données, la solution ici adoptée est celle du regroupement géographique. Cela correspond à faire cinq groupes : les marques américaines et anglaises, les marques françaises, les marques allemandes, les marques européennes autres et les marques asiatiques.

Location

La dernière variable posant une problématique différente des autres est la location géographique. Le groupe ciblant principalement des assurés venant d'une zone géographique précise, il est difficile d'appliquer la méthodologie correspondant à un zonier au niveau national ici. Dans un premier temps, il apparaît que les sept départements centralisant le plus d'exposition sont tous situés dans la même zone géographique. Ils représentent

à eux 7 plus de 85% de l'exposition totale du portefeuille. Avant d'approfondir l'information pour s'intéresser aux codes postaux, une première classification ascendante hiérarchique a été effectuée.

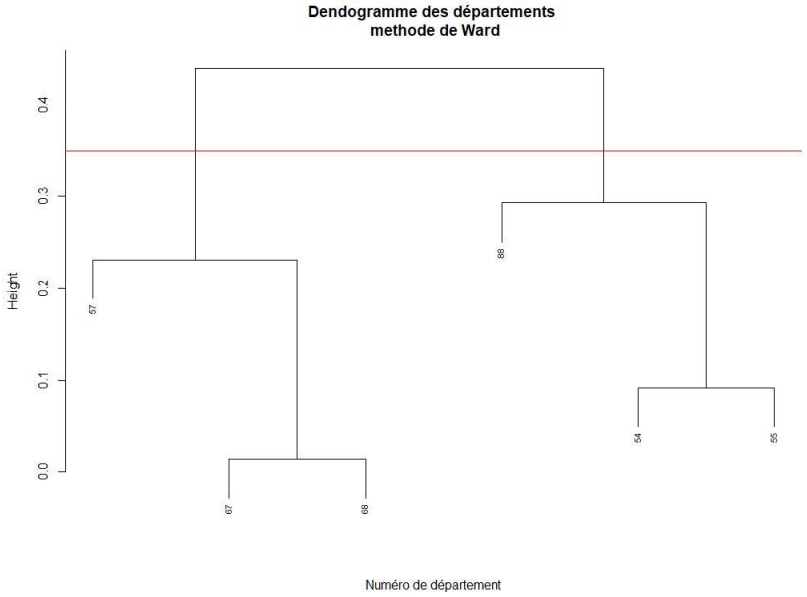

FIGURE 26 – Dendogramme des Départements

Un premier regroupement se fait. Le choix est ainsi donné entre conserver trois ou cinq classes (en comptant celle regroupant les départements ne faisant pas partie de l'étude). En s'intéressant à la position géographique de ces départements, il apparaît que le regroupement par trois apparaît assez homogène puisque les deux groupes seraient composés de trois départements adjacents. La réflexion peut alors être poussée en étudiant, pour les six régions majeures, les fréquences par sous-préfecture.

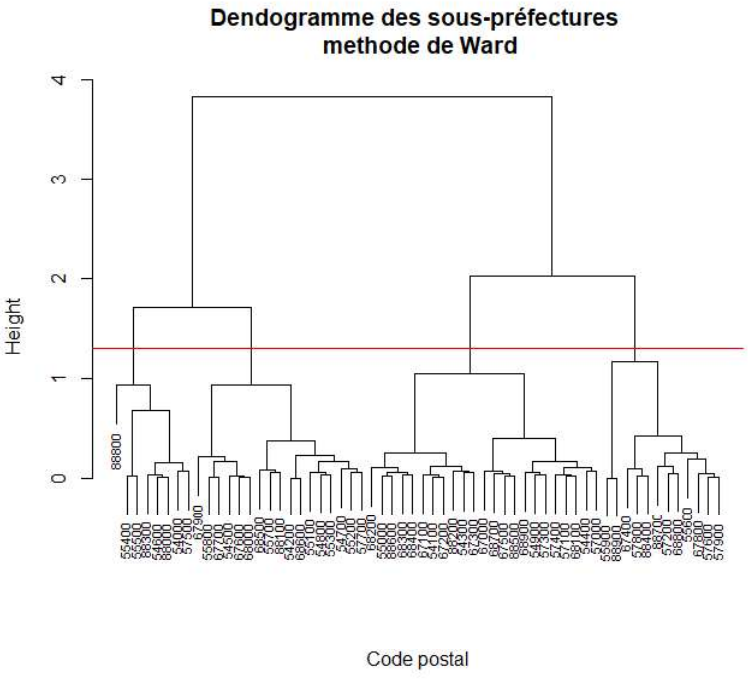

FIGURE 27 – Dendogramme des Sous-préfectures

La subdivision en groupes à partir de la ligne rouge suggèrerait une répartition en quatre groupes. Comme précédemment, en prenant une division supérieure, le clustering peut donner lieu à l'obtention de deux groupes. Il est intéressant de vérifier à quel point cela cadre avec la première répartition par région obtenue, notamment

puisque les deux groupes ont respectivement une taille de 25 et 35 sous-préfectures, et que le clustering par département donnait trois départements par groupe. Il est ainsi observé qu'il y aurait une répartition d'environ deux tiers / un tiers par rapport au groupement établi par département, c'est à dire que deux/tiers des sous-préfectures appartenant au premier groupe de département se trouveraient à nouveau groupées en affinant le clustering.

Si le chiffre peut paraître convenable, il n'est toutefois pas suffisant pour conclure à la suffisance de deux classes pour bien représenter l'origine géographique des polices de la base de données. Ainsi, il est convenu de conserver quatre classes, avec en sus une cinquième correspondant aux sous-préfectures étant hors des six principaux départements.

3.1.4 Études de dépendance

Bien que la tarification se fasse dans un souci d'exhaustivité, elle est également empreinte d'une volonté d'absence de redondance : dans le cas où deux variables donneraient la même information, il est préférable de n'en conserver qu'une seule afin d'alléger le GLM au maximum pour en optimiser son efficacité. Dans cette optique, les corrélations éventuelles doivent faire l'objet d'une analyse poussée. Les clusters de variables ayant été créés, il conviendra de les comparer entre eux, ainsi qu'avec les différentes variables propre à la base de données retenues pour le processus de tarification.

La méthode retenue est le V de Cramér, qui s'applique aux variables quantitatives. Il s'agit d'un test de dépendance, qui n'est pas sensible à l'effectif total de l'échantillon et qui permet de déterminer de manière assez précise quelles variables présentent des liens entre elle. Plus la valeur obtenue est proche de 0, moins il y a de dépendance observée entre les deux variables.

Soient N le nombre total d'observation au sein de la base. Pour chaque couple de variable, une table de contingence est créée. Cette table est de taille $r \times c$, où r est le nombre de lignes / de modalités de la première variable, et c est le nombre de colonnes / de modalités de la seconde variable. Ainsi, chaque élément $O_{i,j}$ de la table va renseigner l'effectif de la base où les modalités i de la première variable et j de la seconde seront simultanément renseignées. Soit maintenant $E_{i,j}$ l'effectif théorique du croisement des modalités correspondantes sous hypothèse d'indépendance, avec $E_{i,j} = N \times p_i \times p_j$ et p_i la probabilité d'appartenir à la modalité i pour la première variable, p_j celle d'appartenir à la modalité j pour la seconde.

Enfin, soit la statistique du χ^2 définie par $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$. Le V de Cramér est alors donné par la formule suivante :

$$\sqrt{\frac{\chi^2}{N \times (r-1) \times (c-1)}}$$

En effectuant le test sur R, la figure résultante est donnée sur la page suivante. Les corrélations des variables suivantes sont étudiées :

1. GroupeAgeVeh : Groupes des âges de véhicule
2. GroupeClasse : Groupes des classes de véhicule
3. GroupeAge : Groupes des âges des assurés
4. GroupePerm : Groupes des anciennetés de permis
5. GroupeGroupe : Groupes des groupes de véhicule
6. GroupePuissance : Groupes des puissances
7. Marque : Groupes des marques de véhicule
8. Garage : Détention d'un garage ou non
9. SR : Nombre de sinistres responsables
10. SNR : Nombre de sinistres non responsables
11. KM : Formule kilométrique
12. GroupeDep : Groupes des départements

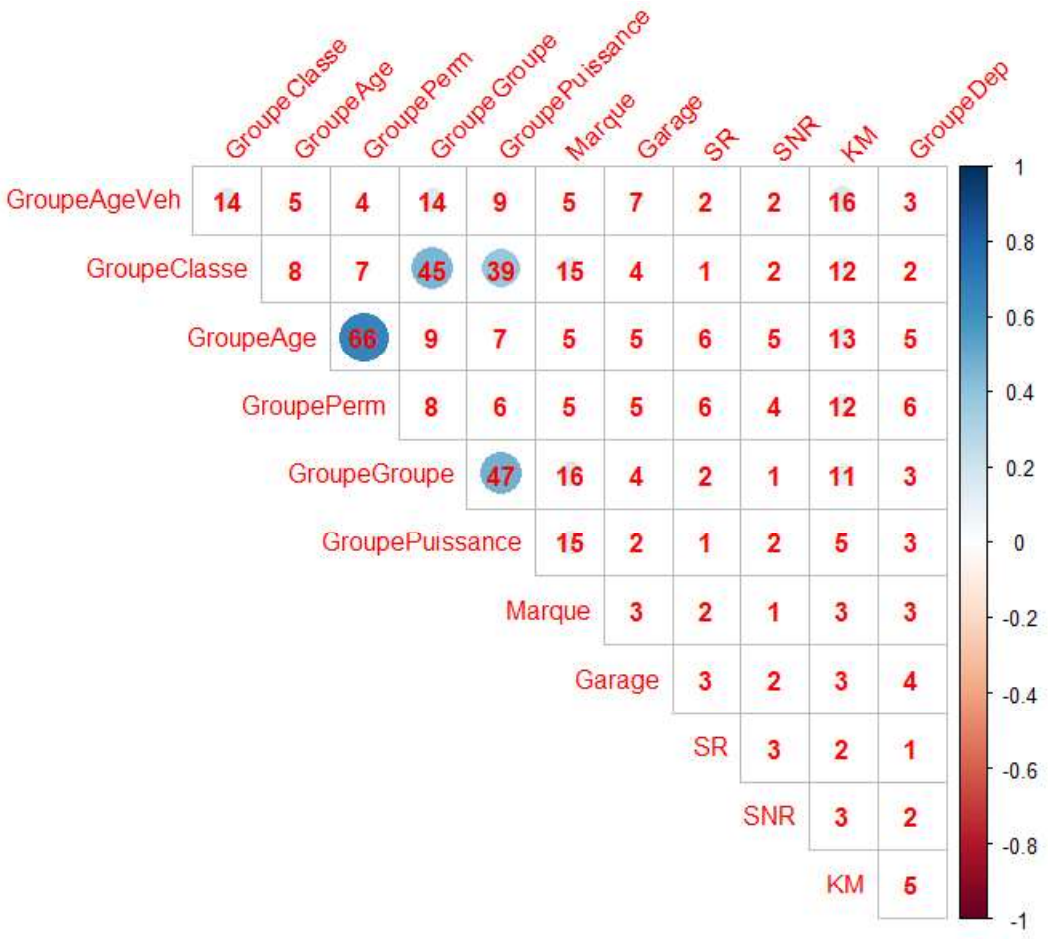

FIGURE 28 – Triangle de corrélations entre les variables retenues

Des corrélations attendues y apparaissent :

- Les variables groupe âge et groupe permis sont corrélées, ce qui atteste également du fait que l'information perdue en regroupant les modalités de ces deux variables n'est pas trop importante puisque la corrélation subsiste.
- De la même manière, des corrélations existent entre les groupes, les classes et les puissance des véhicules. Elles sont moins marquées mais existantes, et une fois de plus amène à penser que la perte d'information des regroupements a bien été limitée.
- Aucune autre corrélation n'est observée. Les autres caractéristiques des véhicules ne font pas voir de liens significatifs entre elles. Encore une fois, il s'agit de résultats attendus par rapport aux études classiques de tarification.

Au vu des résultats, et en fixant un seuil de dépendance à 0,5, usuel dans le contexte, la variable GroupePerm ne sera pas conservée pour la base fréquence. Il est clair que les corrélations des GroupeClasse, GroupeGroupe et GroupePuiss ne sont pas liées au hasard. Il pourrait donc être intéressant de n'en conserver que deux. Toutefois, comme cela sera montré par la suite, elle n'apparaissent jamais tous les trois dans le modèle GLM finalement adopté.

Les corrélations provenant de la base coût, de la garantie Responsabilité Civile, avec des clustering similaires à ceux effectués sur la base fréquence sont également étudiés.

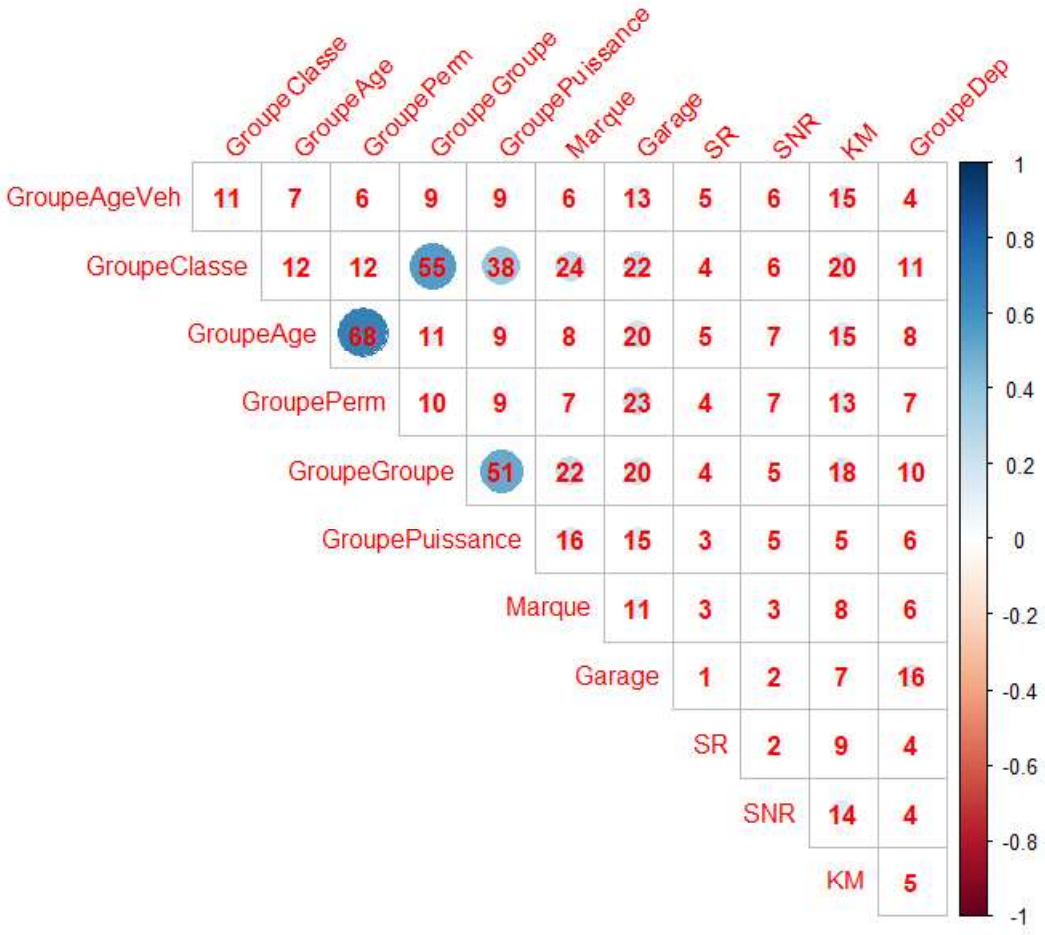

FIGURE 29 – Triangle de corrélations entre les variables retenues

Il y apparaît des corrélations semblables par rapport à celles de la base fréquence. Les conclusions sont donc presque les mêmes, toutefois, la variable GroupeGroupe ne sera cette fois-ci pas conservée en raison de la forte corrélation observée avec les variables GroupeClasse et GroupePuiss. De la même manière, la variable GroupePerm n'est pas non plus conservée.

3.2 Gestion des données manquantes : complétion de la variable Formule Kilométrique

3.2.1 Pré-traitement statistique

Comme énoncé précédemment, la variable Formule Kilométrique apporte des informations cruciales pour préciser le profil de l'assuré. En reprenant le problème des assurés où la formule n'est pas renseignée, les études des variables croisées vont permettre d'affiner le profil des assurés n'ayant pas souscrit à une formule précise. À noter également, l'étude de corrélation de la variable en question avec les autres permet de cibler quelles variables sont plus à même d'être croisées. La formule kilométrique ne présente aucune corrélation importante avec les autres variables de l'étude. Les plus importants coefficients de corrélation correspondent à l'âge et à l'âge du véhicule. Une étude par rapport à l'importance de la classe renseignée a également été explorée pour plus de précision.

L'intérêt aurait également pu être porté sur le Code Postal. En effet, cela laisserait penser que la localisation influerait sur la Formule Kilométrique retenue. Dans le cas où l'assuré résiderait dans une grande ville par exemple, il est possible qu'il souscrive à un formule impliquant une distance annuelle parcourue plus faible. Toutefois, en créant une variable s'intéressant à la taille de la ville, il s'avère que la corrélation avec la Formule Kilométrique n'augmente pas, et l'idée a ainsi été abandonnée.

Les études croisées ont donc été menées, et ont fourni le tableau suivant :

Formule Kilométrique	0	1	2	3	4
Âge de véhicule moyen	14,6	9,5	7,9	5,6	12,0
Âge moyen	52,3	47,0	42,3	45,1	53,3
Indicateur de niveau de classe	2,67	2,93	3,09	3,17	2,71

FIGURE 30 – Statistiques des différentes formules

S'il n'est pas réaliste de chercher à déterminer un modèle à partir de ces seules variables pour déterminer les formules kilométriques correspondant aux lignes où la modalité n'est pas renseignée, la simple moyenne fait voir un âge moyen plutôt élevé, ainsi qu'un âge de véhicule moyen plutôt élevé par rapport à celui des formules 1, 2 et 3. À noter que l'indicateur du niveau de classe correspond à la moyenne de numéro de groupe de classe, ces dernières ayant été réparties en cinq groupe distincts. Il faut donc y comprendre que globalement, les profils des assurés où la formule kilométrique n'est pas renseignée disposent de véhicules à classes faibles en moyenne, et semblent plus proches de ceux ayant souscrit à la formule "0".

Ces chiffres sont davantage des indicateurs de la marche à suivre que des résultats qui seront par la suite utilisés pour résoudre le problème. En moyenne, les lignes où il n'y a pas de formule kilométrique renseignée ont donc des caractéristiques proches de celles où la formule "0" est renseignée. Si aucune tendance ne s'était dégagée, la formule aurait pu être déterminée au hasard au prorata du poids de chaque madalité dans le portefeuille. Dans le cas présent, des travaux supplémentaires sont nécessaires.

3.2.2 Une solution potentielle : la méthode CART

De manière générale, les arbres de décision sont utilisés afin de gérer les situations où des données seraient manquantes. Le principal argument justifiant l'utilisation d'une telle méthode ici est sa lisibilité : l'objectif de ce mémoire étant également de proposer des méthodes assez claires pour les non-initiés aux différentes méthodes, il s'avère que les arbres de décisions produisent des résultats graphiques assez clairs.

Le principe repose sur la construction d'une classification hiérarchique descendante des observations. Chaque classe se voit par la suite attribuer la valeur moyenne des sorties des observations contenues, ou, dans le cas d'une variable qualitative, la modalité la plus fréquente.

L'application de la méthode CART va ainsi poser trois questions :

— Comment définir les subdivisions de l'arbre ?
— À partir de quel moment faut-il arrêter de chercher à étendre l'arbre ?
— Comment déterminer la valeur de la variable à expliquer ?

Afin de bien comprendre les réponses à ces trois questions, et dans l'optique de s'imprégner du fonctionnement de la méthode CART, il a été choisi de développer la théorie mathématique régissant cette dernière.

Dans un premier temps sont introduites les notations suivantes :

— $P(j,t)$: la probabilité qu'une observation soit dans le nœud t et de classe j.
— $P(t)$: la probabilité qu'une observation soit dans le nœud t.
— $P(j|t)$: la probabilité de la classe j sachant que l'observation est dans le nœud t.

Critère de construction
L'intérêt se porte ainsi sur le critère de construction qui amènera à déterminer quelles divisions doivent être effectuées au sein de l'arbre. Soit t un **nœud** de l'arbre, t_g et t_d les nœuds suivants, respectivement à gauche et à droite du nœud initial, résultants d'une **division** δ. Les proportions d'observations envoyées respectivement dans t_g et t_d sont notées $p_g = \frac{p(t_g)}{p(t)}$ et $p_d = \frac{p(t_d)}{p(t)}$.

Par la suite est introduite la notion de **fonction d'hétérogénéité**. Soit h une fonction de $\{(p_1, ..., p_J | p_i \geq 0), \sum_j p_j = 1\}$ avec J le nombre de classes possibles pour l'individu, et les p_i la proportion d'observations appartenant à la classe i. h est une fonction d'hétérogénéité si :

— h est symétrique en $p_1, ..., p_J$
— h est maximale en $(\frac{1}{J}, ..., \frac{1}{J})$
— h est minimale en $(1, 0, ..., 0)$, $(0, 1, 0, ..., 0)$,...,$(0, ..., 0, 1)$

L'hétérogénéité de t est alors donnée par $i(t) = h(p(1|t), ..., p(J|t))$ avec h une fonction d'hétérogénéité. Enfin, est définie la notion de **variation d'hétérogénéité** générée par δ

$$\Delta i(\delta, t) = i(t) - p_g i(t_g) - p_d i(t_d)$$

Ce qui permet enfin d'en conclure que la division optimale du nœud t est donnée par :

$$\delta^*(t) = argmax_\delta(\Delta i(\delta, t))$$

Règle d'arrêt
L'introduction de la règle d'arrêt est plus immédiate. En effet, un nœud t d'un arbre sera déclaré terminal si :

— Une seule observation se trouve dans le nœud t
— Toutes les observations dans t portent le même label.

Lors de l'application de la méthode, il est possible dans R d'augmenter le nombre d'observation minimal à atteindre dans un nœud à l'aide de la fonction rpart du package du même nom. Ainsi, si la valeur est originellement fixée à 1, il est possible de ne pas descendre jusqu'à des groupes de moins de 100 observations par exemple.

Règle d'assignation
Soit t un nœud ayant $j(t)$ comme réponse associé. La probabilité de mauvais classement du noeud t est donnée par

$$r(t) = \sum_{j \neq j(t)} p(j|t)$$

Il vient alors que la réponse $j(t)$ définie par $j(t) = argmax_{j \in \mathcal{J}} p(j|t)$ minimise $r(t)$.

Élagage de l'arbre
Le problème de la méthode appliquée ainsi, notamment en prenant une règle d'arrêt où le nombre d'observation minimal à obtenir dans le nœud va être faible, est que le résultat sera un arbre extrêmement lourd à traiter. En effet, les ramifications seront trop nombreuses pour consister en un résultat clair et utilisable. Qui plus est, puisque tout le modèle sera basé sur la même base de données, il est probable d'observer un phénomène de sur-apprentissage, puisque la méthode s'intéressera à des conclusions trop précises pour ce qui pourrait s'avérer être des cas particuliers.

Ainsi, il convient de convenir d'un critère permettant d'élaguer l'arbre au maximum pour en obtenir le modèle idéal. Élaguer un arbre T revient à s'intéresser à un nœud t non terminal, et à priver l'arbre de tous les nœuds résultants de T, donnant ainsi lieu à un arbre T'. Le but est alors de créer une suite de sous-arbres, dont l'optimal est déterminé par comparaison des erreurs en validant avec un échantillon test ou en procédant par validation croisée. L'objectif étant de minimiser les erreurs.

Par cette procédure, le résultat obtenu est alors un arbre bien plus lisible, et mieux adapté à de nouvelles données.

Application et remarques
Avant de procéder à la méthode CART, il est nécessaire de faire la liste des variables pertinentes pour expliquer la Formule Kilométrique. Toutes les variables n'ont pas été implémentées afin d'obtenir l'arbre. Seules onze d'entre elles ont été intégrées au modèle en plus de la Formule Kilométrique, en se basant sur les potentielles corrélations avec la variable à expliquer ainsi que sur des choix d'expert.

La base de donnée a été séparée en deux bases, une base d'apprentissage et une base de test, la base d'apprentissage prenant 70% des valeurs de la base entière. Le but est ici d'obtenir un premier aperçu de l'efficacité de la méthode.

L'arbre obtenu par la suite via R, après élagage, est schématisé ainsi (à noter que ce n'est pas l'arbre résultant d'une base d'apprentissage aléatoire, mais l'arbre prenant en compte l'ensemble des données, puisque ce sera celui-ci qui sera utilisé par la suite) :

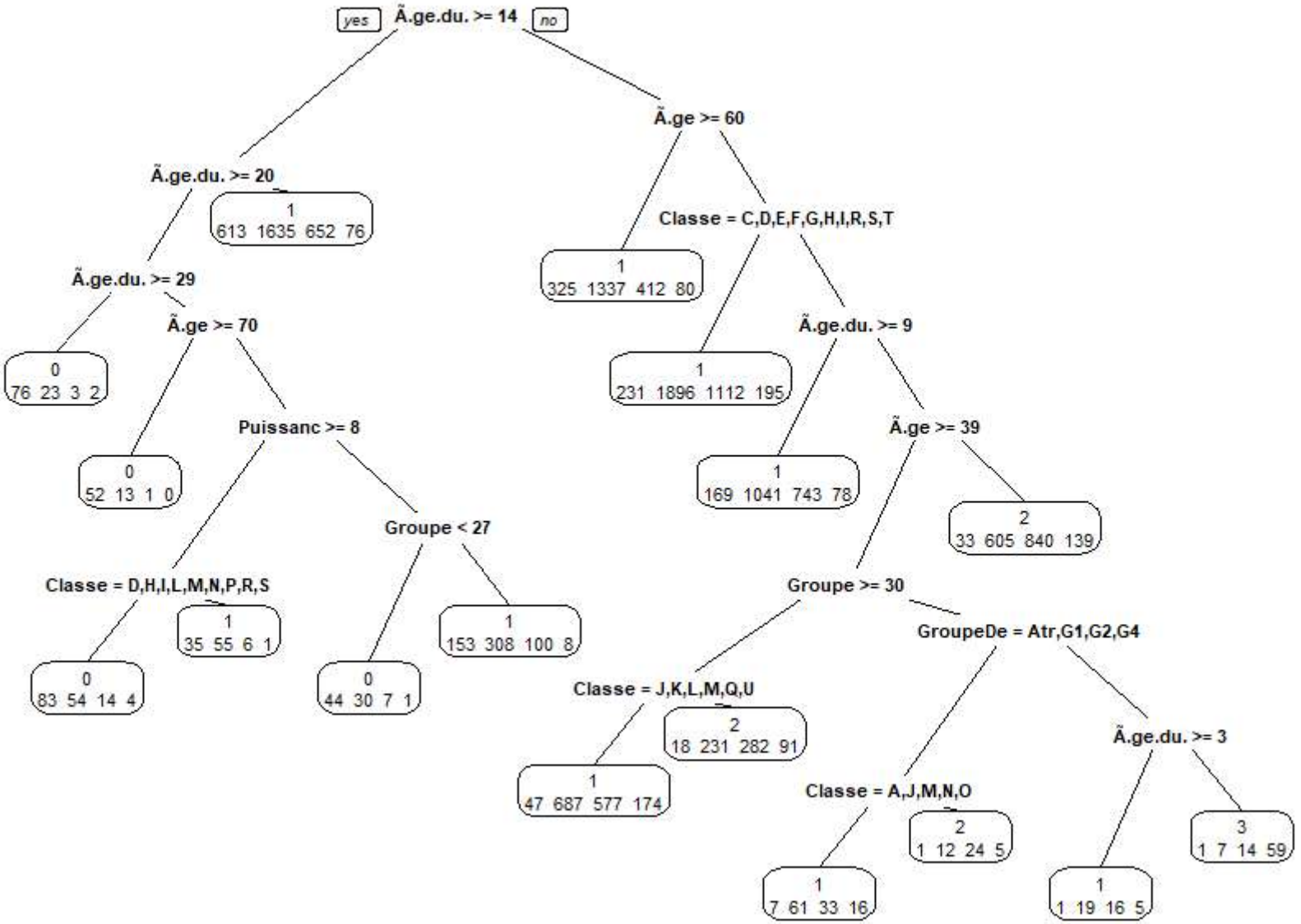

FIGURE 31 – Arbre CART

En observant les ramifications, il ressort que sex variables ont été utilisées pour former les classes :

— L'âge du véhicule
— La classe du véhicule
— La puissance du véhicule
— L'âge de l'assuré
— Le groupe du véhicule
— Le département (après Clustering)

Ces variables ont été retenues par la méthode puisque, comme cela apparaît sur la matrice de corrélation, elles sont les plus corrélées avec la Formule Kilométrique. Il est intéressant de voir que la variable regroupant les département a également été utilisée, malgré sa faible corrélation. Une piste qui expliquerait sa présence parmi les variables précisant l'arbre est que, après des tris préliminaires, la corrélation puisse augmenter vu que l'étude est alors portée sur un profil plus précis d'assurés.

L'application du modèle résultant de l'application de la méthode sur la base d'apprentissage à la base test donne lieu à une qualité de prédiction de près de 55%. Le chiffre peut paraître faible, toutefois, expliquer une variable telle que la Formule Kilométrique avec celles mises à disposition présente des difficultés évidentes en raison du manque de corrélation. C'est d'ailleurs ce qui justifie son intégration en tant que paramètre dans le GLM, puisqu'elle influe évidemment sur la fréquence de sinistre (et donc sur la prime finale à demander à l'assuré), et qu'elle n'est pas facilement reproductible avec les autres variables.

Des résultats intéressants ressortent, notamment le fait que si l'âge du véhicule est élevé et que l'âge de l'assuré l'est aussi, la formule résultant est la 0. Cela rejoint les résultats du pré-traitement statistique qui montrait que ce groupe avait une moyenne d'âge importante, que ce soit pour le véhicule ou pour l'assuré. Le profil des assurés où la classe n'était pas renseignée semblait se rapprocher de ce groupe, il faudra donc vérifier que la part y soit plus importante, en considération de la répartition globale qui tend à favoriser la Formule Kilométrique 1.

S'il pourrait toutefois s'avérer envisageable d'appliquer une méthode ayant une précision supérieure à 50% (en tout cas dans le cas où la variable peut prendre 4 valeurs différentes), il faut bien voir qu'elle associera la majorité des assurés à la Formule Kilométrique 1, puisque cette dernière est la plus fréquente au sein de la base. Dans l'échantillon de test, l'efficacité est bonne pour la classe 1, mais est très mauvaise pour les trois autres classes, négligeables par rapport à celle-ci.

En appliquant maintenant un modèle prenant l'ensemble des données disponibles à la part de la base où la formule n'est pas renseignée, la répartition suivante est obtenue :

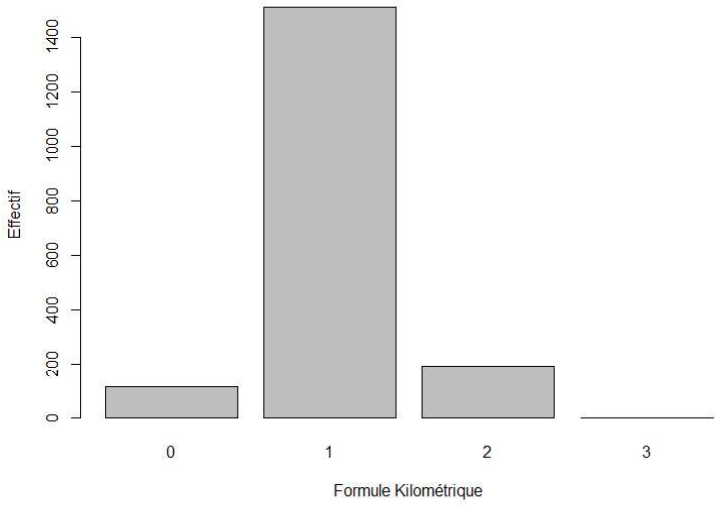

FIGURE 32 – Répartition des polices où la Formule Kilométrique n'était pas renseignée

La répartition montre un pic potentiellement trop accentué pour la Formule Kilométrique "1". Les formules "0"

et "3" sont particulièrement sous-représentées par rapport à ce qui était observé dans la répartition d'origine de la base de données. Cette approximation était attendue en raison de la précision de la méthode obtenue, mais apparaît ainsi trop peu précise pour être utilisée avec confiance.

3.2.3 Une méthode plus précise : Random Forest

L'intérêt est ainsi porté sur une nouvelle méthode également très utilisée dans le milieu de la data science : la méthode Random Forest, présentée en 2001 par le statisticien L. Breiman [2]. Il s'agit d'une variante d'une méthode d'agrégation (*bagging*) appliquée à la méthode CART. Le but étant de simuler plusieurs arbres avec un champ de variables pré-sélectionnées aléatoirement parmi celles disponibles dans l'optique de diminuer la variance liée à la réalisation d'un seul arbre, et à obtenir une méthode efficace pour expliquer une variable avec un grand nombre d'autres variables explicatives. Il est important de voir que cette solution présente plusieurs avantages par rapport à la méthode CART, ainsi qu'un inconvénient majeur :

— Contrairement à la méthode CART, la Random Forest présente moins de risque de sur-apprentissage.
— La précision est plus fine. En effet, l'élagage de la méthode CART n'est pas nécessaire pour la Random Forest.
— La méthode CART permettait une visualisation précise du processus suivi, ce n'est pas le cas pour la Random Forest. Impossible par exemple d'obtenir une représentation graphique de l'application de la méthode.

Concernant cet aspect "boîte noire", il faut comprendre ici que la méthode Random Forest ne donne guère plus d'information sur son fonctionnement interne que les entrées et sorties. Cela rend notamment l'explication et l'utilisation plus difficile. Bien que ce problème ne soit pas négligeable, en appliquant la Random Forest à la base de donnée, la répartition apparaît comme étant bien meilleure en prenant comme base d'apprentissage et de test la base globale (plus de 83% de précision, ce qui correspond a un gain de près de 30% par rapport à la méthode CART, même en prenant les deux mêmes bases pour l'apprentissage et le test).

Les erreurs sont toujours principalement situées pour les formules kilométriques "0" et "3", mais le phénomène est grandement atténué par rapport aux résultats précédents. En effet, il semble à première vue que les formules aux extrémités sont mieux prédites (pour les formules "1" et "3" respectivement, la méthode Random Forest donne 45% et 13% de précision contre 4% et 6% avec la méthode CART). À noter que cette fois-ci, la variable a été traitée en variable numérique (bien que ce ne soit pas réellement le cas), puis la formule associée a été déterminée en prenant l'arrondi de la réponse. Le même processus a été effectué avec la méthode CART mais cela n'améliorait pas les résultats.

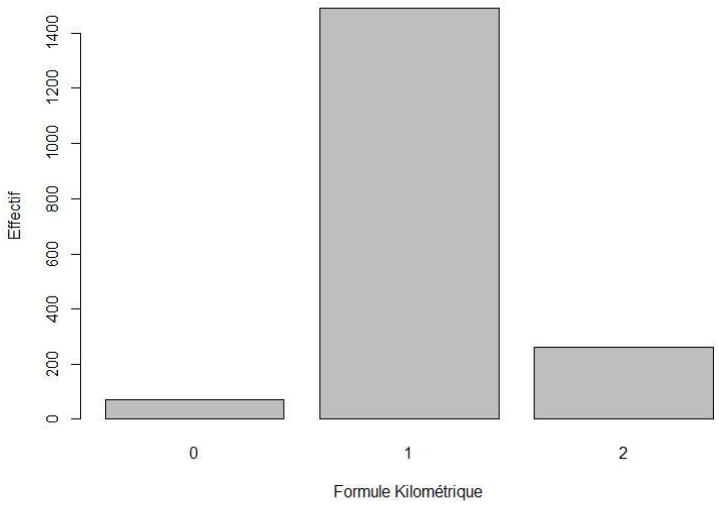

FIGURE 33 – Répartition des polices où la Formule Kilométrique n'était pas renseignée

Malgré l'amélioration de la précision, le modèle ne laisse pas voir un changement significatif quant à la détection des classes "0" et "3". Toutefois, les statistiques montrent notamment que le profil s'approche de ceux ayant

souscrit aux Formules kilométriques "0" ou "1". Il est donc considéré acceptable d'appliquer une telle répartition, bien que la classe "0" semble sous-représentée. La Formule Kilométrique des assurés où la variable n'est pas remplie sera donc complétée via la méthode Random Forest pour les études concernant le GLM.

3.3 Utilisation de modèles linéaires généralisés

3.3.1 Rappels théoriques

Expression du modèle général

Les GLM (Generalized Linear Model) consistent en l'approche traditionnelle d'une problématique de tarification. Leur mise en place en 1972 par Nelder et Wedderburn résulte d'un besoin de résoudre les problèmes liés à l'application des modèles linéaires. Ces derniers présentent en effet le défaut de ne fonctionner que sous certaines limitations, comme la densité gaussienne de la variable à expliquer ou son homoscédasticité (les variances des erreurs stochastiques d'une régression linéaire doivent être égales) par exemple. Malgré la possibilité d'adapter la variable pour mieux répondre à ces contraintes, l'application de la méthode posait plusieurs difficultés significatives.

L'application des GLM permettaient alors de travailler sur des variables dont la loi faisait partie de la famille exponentielle linéaire (et qui ne devait plus nécessairement être normale). Les techniques liés ont ainsi largement supplanté celles des modèles linéaires classiques.

Sa philosophie est basée sur l'utilisation d'un certain nombre de variables explicatives $X = (X_1, ..., X_p)^t$. Elle consiste en l'estimation d'une fonction g telle que la relation suivante, pour une variable aléatoire Y, soit vérifiée :

$$g(\mathbb{E}[Y|x_1, ..., x_p]) = \sum_{k=1}^{p} \beta_k x_k$$

Cette fonction g est appelée fonction de lien du modèle, elle est strictement monotone et dérivable. Elle réalise le lien entre l'espérance conditionnelle et le modèle linéaire de prédiction. En prenant maintenant pour hypothèse que Y suive une loi de Poisson, il est possible de s'intéresser au cheminement mathématique que cela induit.

Dans un premier temps, la famille exponentielle est définie par la densité suivante :

$$f_{\theta,\varphi}(y) = \exp\left(\frac{y \times \theta - b(\theta)}{\varphi} + c(y, \varphi)\right)$$

Il convient de montrer que la loi de Poisson appartient à la famille exponentielle.

Si Y suit une loi de Poisson, alors $\mathbb{P}(Y = y) = \frac{\lambda^k e^{-\lambda}}{k!}$. En l'écrivant sous une autre forme, il en résulte que $\mathbb{P}(Y = y) = exp(y \ln(\lambda) - \lambda + c(y))$. En posant $\theta = ln(\lambda)$, $\Phi = 1$ et $b(\theta) = exp(\theta)$ il apparaît maintenant que la loi appartient bien à la famille exponentielle. Cela permet notamment de faire le lien entre le paramètre de la loi et les variables explicatives :

$$\theta(x) = b'^{-1}(E(Y|x)) = b^{-1}\left(g^{-1}\left(\sum_{k=1}^{p} \beta_k x_k\right)\right)$$

Dans le cas présent d'une tarification d'un contrat d'assurance automobile, il est question de la variable exposition (exposure) qui s'intègre dans le modèle comme une variable *offset*. À l'origine le modèle s'écrit de la façon suivante :

$$\mathbb{E}[N|X] = \exp\left(\sum_{k=1}^{p} \beta_k x_k\right) = \exp(\beta' x)$$

Ainsi, en introduisant la variable exposure comme facteur de la variable à expliquer (ici la prime), le modèle s'adapte et devient :

$$\mathbb{E}[N|X, e] = e \times \exp\left(\sum_{k=1}^{p} \beta_k x_k\right) = \exp(\beta' x + \ln(e))$$

Cela revient alors à rajouter une variable explicative au modèle à l'exception du fait que le β serait connu puisqu'il serait égal à 1.

En pratique, le principe revient donc à déterminer la distribution des erreurs (et des réponses) en fonction du type de réponse. Pour le modèle de fréquence, il s'agit d'une réponse de comptage (renvoyant le nombre de sinistres) alors que pour le modèle de coût la réponse est quantitative continue. Cela peut induire l'utilisation de distributions différentes. Pour chaque modèle, il conviendra de comparer la densité des lois potentielles à celle de la réponse, afin de retenir la mieux adaptée. Généralement, la fonction de lien retenue est la fonction logarithmique, cependant le choix peut différer selon la distribution des réponses.

Estimation des paramètres

Par la suite, l'intérêt est de nouveau porté sur les p variables explicatives $X^1, ..., X^p$. Il en découle la nécessité de l'estimation de $\beta = (\beta_0, ..., \beta_p)$.

L'estimation se fera par le principe du maximum de vraisemblance, permettant d'estimer les paramètres optimisant la vraisemblance de l'échantillon utilisé. Soit f la fonction de distribution retenue. Soit un échantillon de n variables indépendantes Y_i pour $i \in [1, ..., n]$, respectivement d'espérance m_i avec $g(m_i) = \nu_i$. Les expressions de la fonction de vraisemblance et de son logarithme sont données ainsi :

$$L(\beta) = \prod_{i=1}^{n} f(y_i, \theta_i, \phi)$$

$$LogL(\beta) = \sum_{i=1}^{n} ln(f(y_i, \theta_i, \phi)) = \sum_{i=1}^{n} l(y_i, \theta_i, \phi)$$

En s'intéressant à l'expression des dérivées partielles de la log-vraisemblance, les équations de vraisemblance peuvent être obtenues. Elles s'écrivent :

$$\forall j = 1, ..., p, \sum_{i=1}^{n} \frac{(y_i - m_i)x_i^j}{V(Y_i)} \frac{\partial m_i}{\partial n_i} = 0$$

Par la suite, il n'est pas possible de donner une expression précise pour les estimateurs. Des méthodes sont ensuite envisageables dans l'optique de déterminer la solution de ces équations. Il s'agit de l'algorithme de Newton-Raphson et de l'algorithme du score de Fisher. L'exposition de ces méthodes sortant du cadre de ce mémoire, elle ne sont pas développées davantage ici.

3.3.2 Modèle de fréquence

Les deux modèles respectifs de coût et de fréquence seront décrits par la suite. Il est proposé de commencer avec le modèle de fréquence, dont la mise en place de la base a été décrite ci-avant. La variable à expliquer est donc le nombre de sinistre, qui sera par la suite multipliée par le coût moyen pour obtenir la prime pure. Il est de rigueur de procéder à la méthode du *one hot encoding*, qui revient à coder chacune des modalités de la variable en 1 si l'individu présente la caractéristique en question, ou 0 sinon. Une modalité de référence (la plus fréquente) est déterminée et n'est donc pas encodée dans le modèle.

Il y a deux variables *offset* à ajouter au modèle. Il s'agit de l'exposition et du CRM (Coefficient de Réduction-Majoration). La première est essentielle à la mise en place d'un modèle de fréquence, puisque cette dernière dépend forcément de la durée d'exposition. Concernant la seconde, il s'agit d'une contrainte règlementaire dont l'inclusion est obligatoire. À noter qu'en pratique le CRM est utilisé en aval du processus de tarification.

Sélection des variables

Les variables sélectionnées pour le modèle sont les suivantes :

1. Exposition
2. CRM
3. Marque
4. Groupe
5. Puissance
6. Classe
7. Ancienneté du véhicule
8. Département
9. Âge
10. Formule kilométrique
11. Garage
12. Enfants

Celles qui ne sont pas binaires ont été regroupées par modalités comme certains exemples le décrivent dans la partie concernant les clusterings. Concernant chaque variable, une modalité de référence est sélectionnée, ne laissant apparaître dans la matrice des effets que les modalités autres que celle de référence. À ces modalités s'ajoutent la constante *intercept* qui traduira de par son coefficient l'impact des modalités de référence retenues.

La sélection s'est principalement effectuée en considération des corrélations observées précédemment, et du potentiel explicatif de la variable en elle-même.

Choix de la Distribution

Les travaux concernant le choix de la Distribution est ici décrit pour la mise en place du modèle de fréquence de la garantie DTA. Le but est de s'intéresser dans un premier temps à la répartition de la variable NB.sin présente dans la base fréquence. Soit m et v les moyennes et variances de la série considérée.

L'intérêt est alors porté sur trois lois qui pourraient être utilisées pour le modèle final. Les lois binomiales, binomiales négatives et de poisson peuvent être envisagées. L'objectif est de paramétrer les lois telles que leur variance et leur espérance soit égale à celle des données retenues.

Loi binomiale : Soit $X \sim \mathcal{B}(n,p)$. Il en résulte que $\mathbb{E}(X) = np$ et $\mathbb{V}(X) = np(1-p)$.
D'où $np = m \Leftrightarrow v = m(1-p) \Leftrightarrow p = 1 - \frac{v}{m}$ et $n = \frac{m}{1-\frac{v}{m}} = \frac{m}{\frac{m-v}{m}} = \frac{m^2}{m-v}$

Loi binomiale négative : Soit $X \sim \mathcal{J}(r,p)$. Il en résulte que $\mathbb{E}(X) = \frac{r(1-p)}{p}$ et $\mathbb{V}(X) = \frac{r(1-p)}{p^2}$.
D'où $\frac{r(1-p)}{p} = m \Leftrightarrow v = \frac{m}{p} \Leftrightarrow p = \frac{m}{v}$ et $r = \frac{mp}{1-p} = \frac{\frac{m^2}{v}}{\frac{v-m}{v}} = \frac{m^2}{v-m}$

Loi de poisson : Soit $X \sim \mathcal{P}(\lambda)$. Avec $\mathbb{E}(X) = \mathbb{V}(X) = \lambda$, d'où $m = \lambda$ sans qu'il soit possible de régler la variance.

La loi binomiale est peu adaptée puisque la définition mathématique induit une valeur de n entière. Il reste les lois binomiales négatives et de poisson qui peuvent être utilisables dans le cadre de la modélisation. En traçant les densités de simulations de ces lois avec les paramètres précédemment exposés, et en y superposant la densité obtenues avec les données d'origine, le graphique suivant est obtenu :

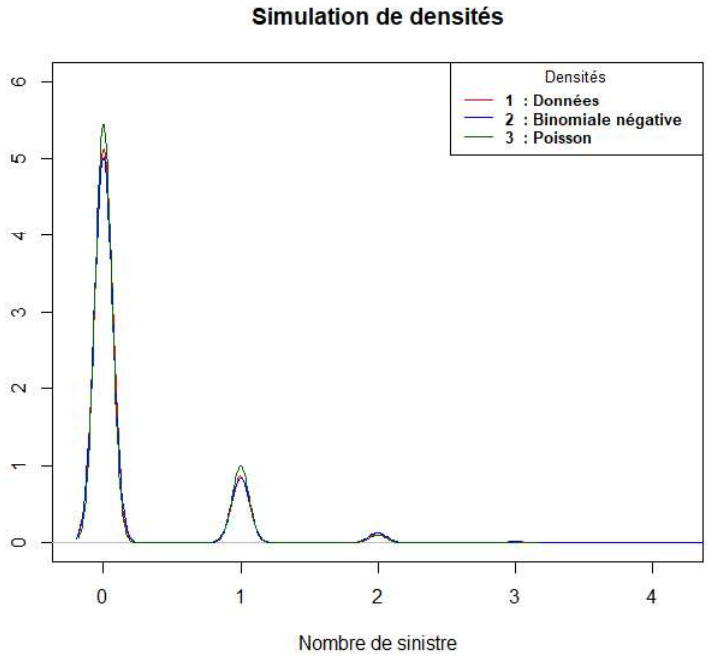

FIGURE 34 – Densités sur les nombre de sinistres

Il ressort que la distribution associée à la loi binomiale négative est plus à même à être utilisée pour le GLM du modèle de fréquence, puisque les deux courbes semblent davantage coïncider qu'avec la simulation résultant de la loi de Poisson. En effectuant le test de qualité de l'ajustement du χ^2 sur ces deux lois, il ressort que l'hypothèse d'une loi de poisson est rejetée, tandis que celle d'une loi binomiale négative est confirmée, avec une précision correcte (p-value supérieure à 0,9 et les résidus de Pearson égaux à 1,05). Ce test a été effectué à l'aide de la fonction *goodfit* du package vcd dans R.

Détermination du modèle retenu

Afin de sélectionner les variables explicatives adaptées au modèle, une méthode dite *stepwise* de type *backward* a été implémentée. Elle suit trois étapes :

- Dans un premier temps, il convient d'ajuster le modèle qui regroupe l'ensemble des variables mises à disposition. Il est question d'un modèle M_n à k variables.
- Par la suite, il convient d'estimer l'ensemble des modèles M_{n-k} avec $k \in [1, n-1]$ privés d'une variable par rapport au modèle M_{n-k+1}, et d'en retenir à chaque étape celui qui minimise la déviance. Cela donne lieu à l'obtention de $n-1$ modèles.
- Enfin, il convient de retenir le modèle optimal parmi ceux obtenus en se basant sur l'AIC[7] (Akaike Information Criterion) résultante. Ce critère est une mesure de la qualité de modèles statistiques. Il s'écrit $AIC = 2k - 2ln(L)$ avec k le nombre de paramètres du modèle à estimer et L le maximum de la fonction de vraisemblance du modèle.

La sortie du modèle est exposée dans le tableau suivant :

```
Call:
glm.nb(formula = Apprentissage$NB.Sin ~ ., data = Expl, weights = Apprentissage$Exp
osure,
    init.theta = 4.046468023, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9918  -1.1544  -0.8208  -0.4318   7.3823

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -2.820e+00  3.515e-01  -8.023 1.03e-15 ***
GroupeClasseE-G     -7.685e-03  9.441e-02  -0.081 0.935125
GroupeClasseH-M      1.495e-01  9.014e-02   1.658 0.097246 .
GroupeClasseN-O      3.307e-01  1.043e-01   3.169 0.001528 **
GroupeClasseP-Z      2.535e-01  1.272e-01   1.992 0.046345 *
Formule.kilométrage1 3.926e-01  7.420e-02   5.291 1.22e-07 ***
Formule.kilométrage2 5.904e-01  7.664e-02   7.704 1.32e-14 ***
Formule.kilométrage3 6.925e-01  8.755e-02   7.910 2.58e-15 ***
GroupeAge>= 76       5.440e-01  1.104e-01   4.928 8.32e-07 ***
GroupeAge26-30       8.280e-02  8.403e-02   0.985 0.324393
GroupeAge31-35      -3.407e-03  8.356e-02  -0.041 0.967474
GroupeAge36-40       3.870e-02  8.455e-02   0.458 0.647138
GroupeAge41-45       1.494e-01  8.238e-02   1.814 0.069690 .
GroupeAge46-50       1.605e-02  8.313e-02   0.193 0.846872
GroupeAge51-55       1.555e-01  8.129e-02   1.913 0.055708 .
GroupeAge56-60       9.269e-02  8.392e-02   1.105 0.269363
GroupeAge61-65       2.013e-01  8.515e-02   2.364 0.018061 *
GroupeAge66-70       3.332e-01  8.781e-02   3.794 0.000148 ***
GroupeAge71-75       2.842e-01  1.033e-01   2.752 0.005922 **
GroupeGroupe>= 36    1.698e+00  3.691e-01   4.601 4.21e-06 ***
GroupeGroupe27-30    1.162e+00  3.307e-01   3.515 0.000440 ***
GroupeGroupe31-32    1.134e+00  3.326e-01   3.410 0.000649 ***
GroupeGroupe32-35    1.290e+00  3.375e-01   3.822 0.000133 ***
GroupePuissance>= 17 -1.697e+00 4.523e-01  -3.752 0.000176 ***
GroupePuissance12-16 -7.887e-01 1.605e-01  -4.914 8.92e-07 ***
GroupePuissance5-6   -1.475e-01  3.783e-02  -3.898 9.70e-05 ***
GroupePuissance7-8   -3.700e-01  5.976e-02  -6.192 5.95e-10 ***
GroupePuissance9-11  -3.969e-01  8.653e-02  -4.587 4.49e-06 ***
GroupeDepG2          -2.252e-01  3.220e-02  -6.993 2.68e-12 ***
GroupeDepG3          -1.697e-02  3.605e-02  -0.471 0.637708
GroupeDepG4          -1.187e+00  2.961e-01  -4.008 6.13e-05 ***
GroupeMarqueDE        2.111e-02  5.261e-02   0.401 0.688191
GroupeMarqueEN-US     3.505e-03  6.206e-02   0.056 0.954964
GroupeMarqueEU       -1.786e-02  6.796e-02  -0.263 0.792635
GroupeMarqueFR       -9.224e-02  4.558e-02  -2.024 0.042993 *
GroupeAgeVeh>= 26    -1.924e+01  3.395e+03  -0.006 0.995478
GroupeAgeVeh11-15    -6.846e-01  5.017e-02 -13.645  < 2e-16 ***
GroupeAgeVeh16-20    -8.167e-01  1.267e-01  -6.446 1.15e-10 ***
GroupeAgeVeh21-25    -6.496e-01  2.645e-01  -2.456 0.014056 *
GroupeAgeVeh6-10     -2.263e-01  2.945e-02  -7.684 1.55e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

FIGURE 35 – Sortie R de l'exécution du GLM Fréquence sur l'ensemble des variables

Par l'application de cette méthode au modèle retenu pour la garantie Dommages Toute Auto, il en résulte donc que les variables suivantes sont retenues :

7. Un autre critère est parfois utilisé, il s'agit du Bayesian Information Criterion (BIC). Il est décrit par la formule suivante : $BIC = -2ln(L) + k \times ln(n)$ avec k le nombre de paramètres estimés et n le nombre d'observation. Toutefois, utiliser les deux critères AIC et BIC en même temps n'est pas intéressant, l'un permettant de retenir les variables pertinentes pour les prévisions et l'autre ayant pour objectif de sélectionner les variables statistiquement significatives.

1. Groupe
2. Marque
3. Classe
4. Puissance
5. Ancienneté du véhicule
6. Département
7. Âge
8. Formule kilométrique

Plusieurs précisions sont à donner pour bien comprendre cette sortie de logiciel quant au modèle étudié. En effet, R permet de bien visualiser les paramètres les plus importants lors de la réalisation d'un GLM, et il convient de détailler les différentes informations mises à dispositions par le *summary* du modèle. La première colonne renseigne le groupement de variable où les statistiques sont données, puis quatre autres colonnes donnent des informations statistiques sur chacun d'entre eux :

— Le coefficient estimé s'affecte au résultat total affectant la fréquence de sinistre, il est indiqué dans la première colonne.
— L'écart-type de la loi suivie par l'estimateur est indiqué dans la deuxième colonne.
— Le résultat de la statistique du test de Wald est indiqué dans la troisième colonne. Le processus menant à l'obtention de cette valeur est détaillé à la suite.
— Enfin, la p-value associée au test de Wald se trouve dans la dernière colonne.

Le choix des groupement de variable va en effet essentiellement se baser sur le résultat du test de Wald, et plus particulièrement à la p-value associée. Le test est utilisé lorsqu'il convient de s'intéresser à la nullité d'un seul paramètre. La statistique de Wald s'écrit ainsi :

$$W_j = \frac{\widehat{\beta}_j^2}{(\widehat{s.e.}(\widehat{\beta}_j))^2}$$

Avec $\widehat{s.e.}(\widehat{\beta}_j))^2$ est l'erreur $\widehat{\beta}_j$. Sous H_0, la statistique de Wald suit une loi du χ^2.

Par la suite, l'hypothèse H_0 se voit rejetée si $\sqrt{W_j} > z_{(1-\frac{a}{2})}$ avec $z_{(1-\frac{a}{2})}$ le quantile d'ordre $1-\frac{a}{2}$ d'une loi normale standard, généralement 0,05. Cela implique que la variable a un effet sur le modèle global.

Les lignes où la p-value n'est pas significative ne font pas figurer d'étoiles dans la dernière colonne. En effet, la modalité peut être considérée comme statistiquement significative si sa p-value est inférieure à 0,05. Elle peut l'être de façon moins marquée pour une p-value inférieur à 0,1 . En raison de ces premiers résultats, il convient de faire des regroupements afin d'arriver à un ensemble de modalité significative. Dans cette optique, les modalités des variables ancienneté du véhicule, marque du véhicule, groupe de département, âge de l'assuré et classe du véhicule subissent des retraitements :

— **Ancienneté du véhicule :** Les modalités 21-25 et >25 sont fusionnées en une unique modalité >20.
— **Âge de l'assuré :** Seules deux modalités sont retenues, selon si l'individu a plus de 75 ans ou non. Cela rejoint d'ailleurs les résultats des statistiques descriptives, qui renseignaient que les individus âgés semblaient causer une sur-sinistralité.
— **Groupe de département :** Comme exposé dans la section concernant la classification hiérarchique ascendante, un choix s'était posé, et le groupement avait été fait selon les sous-préfectures. Toutefois, le degré de détail semble trop important ici, et il a donc été décidé de se baser uniquement sur les six principaux départements pour séparer la variable en trois modalités, en se basant sur le dendogramme réalisé à l'aide des données sur ces six départements. Ce découpage est conservé pour l'ensemble des autres travaux de GLM.
— **Marque du véhicule :** Seules deux modalités sont retenues, selon si la marque du véhicule est française ou non.
— **Classe du véhicule :** Les modalités A-D et E-G sont fusionnées en une unique modalité A-G.

De cette manière, le modèle est résumé par un nouveau tableau :

```
Call:
glm.nb(formula = Apprentissage$NB.Sin ~ Formule.kilométrage +
    GroupeAge2 + GroupeGroupe + GroupePuissance + GroupeDep2 +
    GroupeClasse2 + GroupeMarque2 + GroupeAgeVeh2, data = Expl,
    weights = Apprentissage$Exposure, init.theta = 3.688405187,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9273  -1.1578  -0.8293  -0.4391   7.2251

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -2.78433    0.33777  -8.243  < 2e-16 ***
Formule.kilométrage1  0.35856    0.07411   4.838 1.31e-06 ***
Formule.kilométrage2  0.52325    0.07600   6.885 5.80e-12 ***
Formule.kilométrage3  0.63499    0.08692   7.306 2.76e-13 ***
GroupeAge2>= 76       0.40767    0.08411   4.847 1.25e-06 ***
GroupeGroupe>= 36     1.70135    0.36913   4.609 4.05e-06 ***
GroupeGroupe27-30     1.18124    0.33108   3.568 0.000360 ***
GroupeGroupe31-32     1.15333    0.33311   3.462 0.000536 ***
GroupeGroupe32-35     1.28649    0.33793   3.807 0.000141 ***
GroupePuissance>= 17 -1.62599    0.44984  -3.615 0.000301 ***
GroupePuissance12-16 -0.77812    0.15888  -4.897 9.71e-07 ***
GroupePuissance5-6   -0.13286    0.03755  -3.538 0.000404 ***
GroupePuissance7-8   -0.36219    0.05928  -6.110 9.97e-10 ***
GroupePuissance9-11  -0.36870    0.08573  -4.301 1.70e-05 ***
GroupeDep2G1         -0.06893    0.03650  -1.888 0.058991 .
GroupeDep2G2          0.09259    0.04099   2.259 0.023908 *
GroupeClasse2H-M      0.14176    0.04339   3.267 0.001087 **
GroupeClasse2N-O      0.34375    0.06794   5.059 4.21e-07 ***
GroupeClasse2P-Z      0.25863    0.09859   2.623 0.008708 **
GroupeMarque2FR      -0.07684    0.02805  -2.739 0.006155 **
GroupeAgeVeh2>20     -0.84643    0.26497  -3.194 0.001401 **
GroupeAgeVeh211-15   -0.68755    0.05014 -13.713  < 2e-16 ***
GroupeAgeVeh216-20   -0.82431    0.12671  -6.505 7.76e-11 ***
GroupeAgeVeh26-10    -0.23747    0.02938  -8.082 6.38e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

FIGURE 36 – Sortie R de l'exécution du GLM Fréquence après retraitement des variables

La variable créée avec les département pose toujours un problème de significativité, mais il est considéré pour la suite de l'étude que le dépassement est suffisamment faible pour être toléré. En effet, le découpage étant plus difficile à mettre en œuvre que pour les autres variables, s'obliger à ne conserver que des modalités dont la p-value est inférieure à 0,05 aurait mené à la suppression de la variable dans le modèle.

Un zonier aurait pu être mis en place, pour obtenir des résultats plus précis qu'avec la classification ascendante hiérarchique, toutefois, le groupe étant centralisé sur six départements, cette solution n'a pas été retenue en raison de la spécificité du portefeuille.

Validation du modèle retenu

Après s'être assuré de la significativité des variables du modèle, il convient en parallèle de s'intéresser à la qualité du modèle précédemment envisagé. le critère AIC a déjà été discuté puisqu'il a permis d'optimiser le modèle obtenu. L'objectif ici est principalement de valider les hypothèses qui régissent l'utilisation du GLM. Cela revient à s'assurer de trois caractéristiques concernant les résidus du modèle :

— Ils doivent être indépendants.
— Ils doivent suivre une loi normale de moyenne nulle et de variance résiduelle.
— Ils doivent être homogènes.

Concernant dans un premier temps **l'hypothèse d'indépendance des résidus**, l'attention est tout d'abord portée sur le graphique mettant en évidence l'éventuelle présence d'une auto-corrélation des résidus.

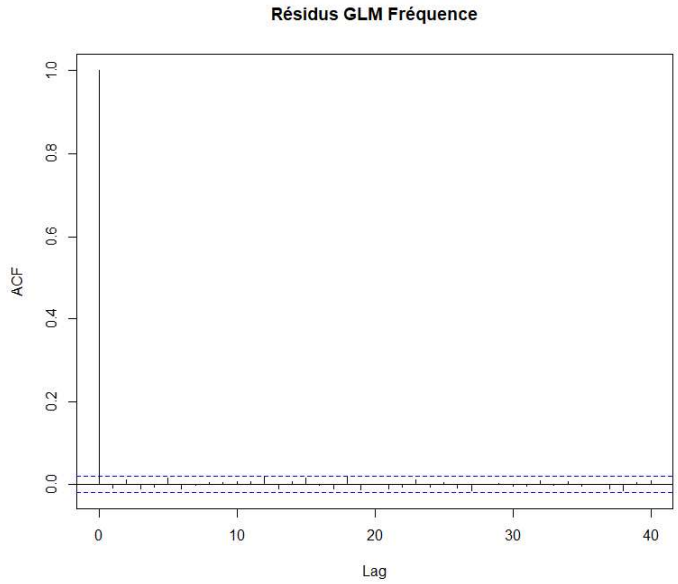

FIGURE 37 – Auto-corrélation des résidus du GLM de fréquence

Il ressort de ce graphique qu'il semblerait ne pas exister de corrélation significative entre les résidus selon le nombre de ligne qui les séparent dans le tableau de résidus (matérialisé par la variable "Lag"). En effet, les traits verticaux qui représentent les coefficients de corrélation entre les résidus de chaque points avec les points de la ligne suivante (pour Lag = 1) sont situés entre les intervalles de confiance du coefficient de corrélation égal à 0.

Concernant désormais **l'hypothèse de normalité des résidus**, l'évaluation de cette normalité est usuellement vérifiée en s'intéressant aux résidus réduits (c'est à dire aux résidus de même variance. La proportion de ces résidus dont la valeur absolue est supérieure à 2 est d'environ 3%, ce qui permet de confirmer la normalité des résidus, de moyenne nulle et de variance résiduelle. En effet, de manière générale, 95% des résidus réduits doivent être compris entre -2 et 2.

Enfin, pour **l'hypothèse d'homogénéité**, la forme de l'évolution des résidus mène à la conclusion comme quoi l'hypothèse n'est pas globalement respectée ; la droite tracée n'est pas horizontale et les résidus font apparaître plusieurs anomalies concernant leur répartition.

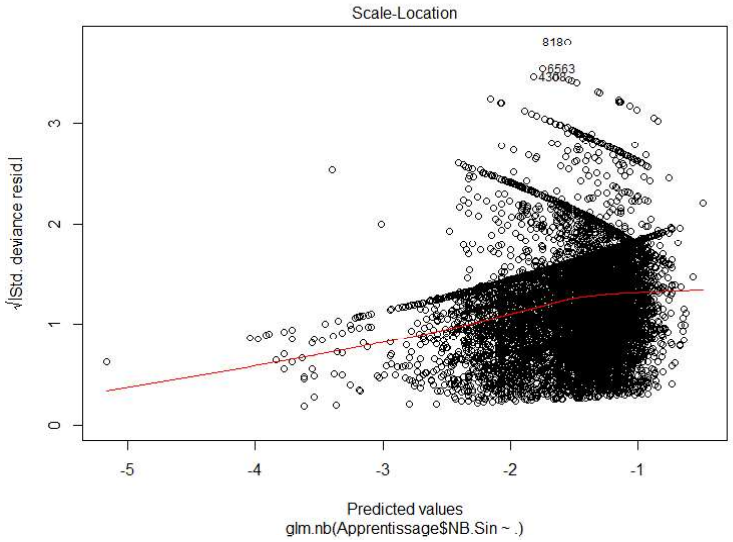

FIGURE 38 – Tracé des racines carrées des résidus de Student

En effet, ces derniers sont regroupés en une même zone du graphique, et des lignes de résidus se dessinent.

Usuellement, ces tests doivent faire voir qu'aucun pattern de résidus ne transparaisse. Ce n'est pas le cas ici. La mauvaise qualité de ces résidus est probablement liée aux difficultés rencontrés par le modèle quant à la bonne appréhension des données, ces dernières étant trop peu nombreuses pour obtenir un bon cadrage.

Dans un premier temps, le modèle est conservé en raison de la facilité de compréhension offerte par le GLM, et de son caractère pratique pour être utilisé par une entreprise d'assurance. Toutefois, selon les résultats obtenus, une alternative sera étudiée en raison du non respect de cette dernière hypothèse.

3.3.3 Modèle de coût moyen

Le mode de fonctionnement est semblable à celui du modèle de fréquence. Les mêmes variables sont sélectionnées pour lancer le modèle.

Choix de la distribution

Trois fonctions sont retenues pour se rapprocher de la densité du coût des sinistres. La loi binomiale négative a été décrite précédemment, est introduite également les loi gamma et log-normale.

Loi gamma : Soit $X \sim \Gamma(k, \theta)$. Il en résulte que $\mathbb{E}(X) = k\theta$ et $\mathbb{V}(X) = k\theta^2$.
D'où $m = k\theta \Leftrightarrow m = \frac{v}{\theta} \Leftrightarrow \theta = \frac{v}{m}$ et $k = \frac{m}{\theta} = \frac{m^2}{v}$

Loi log-normale : Soit $X \sim Log - \mathcal{N}(\mu, \sigma^2)$. Il en résulte que $\mathbb{E}(X) = e^{\mu + \frac{\sigma^2}{2}}$ et $\mathbb{V}(X) = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2}$.
D'où $v = (e^{\sigma^2 - 1})m^2 \Leftrightarrow e^{\theta^2} = 1 + \frac{v}{m^2} \Leftrightarrow \theta = ln(1 + \frac{v}{m^2})$ et $ln(m) = \mu + \frac{\sigma}{2} \Leftrightarrow \mu = ln(m) - \frac{1}{2}ln(1 + \frac{v}{m^2})$

Une représentation graphique des densités de ces fonctions est alors obtenue :

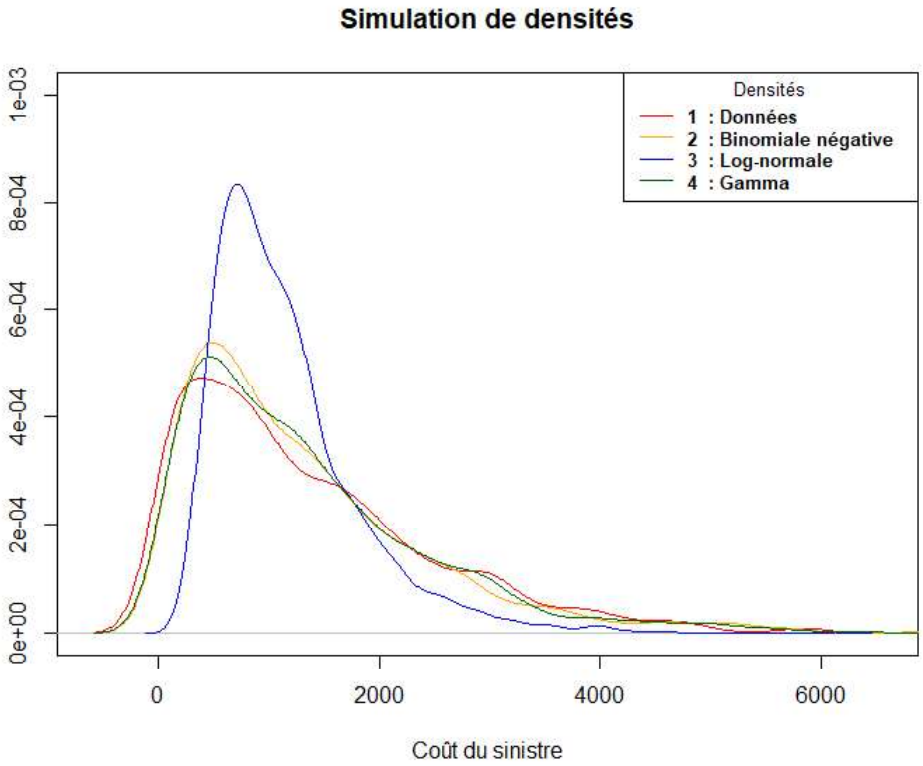

FIGURE 39 – Densités sur les coûts de sinistres

Il semble que l'approximation via une loi Gamma soit légèrement plus adaptée que par l'utilisation d'un loi binomiale négative, tandis que la loi log-normale ne parvient pas à bien correspondre à la courbe. Il est à noter cependant que les courbes ne sont pas aussi proches que ce qui était observé pour le modèle de fréquence. En effectuant par ailleurs le test de Kolmogorov-Smirnov sur ces trois lois, aucune ne ressort comme étant

significativement adaptée pour simuler la variable réponse. Cela laisse penser que l'application d'un GLM sera peu adaptée ici puisque les données ne sont pas assez lissées pour être approchées par une loi usuelle.

Tentative de détermination du modèle retenu

L'application du GLM est similaire à celle du modèle de fréquence. De la même façon, le but est d'obtenir un modèle où les variables sont toutes significatives. Le tableau en page suivante est ainsi obtenu après traitement des variables en conséquence.

```
Call:
glm(formula = Apprentissage$COUTOTDUSINGAM ~ ., family = Gamma(link = "log"),
    data = Expl2)

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-3.0830  -0.8878  -0.2335   0.3860   2.0389

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.15471    0.02702 264.771   <2e-16 ***
data>15     -0.31303    0.15302  -2.046   0.0409 *
data6-15     0.08962    0.03738   2.397   0.0166 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

FIGURE 40 – Sortie R de l'exécution du GLM Coût après retraitement des variables

La variable "Data" correspond à l'âge du véhicule, soit la seule variable jugée significative après application de la méthode destinée à minimiser le critère AIC. Le problème ici est le manque de données. Il semblerait que les lignes disponibles dans la base de données ne permettent pas au GLM de parvenir à estimer la réponse de façon satisfaisante. Cela confirme les soupçons précédemment évoqués. Seules trois valeurs de coûts peuvent être déduites des caractéristiques de l'assuré, et la conclusion en découlant est qu'il n'est pas envisageable d'utiliser un modèle si peu précis. L'autre cause de cette difficulté est aussi causé par la faible masse de données disponibles : en effet, le tracé de la densité du coût n'a pas pu être approché par une loi usuelle pour modéliser le coût d'un sinistre, et cela justifie la difficulté du modèle à simuler cette réponse.

Le problème étant visualisé, il convient désormais d'envisager d'autres solutions. En se basant sur les nombreux travaux déjà effectués dans le domaine, il ressort notamment que le modèle de coût peut parfois poser ce genre de problèmes. Si la méthode d'estimation par GLM a été présentée, il faut savoir que des alternatives existent. Trois pistes seront ici explorées, deux d'entre-elles ayant déjà été abordées dans le cadre de ce mémoire :

— La méthode CART
— La méthode Random Forest
— La méthode Gradient Boosting (GBM)

3.4 Nouvelle approche du modèle de coût

Le GLM ne semblant ici pas suffisamment efficace, l'objectif est ici de déterminer la méthode qui permettra de simuler au mieux le coût du sinistre. Les trois méthodes pré-citées vont ainsi être mises en place et appliquées à la base coût, toujours pour la garantie DTA. Une fois la méthode sélectionnée, la même sera conservée pour modéliser les deux autres garanties majeures du groupe. Deux critères de comparaison seront étudiés, puisque la complexité temporelle de chaque méthode est acceptable en raison de la petite taille des données à traiter. Il s'agit simplement des moyennes et écarts-types des différences entre la prédiction réalisée par le modèle, et de la variable coût disponible dans la base.

Dans cette optique, la base est environ séparée en proportion deux tiers / un tiers pour les bases d'apprentissage et de tests. Ces bases seront les mêmes pour toutes les méthodes utilisées, permettant d'avoir un point de comparaison équitable quelque soit la méthode.

3.4.1 Application des méthodes déjà abordées

GLM

Le GLM mis en place ci-dessus a été étudié afin de fournir une référence, malgré le fait qu'il se rapproche essentiellement d'un modèle où seul le coût moyen global aurait été pris en compte. Cela permettra de voir à quel point les autres méthodes s'écartent et améliorent cette approche ayant donné des résultats peu satisfaisants.

Avec la base apprentissage retenue, le GLM ne retient pas de variables significatives en raison de la privation du dernier tiers des données. Afin de s'intéresser à une référence pour le reste des méthodes, est retenue comme méthode une simple moyenne. La moyenne des coûts de la base d'apprentissage est donc utilisée pour être comparée aux valeurs de coûts de la base test. Cela donne une moyenne des écarts égale à environ 867 euros. Il est évident qu'au delà de cette qualité de prédiction, cette méthode présente l'inconvénient évident de ne considérer aucune caractéristique de l'assuré pour en déduire le coût du sinistre causé.

CART

Deux arbres seront retenus ici. En effet, dans un premier temps sera conservé l'arbre non élagué, afin de voir si cela améliore la précision. Le processus d'élagage est cette fois réglé en fonction du résultat du premier arbre, et non pas par une approche de minimisation de l'erreur. En raison de la taille des arbres, les graphiques résultants de l'application de cette méthode seront présentés en annexes.

En ressort une précision peu satisfaisante comparée à la méthode de la moyenne globale. L'arbre élagué est plus efficace que celui dont les ramifications n'ont pas été otées, et la moyenne est alors égale à 920.

Random Forest

Puisqu'il s'agit de la méthode ayant été retenue lors de la complétion des formules kilométriques manquantes, il convient de s'intéresser à l'efficacité de cette méthode par rapport aux autres. En réalisant une random forest de 10 000 arbres basée sur l'ensemble de la base d'apprentissage, la densité du vecteurs de réponse de la base test est comparée à la densité de la variable à prédire :

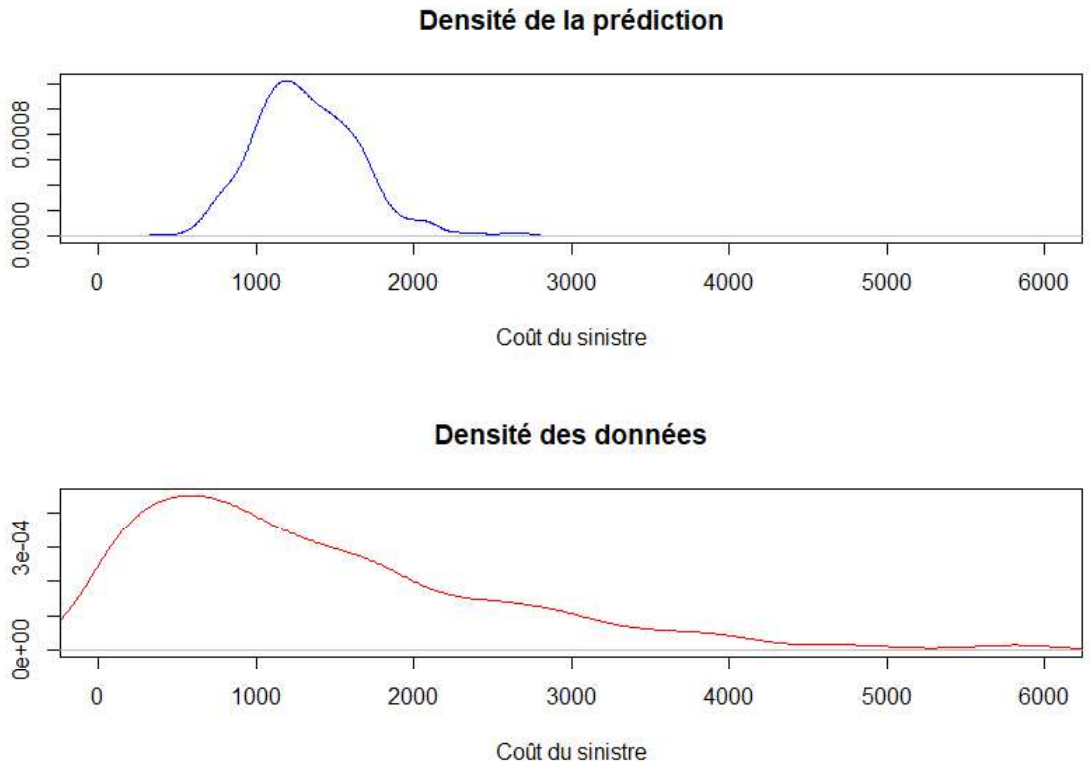

FIGURE 41 – Efficacité de la prédiction - Méthode RF

La méthode arrive à reproduire en partie la forme de la densité de la base test. Le problème réside dans les petites valeurs qui ne sont pas bien reproduites, causant un regroupement autour de la valeur moyenne.

Des tests ont également été effectués en modifiant le nombre d'arbres du modèle, afin de voir si cela impacte positivement la précision de la prédiction. Il en ressort que quelque soit le nombre d'arbre (entre 500 et 10 000 avec un pas de 500), la précision n'augmente pas. Elle reste constante et donne lieu à une moyenne de 892,

soit un résultat supérieur à la simple utilisation de la moyenne. La méthode n'ayant que peu de données sur lesquelles se baser, elle ne parvient pas à prédire efficacement les données de la base test.

3.4.2 Introduction du Gradient Boosting

La GBM (Gradient Boosting Method) est une méthode d'agrégation d'arbres, une variante de *bagging* tout comme la méthode Random Forest. En effet, là où le pouvoir prédictif d'un seul arbre est limité puisque les chemins ne permettent pas de retour en arrière, le gradient boosting n'est rien d'autre qu'un regroupement d'arbres qui vont être amenés à donner des résultats différents, sommés par la suite pour en obtenir la réponse. Cette méthode a été mise en place en 1999 par J.H. Friedman[9].

Plusieurs avantages découlent de l'utilisation de cette méthode. De la même manière qu'avec le modèle CART, le modèle est notamment capable de faire ressortir les variables les plus importantes. Cela fournit à l'assureur l'information des caractéristiques de l'assuré qui vont jouer sur la variable réponse (ici le coût). Un GBM permet également de prendre en compte les interactions entre les différentes variables, contrairement à un modèle GLM où les impacts de chaque variables sont vus de manière individuelle sur la réponse à fournir. Enfin, cela permet une prédiction correcte lorsque le lien entre les variables explicatives et la variable à expliquer n'est pas clair. De manière générale, ce n'est pas le cas pour l'évolution du coût en fonction des caractéristiques de l'assuré. Toutefois, ici, en raison notamment du manque de données, les liens sont moins évidents à réaliser, et de ce fait, cela justifie l'utilisation d'un tel procédé.

De la même manière que pour les Random Forests, l'algorithme de Gradient Boosting fait office de procédé où il est peu aisé d'y comprendre exactement ce qui s'y passe. L'algorithme faisant appel à deux mécanismes différent, soient la méthode de Boosting ainsi que l'algorithme du gradient, il est difficile d'obtenir une exposition claire du fonctionnement en tandem de ces deux mécanismes.

Algorithme du gradient

L'algorithme du gradient est un algorithme d'optimisation différentiable. Ici dans le cadre de l'utilisation pour une approche XGBoost, l'algorithme permet d'obtenir un point stationnaire pour un problème d'optimisation sans contrainte.

Soit maintenant un esace hilbertien $\mathbb{E}$ et f une fonction différentiable avec $x \in \mathbb{R} \rightarrow f(x) \in \mathbb{E}$. En prenant $f'(x)$ et $\nabla f(x)$ les dérivées et gradients de f en x, il résulte que $\forall d \in \mathbb{E}$, $f'(x).d =< \nabla f(x), d >$.

L'algorithme du gradient prend alors un point initial $x_0 \in \mathbb{E}$ et un seuil de tolérance $\epsilon \geq 0$. L'algorithme définit ainsi une suite de valeurs $x_1, x_2, ... \in \mathbb{E}$ qui prend fin lorsqu'une des conditions d'arrêt est activée. Cette condition peut porter sur un nombre maximal d'itérations (définit par l'utilisateur), ou sur un seuil de tolérance. Pour chaque x_k, x_{k+1} est définit selon le processus suivant :

- Simulation de$\nabla f(x_k)$
- Test d'arrêt en vérifiant que $\parallel \nabla f(x) \parallel \leq \epsilon$
- Détermination du pas α_k par une règle de recherche sur f en x le long de la direction $-\nabla f(x_k)$
- La nouvelle itération est alors définie de la manière suivante : $x_{k+1} = x_k - \nabla f(x_k)$

Avec α le taux d'apprentissage.

Méthode de boosting

L'objectif ici n'est pas de décrire mathématiquement le fonctionnement qui régit l'algorithme XGBoost. Les détails peuvent être retrouvés à partir des travaux de T. Chen et C. Guestrin [5], eux mêmes succinctement présentés dans le mémoire d'actuaire de P.Ottou [11] et ne seront pas exposés ici afin de faciliter la lecture. Cette partie a pour but de fournir une explication qualitative de la méthode de boosting, partie intégrante de l'algorithme XGBoost.

Le modèle, destiné à obtenir une prédiction à partir de plusieurs paramètres, va se baser sur une **fonction objectif**. Cette dernière permettra de donner une mesure de performance du modèle obtenu, en fonction de certains paramètres. Elle se décompose en une fonction de perte qui mesure la qualité de la prédiction du modèle, et sur un terme de régularisation qui plafonne la complexité du modèle, et qui évite que le modèle tombe

dans l'écueil du sur-apprentissage.

Comme présenté plus tôt, l'algorithme XGBoost est avant tout un modèle d'agrégation d'arbres. Comme remarqué avec la méthode CART qui se base sur les résultats d'un seul arbre, le pouvoir de prédiction atteint peut être insuffisant, c'est pourquoi la solution adoptée est de sommer la prédiction de plusieurs arbres.

Cela introduit ainsi le concept du **boosting**. La fonction objectif se basant sur d'autres fonctions comme paramètres, les méthodes d'optimisation usuelles ne peuvent ici pas être appliquée, et c'est l'algorithme du gradient qui sera alors utilisé pour minimiser (optimiser) ces fonctions. La philosophie de la méthode est alors de se baser sur l'erreur de prédiction. La méthode procédant par itération, chacune d'entre-elles bénéficiant des enseignements tirés de l'erreur commise précédemment, le but sera d'ajouter un arbre au modèle qui aura pour but de réduire l'erreur de prédiction.

Une **fonction objectif optimale** est alors déterminée mathématiquement, et permet de mesurer la qualité de l'arbre ajouté. En se basant sur cette dernière, il sera possible de déterminer si l'ajout de l'arbre en question est pertinent ou non. Il faut garder à l'esprit qu'il n'est pas possible de simuler tous les arbres possibles pour voir quel serait le meilleur à ajouter au modèle. L'intérêt est alors porté sur une **fonction de gain** qui fait état de quatre informations lors de la création de l'arbre supplémentaire en lui-même :

— Le score de la nouvelle feuille à gauche
— Le score de la nouvelle feuille à droite
— Le score du nœud d'origine

De cette manière, si le gain déterminé est plus faible qu'un terme de régularisation également compris dans la fonction (soit la quatrième information), la subdivision n'est pas effectuée.

Mise au point sur les différences avec la méthode RF

Afin d'éclairer à la compréhension du gradient boosting, il convient également d'expliquer la nature des différences entre les méthodes de gradient boosting et random forest. En effet, toutes deux reposant sur une simulation de plusieurs arbres de décision résultant de la méthode CART, et cela peut prêter à confusion. Quatre différences sont à noter :

— **Le processus de construction des arbres** : pour la GBM, il s'agit d'un processus séquentiel tandis que pour la méthode RF, chaque arbre est construit indépendamment en se basant sur un échantillon aléatoire différent de la base de données initiale.
— **Les domaine d'application** : la GBM s'applique à tout type de situation de façon assez efficace, tandis que la méthode RF peine dans certains cas (telles que les régression de poisson par exemple).
— **Le phénomène de sur-apprentissage** : notamment si la base de donnée est sujette à d'importantes variations, le GBM est plus exposé au risque de sur-apprentissage là où la méthode RF de par son fonctionnement y est plus résistante.
— **La durée d'exécution** : la méthode RF est généralement plus longue que la GBM pour un grand nombre d'arbres.

Globalement, faire appel à une méthode de random forest est plus conseillé pour des données où les corrélations sont visibles et fonctionne également si ces dernières sont bruitées. L'utilisation de l'algorithme XGBoost est conseillée si le but est de déterminer une classe dominante et d'identifier des cas mineurs sans leur accorder un poids trop important puisque leur modélisation n'est pas primordiale. Dans le cas présent, où il a été observé que l'application d'un GLM ne parvenait pas à retenir suffisamment de variables explicatives, et qui par le même raisonnement seraient corrélées au coût du sinistre, il semblerait que l'utilisation de l'algorithme XGBoost soit plus adapté.

Mise en application

Dans un premier temps, le graphique de l'influence relative des variables sur la détermination finale du coût totale est donné par le graphique suivant (après application d'un algorithme de gradient boosting à 10 000 arbres).

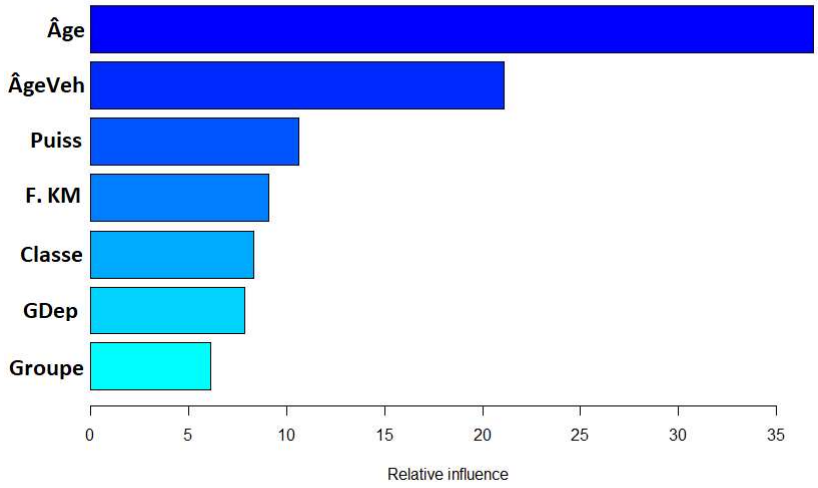

FIGURE 42 – Influence relative des variables sur le coût total

L'âge du véhicule déjà remarqué dans le GLM revient dans les deux variables les plus influentes au sein du modèle. L'âge est la plus influente, en sachant que ces deux variables ont été traitées de façon brutes, c'est à dire que ce n'est pas le regroupement qui a été retenu pour la base d'apprentissage (comme c'était le cas pour l'application des modèles précédents). Le reste des variables se partagent de façon assez homogène le reste de l'influence.

La moyenne obtenue avec une modélisation à 1 000 arbres est de 871, soit la plus faible comparée aux deux méthodes précédentes, mais toujours plus importante que celle obtenue par la simple utilisation de l'espérance. Il est intéressant de noter que dans ce cas, plus le nombre d'arbres est important, moins la méthode est précise. Cela fait écho au problème de sur-apprentissage sur la base prise aléatoirement. En effet, les données étant trop peu nombreuses, l'algorithme peine à trouver un réel lien entre les variables explicatives et la variable à expliquer sans s'adapter de façon excessive à la base d'apprentissage.

3.4.3 Comparaison des trois méthodes et choix final

Les résultats en se basant essentiellement sur les moyennes et écarts-types des erreurs, il convient de donner dans un premier temps le tableau récapitulatif de ces écarts.

Méthode	Espérance	CART	Random Forest	XGBoost
Moyenne	867	920	892	871
Ecart-type	1090	1254	1139	1095

FIGURE 43 – Écarts-types des écarts à la prédiction selon la méthode employée

Comme la méthode de l'espérance fait apparaître de bonnes moyennes et écarts-types mais ne laissera pas la possibilité d'avoir des coûts différents pour les assurés, il convient d'exclure la méthode et de trouver si possible un autre moyen d'évaluer l'efficacité des applications de data science. Afin de déterminer au mieux le choix final à adopter, le graphique des densités donnés par les prédictions des trois différentes méthodes est tracé, ainsi que les densités des bases apprentissage et test.

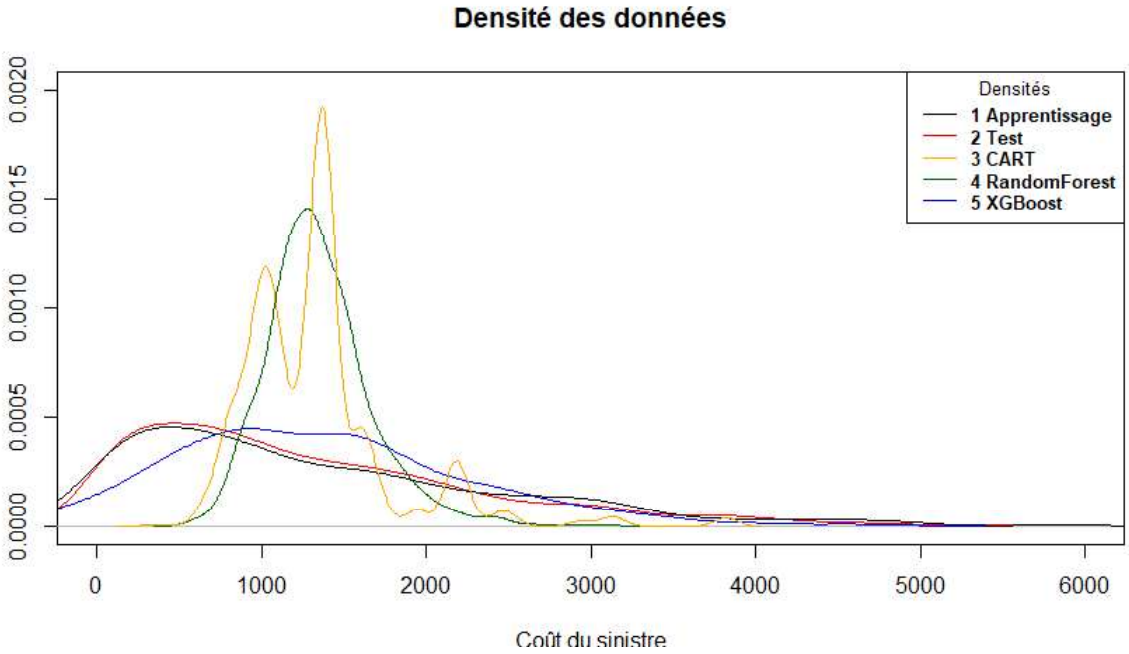

FIGURE 44 – Densités réponse obtenue par les méthodes autres que GLM

Malgré le manque de précision, l'application du modèle XGBoost permet d'obtenir une densité assez similaire à celle observée pour les données, que ce soit pour les données d'apprentissage ou de test. Les méthodes CART et random forest ont des difficultés à simuler un vecteur de prédiction ressemblant à ce qui devrait être obtenu. En se basant sur ces seules observations, conserver la méthode XGBoost semble être la meilleure solution.

Il faut également noter qu'en prenant un ratio 80% pour la base d'apprentissage et 20 % pour la base de test, les méthodes de gradient boosting et de random forest deviennent plus intéressantes, la moyenne et l'écart-type baissent pour atteindre respectivement 626 et 925 pour le gradient boosting, et 720 et 937 pour la random forest. La base retenue faisait voir des résultats moins intéressants en prenant simplement une moyenne arithmétique (moyenne des écarts de 890). Il est donc décidé de conserver la méthode de gradient boosting pour simuler les coûts associés au modèle de coût-fréquence mis en place. En effet, il semblerait qu'avec davantage de données la mise en place du modèle soit plus efficace, et de ce fait, il est justifiable de faire appel à une telle méthode pour simuler le coût d'un sinistre causé par l'assuré en question.

A noter que dans ce cas où l'algorithme dispose de plus de données mise à disposition, augmenter le nombre d'arbres a un effet positif. Le résultat donné ici résulte de l'exécution de la méthode de gradient boosting avec 50 000 arbres. En effet, jusqu'à un certain point, et en raison de l'augmentation de la masse de données mises à disposition pour l'entraînement de l'algorithme, le GBM aura plus de facilité à obtenir la tendance qui lui permettra de simuler les coûts par rapport aux profils des assurés. Toutefois, en augmentant de façon trop importante le nombre d'arbre, le modèle aura tendance à s'adapter de façon trop précise aux données d'entrée, et perdra ainsi de l'efficacité lorsqu'il s'agira de prédire une nouvelle série de profils.

3.5 Perspectives d'utilisation

La mise en place de la comparaison de plusieurs méthodes permet d'en retenir celle qui sera la plus adaptée par la suite pour amener à déterminer la prime finale. Ici donc, il ressort que l'utilisation du GLM sera adaptée pour le modèle de fréquence, et les résultats font d'ailleurs voir que cela s'applique aux trois principales garanties du groupe, mais qu'il faudra s'intéresser à la méthode de gradient boosting pour ce qui est du modèle de coût.

Ces travaux et réflexions, au delà du simple objectif qu'est l'estimation de la prime à réclamer à l'assuré, permettent de mettre en lumière le problème inhérent aux cédantes de petite taille. En effet, le manque de masse de données rend plus difficile la modélisation via les méthodes classiques ou plus modernes. Cela laisse assez peu d'informations aux cédantes pour en tirer les conclusions nécessaires à l'adaptation du tarif. Le résultat, comme le montrera l'ultime partie de ce mémoire, n'en est pas étonnant, puisqu'une nécessité de s'aligner sur

les prix du marché se fait ressentir, alors que l'étude de tarification fera ressortir la nécessité de demander plus aux assurés.

Plus que le résultat en terme de chiffre obtenu par ces travaux, il s'agira davantage de cibler les variables dont l'impact est visible pour la détermination du tarif. En effet, le groupe ayant basé son tarif sur une construction de correctifs, une simplification de la grille tarifaire pourrait être effectuée afin d'obtenir un résultat plus proche de la réalité. En prenant en compte ces variables dont l'intérêt ressort régulièrement dans les GLM pour la fréquence, et en tant que variables à influence dans les GBM, un tarif plus simple et plus juste pourra être obtenu. La priorité est donc de s'assurer de la bonne saisie de ces variables afin qu'elles puissent être correctement prises en compte par la suite pour l'élaboration du tarif.

4 Partie IV : Résultats et mise en relation avec la réassurance

4.1 Modélisation de l'ensemble des garanties

4.1.1 Les trois garanties principales

En s'appuyant sur les travaux de la partie précédente, et en généralisant la méthode aux deux autres garanties majeures du groupe (Responsabilité Civile et Bris De Glace), six modèles sont obtenus.

Trois **GLM de fréquence**, prenant tous comme base explicative des groupements de variables différents. La Formule Kilométrique, la Classe du véhicule, la Puissance du véhicule, le Groupe du véhicule et l'Âge du Véhicule en sont les principales. Les caractéristiques de l'assuré semblent moins jouer. En effet, l'âge est pris en compte mais est souvent sujet à de nombreux regroupements pour faire partie du modèle. L'emplacement géographique joue davantage sur la fréquence obtenue mais il semblerait que ce soit moins important que les caractéristiques du véhicule.

Le fait que la Formule Kilométrique soit une variable influente est compréhensible puisque le temps de conduite augmente l'exposition au risque. Le type de véhicule est également un facteur déterminant puisque certaines voitures sont plus à même de causer des sinistres que d'autres. Enfin, les caractéristiques de l'assuré jouent également mais sont moins déterminantes puisqu'il est souvent difficile de corréler le niveau de conduite à la personne en elle-même. Le fait que l'âge agisse est tout de même rassurant puisque les personnes âgés ou les jeunes conducteurs sont davantage porteuses de risques que les autres.

À noter que la situation géographique est un paramètre intéressant à étudier, les zoniers étant une étude classique de tarification, il est bienvenu de voir que le paramètre soit pris en compte même pour un portefeuille concentré au sein d'une même région. Il est évident que la variable n'aura pas la même portée que pour une société d'assurance automobile ayant des contrats dans la France entière, mais la situation géographique a tout de même une portée dans un cas plus limité.

Trois **GBM de coût** utilisant sensiblement les mêmes variables explicatives que les modèles de fréquence. L'âge est ici davantage utilisé, ainsi que l'âge du véhicule qui a également un rapport direct avec le coût. Les caractéristiques du véhicule sont également utilisées puisqu'elles seront directement liées au coût des réparations.

4.1.2 Discussion autour des résultats

Les autres garanties n'ont pas été traitées ici. En effet, les garanties principales ne bénéficiaient déjà pas d'une masse de données importantes, et ce problème est accentué pour les garanties mineures, qui présentent notamment une masse de sinistres très limitée.

Pour cette raison, le montant de prime affecté à chacune de ses garanties se base sur une méthode de prorata. La référence est prise avec la prime RC estimée, puisqu'il s'agit de la garantie souscrite par l'ensemble des assurés. Le montant total de la prime réelle de chaque garantie non étudiée est par la suite rapporté au total de celui affecté à la RC, puis le prorata est affecté à la prime RC estimée de chaque assuré.

Les résultats présentés ensuite prennent alors en compte l'exposition à multiplier à la fréquence, et donc à la prime totale puisque le CRM est ici utilisé comme variable d'ajustement. Le portefeuille étant composé en quasi-totalité d'individus présentant un CRM égal à 0,5, la prime résultant est indexée sur cette valeur de CRM comme référence. De ce fait, les individus ayant un CRM supérieur à ce nombre se voient affecté un coefficient augmentant leur volume de prime demandé.

Les garanties réelles sont considérées après la prise en compte du CRM (comme c'est le cas dans les analyses multivariées). La raison est que l'outil du groupe fonctionne différemment de celui exposé dans ce mémoire et avait pour objet d'utiliser le CRM en valeur multiplicative. En pratique, le CRM sert davantage à adapter le tarif selon le profil de risque de l'assuré, et il est plus simple pour un assureur de le considérer a posteriori pour ajuster la prime selon le profil.

En effet, bien qu'il n'ait pas été pris en compte dans l'étude de tarification (il a volontairement été exclu du GLM mais aurait en théorie pu être considéré comme poids), les primes ont été calculées de manière à représenter au

mieux la réalité. De ce fait, il ne peut pas être considéré de la même façon.

Récapitulatif Moyenne	RC	DTA	BDG	Sous-Total
Réel	86,86	113,35	28,67	**228,88**
Estimé	63,38	185,20	73,39	**321,97**
Estimé YC CRM	67,33	196,28	77,99	**341,59**

Récapitulatif Moyenne	Accessoires	Cat Nat	Cat Tech	GC	Incendie	PM	SM	Tempête	Vol	Total Général
Réel	3,79	2,78	1,99	26,33	5,49	42,92	5,45	3,53	25,40	**346,56**
Estimé	2,39	1,55	1,25	16,69	3,44	29,85	3,39	2,18	15,93	**398,65**
Estimé YC CRM	2,53	1,64	1,32	17,67	3,64	31,48	3,60	2,32	16,87	**422,67**

FIGURE 45 – Tableaux récapitulatifs des résultats

La ligne renseignant le "réel" donne les montants de prime renseignés par l'outil sur les données de la base retraitée utilisée pour calibrer les modèles de coûts et de fréquence. La ligne renseignant l'"estimé" correspond aux résultats donnés par la nouvelle méthode de tarification.

L'estimation des primes des garanties DTA et BDG sont bien supérieures à celles originellement demandées par le groupe, normalement compensées par les polices des autres garanties. L'écart au total est par la suite atténué puisque la prime estimée pour la garantie RC est inférieure à la prime réelle. Il faut noter qu'un montant de 350 est imputé aux aménagements proposé par le groupe, réduisant le total de primes observé sur le réel. Puisque les garanties non estimées par les méthodes exposées dans la partie précédente sont indexées sur la RC, celles-ci sont également plus faibles que les réelles.

La ligne la plus intéressante correspond aux primes où le CRM est inclus qui correspond à la prime à demander à l'assuré. C'est cette prime qui sera conservée pour la suite de la partie.

Le modèle de fréquence résultant du GLM sur la garantie RC fait apparaître que la fréquence moyenne, une fois l'exposition prise en compte, est égale à 0,047 contre 0,064 dans la base d'origine. Cette imprécision provient directement du fait que peu de sinistres sont observées pour la garantie RC, ce qui se traduit par ailleurs par un GLM moins solide que pour la garantie DTA (moins de variables significatives, et p-value plus élevée). Par ailleurs, cela rejoint le problème observé au niveau des résidus des GLM (qui était également observé pour les garanties RC et BDG), et montre que le GLM n'est pas adapté à la situation. De ce fait, une méthode de random forest est utilisée pour simuler la fréquence (plus adaptée car l'objectif est ici se protéger au mieux du sur-apprentissage au vu du nombre de données plus important que pour la base coût). Cela donne cette fois une police moyenne approchant les 92 euros, pour un SP baissant à 97% (soit amélioré par rapport au réel).

Cette méthodologie est également appliquée aux deux autres garanties DTA et BDG. Bien que le résultat soit plus lisible et plus facilement applicable avec un GLM, dans l'optique d'une utilisation de l'entreprise du modèle, cette nouvelle méthode est plus adaptée, et mathématiquement plus juste en raison des résidus peu engageants observés. L'objectif de ce mémoire est de fournir des résultats satisfaisants par rapport à la problématique d'origine, qui est d'améliorer le ratio global, l'utilisation d'une méthode de Data Science permet ainsi d'améliorer la qualité de notre prédiction, et les résultats seront conservés pour la suite de l'étude.

Récapitulatif Moyenne	RC	DTA	BDG	Sous-Total
Réel	86,86	113,35	28,67	**228,88**
Estimé	92,34	120,69	49,94	**262,96**
Estimé YC CRM	98,93	129,06	53,52	**281,51**

Récapitulatif Moyenne	Accessoires	Cat Nat	Cat Tech	GC	Incendie	PM	SM	Tempête	Vol	Total Général
Réel	3,79	2,78	1,99	26,33	5,49	42,92	5,45	3,53	25,40	**346,56**
Estimé	4,10	2,66	2,13	28,71	5,90	49,47	5,84	3,83	27,30	**392,89**
Estimé YC CRM	4,39	2,85	2,29	30,75	6,31	52,59	6,25	4,11	29,23	**420,28**

FIGURE 46 – Tableaux récapitulatifs des résultats après changement de méthodologie

Il convient de constater qu'en adaptant la méthode de prédiction des trois garanties, la différence devient moins importante pour les trois garanties, hormis la garantie BDG qui fait apparaître une importante augmentation. L'intérêt est par la suite porté aux nouveaux ratios S/P obtenus avec la prime estimée par rapport à la prime réelle. Les sinistres pris en compte correspondent à ceux de la base après retranchement des valeurs extrêmes puisque la tarification était adaptée pour ces valeurs.

	RC	DTA	BDG	TOTAL
Montant Sinistre	3 601 k€	3 074 k€	1 560 k€	8 946 k€
Montant Police réel	3 820 k€	3 241 k€	1 202 k€	12 669 k€
Montant Police estimé	4 061 k€	3 691 k€	2 244 k€	15 164 k€
S/P Réel	94%	95%	130%	71%
S/P Estimé	89%	83%	70%	59%

FIGURE 47 – Tableaux récapitulatifs des ratios S/P

Seules les garanties principales sont affichées mais le total prend en compte l'ensemble des garanties. Si les résultats apparaissent relativement satisfaisant pour les garanties DTA et BDG où l'augmentation de la part de la prime allouée a amélioré les choses, de façon très notable pour la garantie BDG, il convient de porter un jugement quant aux primes allouées aux garanties secondaires, dont les résultats sont présentés dans le tableau suivant :

	Accessoires	Cat Nat	Cat Tech	GC	Incendie	PM	SM	Tempête	Vol	TOTAL
Montant Sinistre	15 k€	1 k€	0 k€	445 k€	61 k€	34 k€	0 k€	0 k€	154 k€	8 946 k€
Montant Police réel	159 k€	117 k€	84 k€	1 154 k€	230 k€	1 223 k€	240 k€	134 k€	1 065 k€	12 669 k€
Montant Police estimé	185 k€	120 k€	96 k€	1 347 k€	265 k€	1 499 k€	275 k€	156 k€	1 226 k€	15 164 k€
S/P Réel	10%	1%	0%	39%	27%	3%	0%	0%	14%	71%
S/P Estimé	8%	1%	0%	33%	23%	2%	0%	0%	13%	59%

FIGURE 48 – Tableaux récapitulatifs des ratios S/P - Garanties mineures

Ces garanties sont sur-tarifées. En effet, la tarification appliquée se base sur un prorata tiré de l'outil originel, qui prenait probablement des montant déterminés arbitrairement et donc peu adaptés aux sinistres réellement observés. Dans une optique commerciale, il peut être judicieux de réduire la part de prime allouée à ces garanties pour limiter l'augmentation induite par le changement de méthode sur les trois garanties principales. En effet, en gardant ces résultats, le SP global de 59% obtenu est amplement suffisant pour limiter le profit sur ces branches.

De ces premiers résultats sont tirés plusieurs conclusions. Bien évidemment, il faut bien comprendre que dans un tel contexte, la faible quantité de données ainsi que leur qualité approximative entrave grandement le bon traitement de ces dernières. Cela donne des résultats porteurs d'information certes, mais au prix de l'utilisation de méthodes de data science dont l'application dans le monde réel est difficile. Pour un groupe de moyenne taille, le recours au GLM est plus adapté puisque cela permet une mise en place plus aisée du tarif à demander à l'assuré. Toutefois, la situation ne permet par l'obtention de résultats envisageables et nous poussent à utiliser d'autres méthodes afin d'obtenir des résultats utilisables pour les derniers travaux destinés à répondre à la problématique posée en début de mémoire.

Ces problèmes communs aux petits organismes sont usuellement résolus par l'acquisition de base de données plus importantes et de ce fait plus aisées à traiter pour en obtenir une tarification cohérente et permettant de se rapprocher de la réalité. Cela permet un meilleur fonctionnement des modèles de GLM, et simplifie le fonctionnement des modèles par rapport aux variables qu'ils n'arrivent pas à prendre en compte en raison du manque de données. Les méthodes présentées restent toutefois pertinentes dans le cadre d'une étude de tarification, et cela permet de mettre en lumière les problèmes auxquels sont confrontées les petites et moyennes sociétés d'assurance automobile quant à la mise en place d'une tarification juste et précise.

4.2 Application de la réassurance

4.2.1 Résultats bruts

Afin de s'intéresser à l'impact sur le réassureur du changement des primes, il convient de préciser le mode d'action qui sera retenu afin de limiter au plus les écarts de résultats observés et expliqués dans la partie précédente. Les résultats de S/P seront conservés, et la part allouée aux garanties secondaires est réduite. En effet, la différence entre les primes et les sinistres de ces garanties fait apparaître une marge de plus de 4 millions d'euros. Ainsi, il est proposé de réduire d'un quart la part de prime pour ces garanties.

Idéalement, il conviendrait d'étudier plus en amont ces garanties, mais le faible nombre de sinistre, et l'éventuelle gravité de certains (notamment pour les garanties Catastrophes naturelle et technologiques) rend une approche par une modélisation peu pertinente. En s'intéressant à l'historique du groupe, ces garanties sont usuellement

très rentables, et diminuer le volume de prime affecté à ces dernières semble donc être une option justifiable. Les ratios S/P globaux bruts de réassurance et de frais en ôtant ces garanties deviennent respectivement 64% et 72% en simulé et en réel.

Par ce retranchement, la prime totale est légèrement supérieure à celle d'origine, mais tous les ratio S/P semblent acceptables, ce qui permettra de faciliter une éventuelle modification de tarif en cas d'application des travaux. L'objectif ici est de pousser la réflexion un peu plus loin en observant dans un premier temps l'impact sur le réassureur qu'à cette nouvelle prime, et ensuite de s'intéresser à un processus d'optimisation permettant de modifier la répartition de la prime totale pour en visualiser l'impact sur les résultats de l'assureur et du réassureur (qui bien évidemment se partagent le résultat global, ce qui implique que la somme des deux résultats sera toujours égale).

			RC	DTA	BDG	Total
Vision assureur (Net)	Simulé	Primes	2 371 k€	2 446 k€	1 487 k€	9 348 k€
		Sinistres	2 881 k€	2 459 k€	1 248 k€	7 172 k€
		SP	121%	101%	84%	77%
	Réel	Primes	2 226 k€	2 133 k€	791 k€	8 330 k€
		Sinistres	2 881 k€	2 459 k€	1 248 k€	7 172 k€
		SP	129%	115%	158%	86%

			RC	DTA	BDG	Total
Vision Réassureur (Cédé)	Simulé	Primes	1 690 k€	1 245 k€	757 k€	4 642 k€
		Sinistres	1 620 k€	1 383 k€	702 k€	4 010 k€
		SP	96%	111%	93%	86%
	Réel	Primes	1 594 k€	1 109 k€	411 k€	4 139 k€
		Sinistres	1 620 k€	1 383 k€	702 k€	4 010 k€
		SP	102%	125%	171%	97%

FIGURE 49 – Application du traité de réassurance - Garanties majeures et total

Il convient de préciser à nouveau que ces montants correspondent à l'ensemble des polices et sinistres de la garantie automobile du groupe sur les quatre ans étudiés. Les frais ont été évalués à 25% des prestations à payer, attribués à chaque garantie sans réellement se soucier des spécificités de chacune concernant les frais. Aucun frais n'a été affecté concernant l'acquisition des primes. Le ratio S/P brut de réassurance en prenant les frais en compte passe de 72% à 90% en réel, et de 64% à 80% en simulé. En ajoutant les frais non pris en compte dans l'étude, il apparaît bien vite que le contrat ne donne pas lieu à une rentabilité importante pour le groupe.

Afin de bien comprendre les différences entre les ratios résultants des primes et des sinistres cédés, dans le cas d'un quote-part, et les ratios bruts de réassurance, le fonctionnement du contrat est ainsi décrit :

- Après que la quote-part ait été appliquée, une **commission de réassurance** est appliquée au montant de primes cédées par l'assureur, et est reversée à ce dernier. Elle s'élève à 6% pour les garanties type RC (Responsabilité Civile et Garantie du Conducteur) et 24% pour les garanties type DTA (l'ensemble des autres garanties, à l'exception de Panne Moteur et Secours Mutaliste qui ne sont pas réassurées). Les commissions ont pour but de représenter la participation du réassureur aux frais de l'assureur.
- Un **compte de résultat** est ensuite tenu, il est calculé à partir de l'ensemble des primes cédées au réassureur, auxquelles sont retranchées les sinistres cédés et les commissions correspondantes.
- Si ce compte est positif, 10% du compte de résultat est reversé à l'assureur en tant que **participation aux bénéfices**. Ce montant de participation de bénéfices est ensuite réparti au prorata des primes sur chaque garantie concernée.

Les résultats sont intéressants, puisqu'ils font apparaître un S/P plutôt positif au global, principalement alimenté par les garanties secondaires qui n'ont pas été représentées dans le tableau récapitulatif de l'effet de la réassurance. La différence entre les primes basées sur la prédiction et celles observées réellement est déjà assez parlante pour l'assureur, notamment si l'attention est porté sur la garantie Bris De Glace qui apparaît comme très nettement déficitaire.

La problématique de la réassurance apparaît également assez clairement ici. Les trois garanties principales sont déficitaires pour le réassureur, et dégradent fortement son ratio S/P, bien qu'avec les primes simulées l'effet soit moins marqué. Les deux garanties non concernées par le contrat de réassurance sont parmi les plus rentables

comme cela a été montré dans la sous-partie précédente, et ne sont pas captées par le réassureur. En améliorant, et lissant par la même occasion le résultat, le réassureur a une sinistralité moins importante à endiguer. La répartition des primes a donc son importance, et en faisant en sorte de donner un S/P cible pour l'ensemble des garanties, et en répartissant le surplus pour combler la sinistralité des garanties déficitaires, l'assureur perdra une part de bénéfice au profit du réassureur.

Cette modification de la répartition impactera très faiblement le consommateur. En effet, à l'exception de la garantie DTA, l'ensemble des garanties sont souscrites par l'entièreté du portefeuille. La prime ne sera pas sujette à d'importantes modifications, l'affectation aux différentes garanties, elle, sera modifiée. Globalement, le surplus du volume de prime causé par la nouvelle tarification simulée sera ainsi réparti entre l'assureur, qui ne verra pas son résultat être significativement modifié, et le réassureur, qui retrouvera un intérêt à couvrir le portefeuille.

4.2.2 Mise en application d'une nouvelle répartition

Le but est ici d'obtenir le ratio S/P limite minimal tel que l'ensemble des S/P de chaque garantie soit inférieur à ce seuil. Certaines contraintes sont mises en place, notamment l'affectation d'un seuil minimal de prime à chaque garantie même dans le cas où celle-ci est très peu sinistrée. Comme précédemment, les garanties Catastrophes naturelles et technologiques sont exclues, en raison du caractère ponctuel de tels sinistres. Il est proposé de ne pas toucher au volume de prime alloué à ces garanties puisqu'il est difficile de se prononcer sur une sinistralité qui peut épisodiquement être très élevée. Les autres garanties font voir des résultats assez largement bénéficiaires sur les quatre années, et sont donc étudiées en conséquence.

Le ratio limite est fixé à 96% pour les montants réels, et à 85% pour ceux résultant de la simulation. Il s'agit du plus petit S/P où le montant total de la prime peut-être réparti de manière à ce que aucune garantie ne donne lieu à un S/P supérieur à ce seuil. Un minimum de 1% de la prime totale doit être alloué à chaque garantie. Puisque les primes simulées impliquent un volume plus important, le ratio limite peut-être plus bas que pour les montants réellement observés.

			RC	DTA	BDG	Total
Vision assureur (Solveur)	Simulé	Primes	3 389 k€	2 893 k€	1 469 k€	8 899 k€
		Sinistres	2 881 k€	2 459 k€	1 248 k€	7 172 k€
		SP	85%	85%	85%	81%
	Réel	Primes	3 001 k€	2 562 k€	1 300 k€	7 889 k€
		Sinistres	2 881 k€	2 459 k€	1 248 k€	7 172 k€
		SP	96%	96%	96%	91%

			RC	DTA	BDG	Total
Vision Réassureur (Solveur)	Simulé	Primes	2 361 k€	1 481 k€	750 k€	5 091 k€
		Sinistres	1 620 k€	1 383 k€	702 k€	4 010 k€
		SP	69%	93%	94%	79%
	Réel	Primes	2 119 k€	1 329 k€	675 k€	4 581 k€
		Sinistres	1 620 k€	1 383 k€	702 k€	4 010 k€
		SP	76%	104%	104%	88%

FIGURE 50 – Application du traité de réassurance - Sortie du solveur

Le ratio de rentabilité global du réassureur dépend de celui obtenu à la base. La répartition modifiée permet de transférer une part du résultat de l'assureur au réassureur. Dans le cas du résultat de la prédiction effectuée lors de ce mémoire, le ratio passe légèrement en dessous du seuil des 80%, et en dessous des 90% pour les primes réelles.

Comme attendu, les garanties les plus impactantes pour le réassureur sont les trois principales représentées dans le tableau. La différence de ratio entre la garantie RC et les garanties BDG et DTA provient du fait que le pourcentage de commission ne soit pas le même. La commission pour les garanties type RC étant plus faibles, le réassureur conserve davantage de résultat. Finalement, dans les deux cas, optimiser la répartition pour lisser les ratios S/P implique un transfert d'une part du résultat de l'assureur au réassureur.

Ces chiffres permettent donc de constater qu'une troisième part du problème, en sus de la conservation de mauvais profils au sein du portefeuille et d'une tarification dépassée, provient de la répartition des primes au sein de chaque garantie. Le résultat du réassureur est certes lié au résultat brut, mais peut s'améliorer en acceptant de

mieux répartir les primes au sein de chaque contrat. Les résultats observés montrent de façon assez claire que les garanties mineures souffrent de tarification approximative, et devraient être retraitées afin de ne pas prendre le risque de désintéresser totalement le réassureur du contrat automobile.

Au delà de ces conclusions, le ratio de 85% obtenu pour chaque garantie après optimisation est plutôt satisfaisant, puisque l'augmentation liée à l'application de la réassurance n'a pas un impact trop important. La réassurance en quote-part fait sens dans la mesure où les montants renseignés agrègent quatre années de prime et de sinistralité. L'observation sur 2016 et 2017 faisait voir un ratio très bon, tandis que 2015 et 2018 donnaient lieu à un S/P plus proche des 100%. Il faut bien préciser que certains frais n'ont pas été pris en compte, et qu'il faut se montrer critique envers le ratio final obtenu. La réassurance a pour but originel de faciliter les années plus sinistrées, et mérite d'être conservée dans la situation actuelle du groupe, d'autant plus que son coût reste suffisamment faible pour être accepté par le groupe.

4.3 Prolongements dans le contexte règlementaire actuel

4.3.1 L'ORSA au sein de Solvabilité 2

La Directive Européenne Solvabilité II, mise en œuvre depuis janvier 2016, s'applique à l'ensemble des États membres de l'Union Européenne. L'objectif initial est de définir un cadre règlementaire destiné à adapter les Fonds Propres des organismes concernés aux risques auxquels ils sont exposés.

Cette réforme a été mise en place en palliant aux limites de la Directive Solvabilité I, précédemment appliquée. Plusieurs obligations découle de cette Directive, dont certaines étaient déjà présentes au sein de Solvabilité I :

- La constitution de provision prudentes, de manière à s'assurer que les primes aient pour but de faire face aux coûts de prestation importants.
- La valorisation adaptée des actifs, à vision actuelle plutôt qu'historique.
- Par l'estimation de l'ensemble des risques auxquels est soumise l'entité, la nécessité de disposer d'un niveau de Fonds Propres nécessaires dans l'hypothèse d'un risque bicentennaire.
- La mise en place d'un système de Gouvernance organisé
- La mise en place de dispositifs obligatoires concernant le rendu des informations envers le public et le superviseur

Pour répondre à ces exigences, la Directive s'articule autour de trois piliers définis de manière assez précise dans les textes règlementaires. En annexe sont données quelques précisions sur le contenu des trois piliers et la définition du **SCR** (Solvency Capital Requirement).

La mise en place d'une nouvelle tarification pour le contrat automobile pourrait être considéré comme un stress-test au sein du processus **ORSA**. En effet, l'ORSA (Own Risk & Solvency Assessments) se décompose en plusieurs partie.

Dans un premier temps, il est question de la critique de la formule standard dans le cadre de l'organisme assurantiel étudié. Il s'agit par exemple de réévaluer le risque immobilier au sein du SCR Marché, ou de calculer autrement les risques de primes et réserves par l'application d'autres méthodes statistiques adaptées aux données de l'entreprise. Par l'application de plusieurs retraitement, le SCR vision solvabilité 2 s'adapte pour devenir le SCR Vision ORSA.

Par la suite, une estimation des risques propres à l'entreprise qui ne sont pas pris en compte dans la formule standard est ajoutée au capital déterminé précédemment. Cela peut par exemple concerner le risque homme clé, la prise en compte de nouvelles règlementations, ou le risque de cyber-attaque, plus actuel. Il est également question du risque RGPD (Règlement Général sur la Protection des Données) actuellement, qui concerne d'autant plus les contrats d'assurance automobile puisque les caractéristiques des assurés y sont renseignées. Résulte de cet ajout l'obtention du Besoin Global de Solvabilité, ou BGS, qui rapporté aux fonds propres donne un nouveau ratio vision ORSA.

L'ORSA a pour but par la suite de montrer que les exigences de règlementation seront assurées par la suite, par le biais d'une projection du ratio sur cinq ans suivant les hypothèses d'un Business Plan. Enfin, des **scénarios de stress** sont mis en place pour tester la solidité du ratio de la société suivant divers évènements tels que la hausse ou la baisse du chiffre d'affaire, la dégradation de la sinistralité ou même la mise en place d'une nouvelle

règlementation, ou d'un changement de tarification.

C'est cet aspect sur lequel l'accent sera porté pour cette dernière partie. Deux stress test seront étudiés afin de conclure sur les tenants et aboutissants de l'ensemble de l'étude concernant la tarification du portefeuille. L'un concernera la mise en place de la nouvelle tarification, tandis que l'autre aura pour sujet la conservation de la tarification actuelle, mais la perte de la réassurance à horizon 2021.

4.3.2 Mise en place des stress-test

Le groupe ayant une activité assez diversifiée, le poids global de l'assurance automobile n'explique pas toutes les variations observées. Il est proposé de s'intéresser à l'évolution globale de deux paramètres : le ratio SCR, dont la composition est exposée en annexe, ainsi que le résultat du groupe, afin dans un premier temps de constater l'impact de l'abandon de la réassurance du secteur automobile.

Réassurance

L'hypothèse traite le désistement du réassureur en 2020 concernant les contrats automobiles du groupe. Cela signifie que la réassurance est conservée sur l'ensemble des autres activités.

Pour ce qui est du résultat en lui même, la conséquence est que le groupe doit supporter l'ensemble des coûts des garanties déficitaires sans passage par quote-part, cela en induit une dégradation. La dégradation se reporte par la suite sur les fonds propres du groupe, mais le véritable impact sur le SCR concerne les sous-modules catastrophe et prime du module non-vie.

Pour précision, le **SCR Catastrophe** estime le montant des dommages applicables aux garanties couvertes par l'assureur pour le risque d'accident de masse et le risque de concentration d'accidents. Ce montant fait généralement l'objet d'une atténuation (*mitigation*) importante par le biais du réassureur.

Le **SCR Prime** consiste en l'estimation du risque de sous-tarification. Ce montant dépend des cotisations sur lesquelles la société d'assurance s'est engagée.

La réassurance ne concerne désormais plus que les garanties hors auto du groupe par le biais de l'hypothèse, cela limite l'atténuation normalement appliquée pour obtenir le SCR catastrophe, et augmente le volume de prime nettes touchées par l'assureur. Cet effet est contrebalancé par la baisse du SCR contrepartie, en raison de la réduction de la part cédée au réassureur.

En effet, le **SCR Contrepartie** cherche à évaluer le risque de faillite des débiteurs de l'entité, qui causerait donc une perte pour cette dernière. Puisque le volume de provisions cédées baisse, le risque induit par la faillite de l'assureur si avéré donnerait lieu à une perte moins importante.

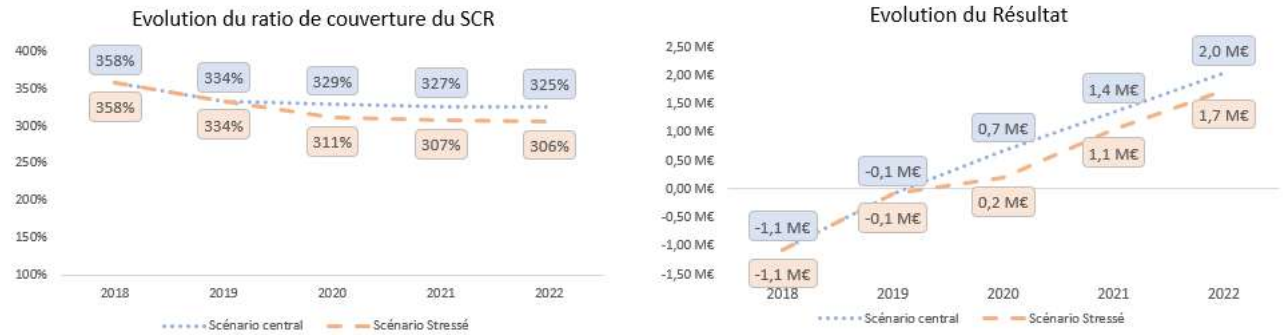

FIGURE 51 – Mise en place du stress-tess 1 : Perte de la réassurance en 2020

Les résultats montrent que la hausse du SCR non-vie est plus importante que la baisse du SCR contrepartie, puisque le ratio SCR global subi un décrochage important en 2020 sans que les fonds propres ne subissent une importante diminution. Le résultat baisse également comme prévu originellement, mais se rapproche peu à peu du scénario central en raison des hypothèses de baisse du ratio P/C émis par le groupe.

Mise en place de la nouvelle prime

L'hypothèse traite désormais de la mise en place de la nouvelle tarification pour le groupe à compter de 2019. Il est proposé de modéliser cela en modifiant les cotisations sans modifier le risque de rachat. L'hypothèse est certes forte, puisque cela suppose une augmentation du coût pour l'assuré qui pourrait s'intéresser à la concurrence, mais l'objectif ici est davantage de visualiser la différence de résultat avec l'ancienne tarification.

En vu des résultats globaux obtenus avant nouvelle répartition avec la nouvelle prime, et par souci de simplification, le stress-test consiste à augmenter les cotisations de 8%. Cela impliquera une augmentation du chiffre d'affaires sans pour autant modifier le montant de prestations à payer.

FIGURE 52 – Mise en place du stress-tess 2 : Nouvelle tarification à partir de 2019

Davantage que l'effet positif qui ressort nettement de ces deux graphiques, qui était attendu, il est intéressant de voir l'impact de la garantie auto sur le SCR et le résultat de l'entreprise. L'augmentation relative du ratio SCR est assez faible, tandis que l'impact sur le résultat est clairement visible selon les hypothèses de projection.

4.4 Réponse à la problématique et prolongements

Les résultats montrent l'utilité d'une nouvelle tarification. Celle-ci est double car en plus d'améliorer le résultat de l'entreprise sans détériorer son ratio de solvabilité, la conservation de la réassurance constitue un enjeu majeur pour le groupe.

Toutefois, les limites sont assez claires. Dans un premier temps, la mise en place de la tarification de manière précise est difficile à réaliser. Le GLM n'étant pas fonctionnel sur une masse si peu importantes de données, une grille tarifaire n'est pas envisageable à partir des modèles de random forest et de gradient boosting considérés dans le mémoire. Comme cela a été précisé avant, le changement nécessiterait idéalement l'acquisition d'une base de donnée plus large et plus propre pour calibrer un modèle de GLM plus performant.

Il est évident que cela représente un coût non négligeable pour le groupe. De plus, la mise en place de cette nouvelle tarification ne pourrait pas se faire immédiatement, et en commençant les travaux dès maintenant, il pourrait être envisageable de faire l'entrée en vigueur du nouveau tarif en septembre 2020.

L'autre aspect qui pourrait être évoqué comme limite est que l'augmentation du tarif ne se fait pas aisément dans un tel secteur. En effet, comme cela était présenté dans l'introduction, l'environnement de la tarification automobile est hautement concurrentiel. Commercialement parlant, il convient de proposer un service de qualité pour justifier une augmentation de tarif. Des aspects tels que l'approche commerciale pour attirer de nouveaux profils, et le traitement des remboursements pourraient faire l'objet d'études approfondies.

En effet, l'étude portée sur les profils porteurs de sinistralité a fait comprendre que les personnes âgées du portefeuille, notamment, donnaient lieu à des ratios SP assez élevés. Il est également ressorti de ces études que le portefeuille vieillissait assez significativement d'une année sur l'autre. La tarification pour les jeunes conducteurs semblant de primer abord adaptée à la sinistralité accrue qui leur correspond, s'intéresser à une manière de capter ces nouveau profils, également pour renouveler le portefeuille, est une première piste pour améliorer la rentabilité sans pour autant augmenter le tarif.

De manière globale, et en suivant une approche plus stratégique que statistique, il conviendrait de revoir la manière de tarifer et en même temps de renouveler le portefeuille. Le changement de portefeuille impliquera que les modèles souffrant de formes de sur-apprentissage deviendront rapidement inadaptés. De ce fait, l'achat d'une base de données globales est d'autant plus intéressant puisqu'elle prendra correctement en compte la captation de ces nouveaux profils qui déséquilibreraient la base de données du groupe.

Concernant le problème de la réassurance, qui sera en partie résolu par l'amélioration de la rentabilité du contrat automobile puisque le contrat est un quote-part, sa conservation peut être assurée en jouant sur le levier de la répartition du volume de prime. Le choix de conserver la rentabilité des primes telles que, comme cela a été montré précédemment, les garanties Panne Moteur et Secours Mutualiste peut à tout moment être remis en question si l'enjeu de la conservation du réassureur se fait prioritaire.

Au delà de cet objectif de conserver le réassureur avec un quote-part élevé, si le portefeuille s'améliore par la suite de façon significative, et que de la rentabilité ressort de ce changement, le groupe sera en mesure peu à peu de reconsidérer le contrat pour baisser la part de quote-part en conservant le contrat XS. En effet, dans la simulation de stress-test correspondante, l'impact sur le SCR est principalement liée à la perte du contrat XS auto qui endiguait fortement le risque catastrophe. La baisse du quote-part pourrait avoir un effet moindre puisque cela laisserait seulement l'effet de l'augmentation du SCR Primes, contrebalancée en partie par la baisse du SCR Contrepartie.

Bien que le secteur automobile reste dominé par des géants du milieu, et qu'il soit difficile de s'y faire une place quand les moyens manquent, les études statistiques et tarifaires montrent que des failles existent dans le fonctionnement du contrat automobile du groupe. En mettant cela en relation avec la qualité des données discutable, et donc les problèmes en résultant pour la calibration d'un modèle utilisable pour l'entreprise, la problématique ne peut être résolue que par la mise en place d'une stratégie précises. Ce mémoire donne une méthodologie destinée à donner des informations sur la marche à suivre et sur les points à améliorer, et fournit un aperçu des résultats qui en ressortirait, mais les solutions adaptées ne peuvent être obtenues qu'en effectuant des choix destinés à agir concrètement sur les faiblesses mises en lumière.

Conclusion

Les travaux menés dans le cadre de ce mémoire ont permis de mettre en lumière plusieurs techniques usuelles appliquées aux problématiques de tarification, ainsi que quelques prolongements lorsque des problèmes telle qu'une masse de données insuffisante se posaient.

La place d'un groupe de petite taille au sein du marché automobile est en effet difficile à cerner. Les résultats sont éloquents à ce sujet puisqu'ils nécessitent l'utilisation de méthodes difficilement utilisables en état pour le groupe. Ce problème pourrait être compensé en effectuant des GLM sur des bases de données achetées. Il aurait également pu être question d'une simulation des données à partir des bases de départ, mais au vu de la qualité de ces dernières, la première option semble préférable. Les résultats font également apparaître la nécessité d'une hausse de la prime, potentiellement difficile à demander en raison de la forte concurrence induite par la marché.

Dans ce cadre, le réassureur n'est pas totalement satisfait du contrat en raison de la nature déficitaire des garanties prévues dans le traité de réassurance. Les déficits observés ont ainsi pour source la répartition des primes au sein des garanties en plus de la nature peu rentable du contrat automobile réassuré avec la tarification actuellement employée par le groupe. Il faut également garder à l'esprit que les travaux présentés dans ce mémoire ne prennent pas en compte l'entièreté des frais imputables aux montants de sinistres, et à l'acquisition des primes, ce qui devrait dégrader les ratios combinés de l'assureur et du réassureur.

La modélisation par le biais des méthodes de Data Science permet toutefois de trouver d'autres solutions pour modéliser les variables telle que la formule kilométrique, aidant à surmonter le problème de qualité des données, et même la fréquence et le coût du sinistre, lorsque pour ces derniers le GLM n'aboutit pas à une estimation satisfaisante. Puisque ce processus permet de compléter la base, cela permet d'affiner le modèle de tarification et de présenter un réel intérêt même dans le cas d'un groupe de petite taille. Ces méthodes peuvent fonctionner avec peu de données, et, malgré la difficulté liée à leur mise en application, permettent au groupe d'obtenir des informations pertinentes sur les problèmes rencontrés, et d'en tirer les conclusions induites.

Suivant ces opérations, à l'aide des observations fournies par les études statistiques descriptives et des résultats des analyses des variables et des primes estimées, il apparaît que les raisons justifiant le manque de rentabilité du côté de l'assureur sont claires, en plus de la mauvaise répartition des primes. D'un côté, la mauvaise tarification de certains profils notamment par rapport aux assurés les plus âgés et ceux n'étant pas situé dans la zone géographique principale du groupe dégrade le ratio S/P, et d'un autre côté, le manque de données rend difficile la mise en place d'un tarif précis et représentatif de la réalité.

Finalement, la mise en place de plusieurs techniques actuarielles a permis de répondre en partie à la problématique, c'est à dire de proposer des solutions aux problèmes de rentabilité du réassureur. Des zones d'ombres subsistent, notamment sur l'obtention d'un outil utilisable de façon réaliste pour un groupe d'une taille aussi réduite, et il apparaît que les résultats notamment pour la modélisation du coût du sinistre peuvent faire l'objet d'améliorations. Le manque de données est toujours difficile à bien prendre en compte malgré le nombre grandissant de méthodes mises à disposition des actuaires, bien que celles-ci permettent l'obtention de résultats provisoires, surmontant les difficultés imposées.

Références

[1] Bertrand, F., Maumy, M. (2008) : Choix du modèle *IRAM, Université de Strasbourg*

[2] Breiman, L. (2001) : Random Forests. *Machine Learning*

[3] Breiman, L., Friedman, J., Olshen, R., Stone, C. (1984) : Classification and regression trees *Wadsworth & Brooks*

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

[5] Chen, T., Guestrin, C. (2016) XGBoost : A Scalable Tree Boosting System *Cornell University.*

[6] Cohen, P. (2016) : De l'appétence aux risques déclarée par les dirigeants d'assurance à la mise en œuvre opérationnelle *Université Paris-Dauphine*

[7] Faugere, N. (1985) : Tarification d'un portefeuille automobile sur le marché luxembourgeois. *CEA.*

[8] Finas, C. (2015) : Les tarifs des réassureurs sont-ils crédibles ?. *ISFA*

[9] Friedman, J. (1999 - 2001) : Greedy function approximation - A gradient boosting machine *IMS Lecture.*

[10] Gonnet, G. (2010) : Etude de la tarification et de la segmentation en assurance automobile *ISFA*

[11] Ottou, P. (2017) : Méthodes d'apprentissage automatique appliquées au provisionnement ligne à ligne en assurance non-vie *Université Paris-Dauphine*

[12] R Development Core Team (2019) : The R Project for Statistical Computing *The R Foundation*

[13] Rakotomalala, R. : Techniques ensemblistes pour l'analyse prédictive *Université Lumière Lyon 2*

[14] Planchet, F., Serdeczny, G. (2014) : Modèles fréquence – coût, quelles perspectives d'évolution ? *Prim'act, MAIF, ISFA*

[15] Tuleau-Malot, C. : Présentation de l'algorithme CART *Université Nice Sophia Antipolis*

Annexe 1 : Exemple de combinaison de deux contrats XS par risque et par évènement

L'exemple se basera sur un contrat XS par risque à deux tranches : 2 000 000 XS 250 000 et 1 750 000 XS 2 250 000. De la même manière, un contrat est parallèlement établi par évènement à trois tranches : 1 500 000 XS 500 000, 8 000 000 XS 2 000 000 et 10 000 000 XS 10 000 000. Le contrat par évènement survient après l'application du contrat par risque. Cela permet de distinguer 3 cas : le cas sans réassurance, le cas avec seulement le contrat XS par risque et le cas avec les deux contrats. Afin de bien visualiser l'intérêt du dernier cas, 4 scénarios sont envisagés :

— Un unique sinistre à 8 000 000, par exemple le vol d'une voiture de collection.
— Plusieurs sinistres de montants assez différents (6 000 000, 250 000, 1 250 000, 2 500 000 et 5 000 000)
— Plusieurs sinistres dont les montants sont assez homogènes (4 000 000, 3 000 000, 500 000, deux fois 2 500 000, 1 000 000 et 1 500 000)
— Quatre sinistres de montants très élevés (8 000 000, 11 000 000, 9 000 000 et 6 000 000)

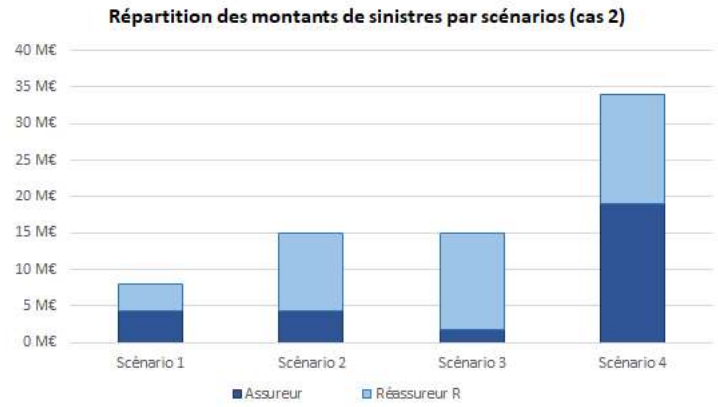

FIGURE 53 – Application dans le deuxième cas

Ce premier graphique montre le poids de la réassurance XS par risque pour endiguer le sinistre. Si le 4ème scénario fait apparaître un montant énorme pour l'assureur puisque le contrat n'est pas prévu pour des montants aussi élevés, l'intérêt réside dans la comparaison des graphiques deux et trois, où il apparaît dans un premier temps l'impact de sinistres plus élevés dans un même évènement, en sachant que le coût total des deux évènements est égal.

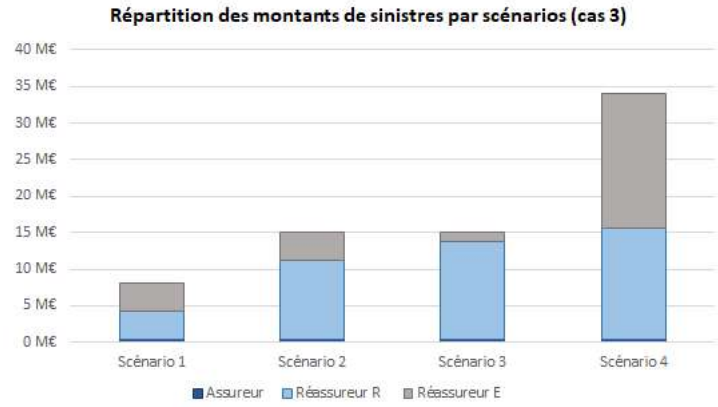

FIGURE 54 – Application dans le troisième cas

Ce graphique introduit le contrat par évènement en sus. Dans un premier temps, il permet de gérer les sinistres

extrêmes qui dépassaient la limite du contrat par risque, comme cela est montré par le premier scénario. Il permet surtout de laisser une charge égale à l'assureur quelque soit le scénario (soit 500 000 euros). Même dans le quatrième scénario, il absorbe totalement les montant particulièrement élevés des quatre sinistres. Il permet également d'éviter à l'assureur de payer chaque minima par sinistre dans le cas où le sinistre entre dans le domaine d'action du traité par risque. Le traité par évènement s'appliquant après celui par risque, il lisse au mieux le montant résiduel laissé par le premier contrat, même dans le cas de nombreux sinistres à montants relativement élevés (cas du troisième scénario).

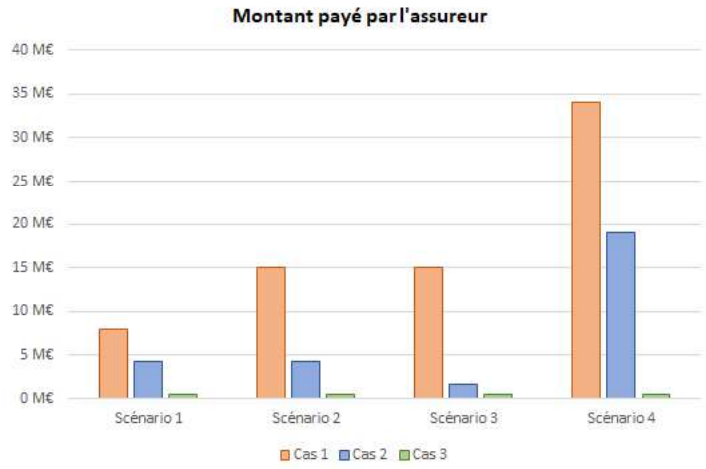

FIGURE 55 – Comparaison des trois cas

Pour finir, ce graphique montre l'évolution du coût pour l'assureur dans chacun des trois cas. L'ajout de la couverture par évènement, en complément de la couverture par risque, permet de balayer davantage de cas qu'en appliquant seulement une couverture par risque. L'intérêt principal maintenant de passer dans un premier temps par une couverture par risque est de simplifier le fonctionnement de la réassurance dans les cas fréquents. Ici, seuls les cas extrêmes parmi les sinistres exceptionnels ont été traités, et il ne fait pas de doute que la couverture par évènement n'entre pas souvent en jeu. Le scénario 3 d'ailleurs ne déclenche que la première tranche du contrat par évènement. De plus, il renforce cette dernière en cas de série de sinistres importants, puisque sans l'application préalable de la couverture par risque, le scénario 4 laisserait à l'assuré un coût de 14 500 000 euros.

La combinaison des deux protège donc d'un grand nombre de sinistre sous un même évènement, ce qui est le but de la couverture par évènement, mais aussi d'un grand nombre de sinistres extrêmes sous un même évènement, et cela passe par l'application d'un contrat par risque en premier lieu.

Annexe 2 : Arbre CART utilisé pour le modèle coût

Ci-après est affiché l'arbre élagué utilisé pour la modélisation du coût du sinistre pour la garantie DTA. Il y apparaît un problème assez caractéristique du CART, déjà apparu lors de l'étude pour la formule kilométrique. En effet, les valeur faibles de sinistres sont mal prédites, ce qui explique l'écart-type important observé en appliquant une prédiction sur les données de test avec ce modèle. La quasi-totalité des variables sont utilisées, malgré le manque de significativité de certaines comme précédemment observé (Enfants et Garage). En effet, le but était de mettre le maximum d'informations à disposition de l'algorithme pour améliorer sa qualité prédictive. Le modèle CART n'est pas retenu en raison de l'efficacité plus intéressante apportée par la méthode du gradient boosting.

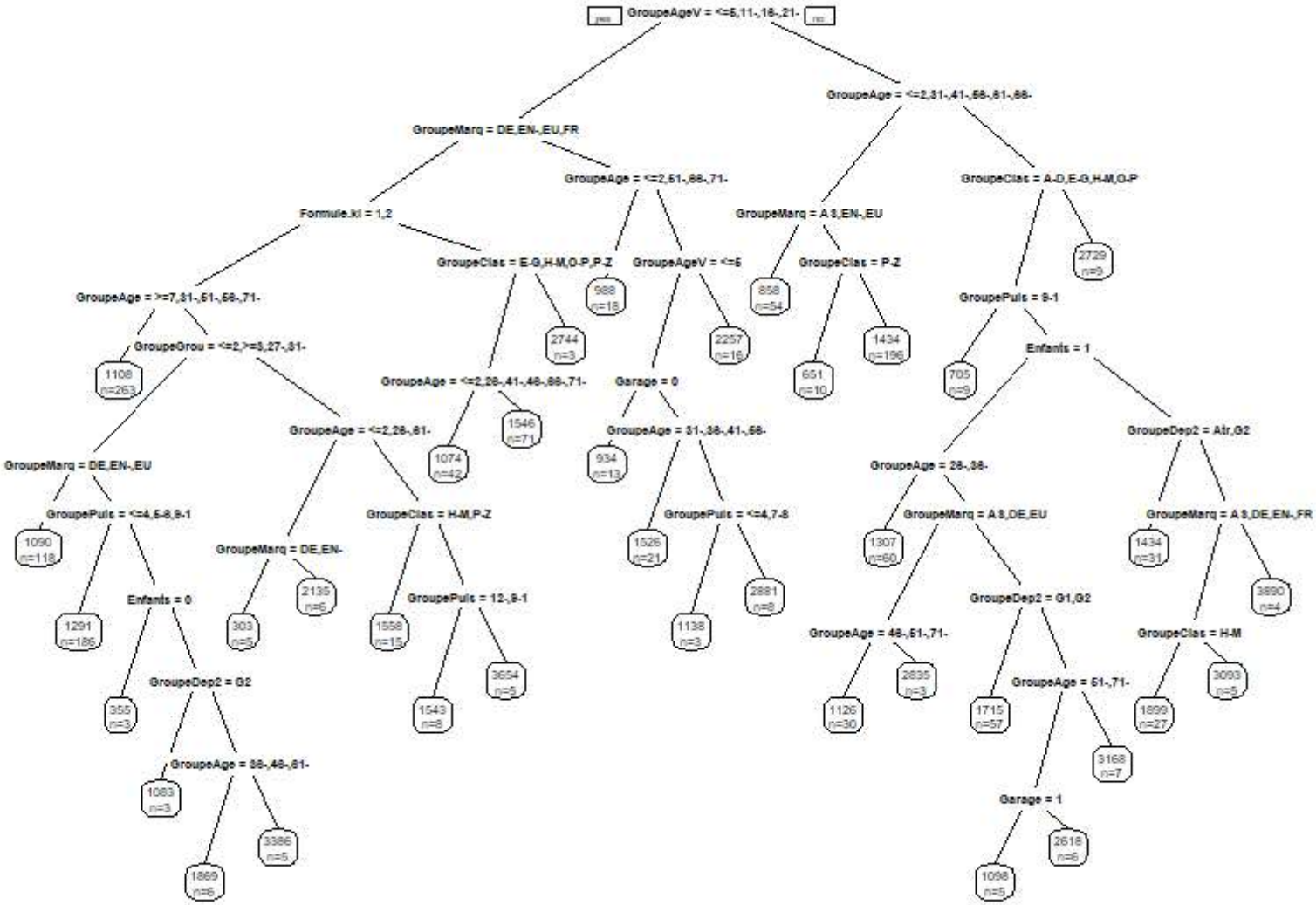

FIGURE 56 – Arbre CART

Annexe 3 : Les trois piliers de Solvabilité II et le SCR

La Directive se repose sur trois piliers distincts formalisant l'ensemble des exigences auxquelles doivent se soumettre les entreprises d'assurance :

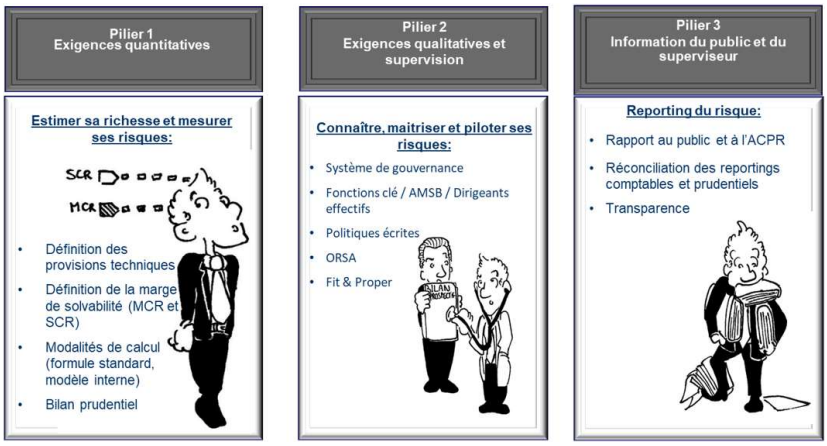

FIGURE 57 – Composition des trois piliers

L'objectif de ces trois piliers est de donner une logique plus organisationnelle, et de ce fait, plus compréhensible auprès des entreprises soumises à la Directive. Les piliers sont liés entre eux malgré les différences d'exigences qui sont impliquées. En effet, l'ORSA se base sur les résultats du pilier 1, et les règles de contrôle interne définissent les normes de calcul du Pilier 1. Enfin, les informations véhiculées dans le cadre du Pilier 3 se basent sur les résultats des Piliers 1 et 2. Ce transfert d'information permet aux administrateurs d'adapter les prises de décisions aux situations rapportées, ce qui par la suite impactera les futurs Piliers 1 et 2.

Pour information, la composition du SCR calculé dans le cadre du pilier 1 est donnée par le schéma suivant :

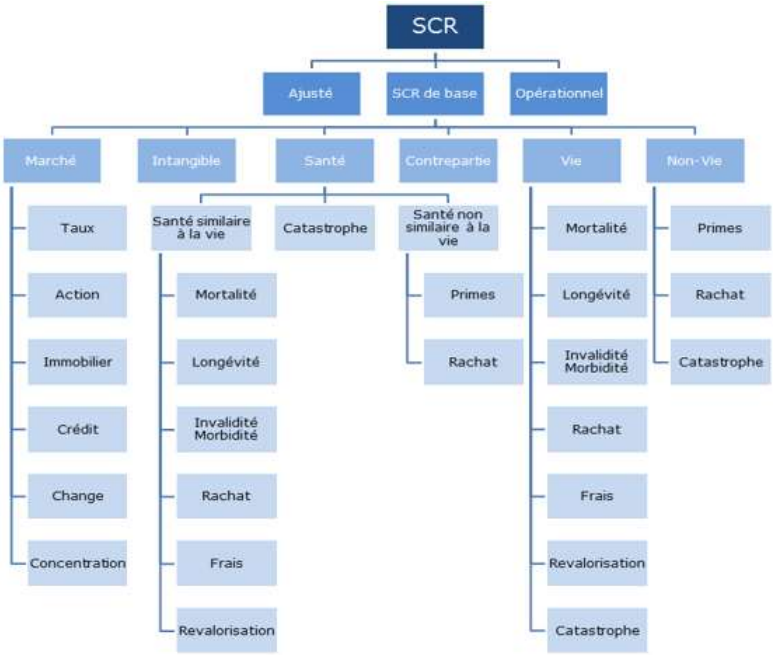

FIGURE 58 – Décomposition du SCR

Annexe 4 : Principaux packages R et utilisation correspondante

— *MASS* : Algorithme de GLM
— *rsq* : Travaux relatifs à l'étude du Chi2
— *corrplot* : Obtention graphique des corrélations
— *ggplot2* : Visualisation de données
— *rpart* : Algorithme de modèles CART
— *rpart.plot* : Obtention graphique des arbres CART
— *caret* : Obtention de matrices de confusion
— *randomforest* : Algorithme de modèles de random Forest
— *gbm* : Algorithme de modèles de Gradient Boosting

L'ensemble de la documentation quant à l'utilisation de ces fonctions peut être trouvée sur le site The Comprehensive R Archive Network.

Table des figures