Mémoire présenté le : 14 mars 2023

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

Par : Nolwenn GUINET

Titre : Révision du tarif sur les flottes automobiles de taille intermédiaire

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

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

M. LOUCHARD

.....

.....

Membres présents du jury de l'ISFA

D. CLOT

.....

.....

Entreprise
Nom : MMA
Signature :

Directeur de mémoire en entreprise
Nom : DUGAST Anne
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

Confidentialité

Pour des raisons de confidentialité, les données (chiffres, graphiques) présentées dans les pages suivantes ont été volontairement modifiées. Les modifications apportées permettent toutefois d'apporter les mêmes conclusions, bien qu'elles ne reflètent pas les données réelles.

Résumé

Lorsqu'il s'agit de tarifer une flotte automobile, les informations relatives au conducteur ne sont pas disponibles. Cependant, d'autres critères tarifaires peuvent être utilisés, comme les caractéristiques de l'entreprise (par exemple son activité ou sa zone géographique) ou celles des véhicules, pour alors obtenir un tarif forfaitaire. Par ailleurs, plus la flotte est de taille importante, plus il y a d'informations sur son profil de risque. Une flotte comportant un nombre important de véhicules sur plusieurs années reflétera, statistiquement, bien sa sinistralité. À l'inverse, pour une flotte comptant un nombre de véhicules plus réduit ou dont l'historique est plus faible, son poids ne peut pas être considéré comme suffisant pour expliquer sa propre sinistralité et son tarif sera plutôt collectif. Nous cherchons donc à ce que le tarif soit progressivement individualisé à partir du tarif forfaitaire, en fonction de l'exposition au risque, pour prendre en compte une partie de l'expérience propre à la flotte. Ce principe d'individualisation du tarif se retrouve dans les méthodes dites de crédibilité.

Ce mémoire a pour objectif de réviser le modèle actuel de crédibilité, mis en place pour tarifer les flottes de taille intermédiaire. Nous nous intéressons tout particulièrement à la composante fréquence du tarif. Les modèles implémentés pour la fréquence collective sont des Modèles Linéaires Généralisés (MLG) avec des lois usuelles (Poisson, Négative Binomiale), mais aussi des MLG avec des lois Zéro-Inflaté ainsi que des forêts aléatoires.

Pour déterminer le degré d'individualisation, appelé coefficient de crédibilité, les méthodes de crédibilité évoquées dans ce mémoire sont le modèle de Bühlmann-Straub et le modèle hiérarchique de Jewell. Cette dernière méthode nécessite une classification de nos données, effectuée avec les méthodes d'Analyse des Correspondances Multiples (ACM) et de Classification Hiérarchique sur Composantes Principales ($HCPC$). Pour être en cohérence avec le modèle de tarification actuel, nous cherchons à modéliser un coefficient de crédibilité sous une forme explicite : une fonction qui dépend de l'exposition au risque de la flotte. Ce tarif révisé a pour objectif d'être plus adapté à la réalité du risque et d'être facilement et rapidement implémentable pour des révisions futures.

Mots clés : flottes automobiles, tarification, base de données, modélisation, MLG, forêt aléatoire, crédibilité, Bühlmann-Straub, Jewell, Analyse des Correspondances Multiples, Classification Hiérarchique sur Composantes Principales

Abstract

When it comes to pricing a vehicle fleet, we don't have any information regarding the driver. However, other pricing criteria can be used, such as the characteristics of the company (for example its activity or its geographical area), or the characteristics of the vehicles to obtain a flat rate on this market. Furthermore, the larger the fleet, the more information there is on its risk profile. A fleet having a large number of vehicles over several years will, statistically, accurately reflect its loss experience. Conversely, for a fleet with a smaller number of vehicles or with a smaller history, we will not be able to consider that its weight is sufficient to explain its own loss experience and its rate will be rather collective. We are therefore looking for the rate to be gradually individualized from the flat rate, depending on the risk exposure, to take into account a part of the experience specific to the fleet. This principle of tariff individualization is found in so-called credibility methods.

This essay aims to revise the current credibility model, implemented to price mid-sized fleets. We are particularly interested in the frequency component of the tariff. The models implemented for the collective frequency are Generalized Linear Models (GLM) with usual laws (Poisson, Negative Binomial), GLMs with zero-inflated laws as well as random forests.

To determine the degree of individualization, called the credibility coefficient, the credibility methods that interest us in this thesis are the Bühlmann-Straub model and the hierarchical Jewell model. This last method requires a classification of our data, carried out with the methods of Multiple Correspondence Analysis (MCA) and Hierarchical Classification on Principal Components (HCPC). To be consistent with the current pricing model, we seek to model a credibility coefficient in an explicit form : a function that depends on the risk exposure of the fleet. This revised tariff aims to better match with the reality of the risk and to be easily and quickly implemented for future revisions.

Keywords : vehicle fleet, pricing, database, modeling, MLG, random forest, credibility theory, Bühlmann-Straub, Jewell, Multiple Correspondance Analysis, Hierarchical Clustering on Principle Components

Remerciements

Mes premiers remerciements vont à ma tutrice Anne DUGAST, actuaire au sein de l'équipe Actuariat Circulation chez MMA, pour son précieux accompagnement tout au long de la réalisation de ce sujet. Je la remercie également pour sa disponibilité et sa confiance à mon égard qui m'a permis de me développer, en autonomie, sur ce sujet technique. Nos riches échanges ont été l'occasion pour moi d'en apprendre beaucoup sur le vaste marché des flottes et de l'actuariat et sont une source d'encouragements.

Je tiens à remercier chaleureusement Sabrina SAVARRE, manager de l'équipe Actuariat Circulation MMA, pour avoir renouvelé sa confiance en moi à la suite de mon stage et de ma première alternance et tout particulièrement de m'avoir encouragée à intégrer l'ISFA à la suite de mon premier master. Merci également de m'avoir proposé un sujet à enjeu si enrichissant. Je tiens à la remercier aussi pour sa disponibilité et ses précieux conseils qui m'ont permis de mener à bien cette étude.

Mes remerciements vont aussi à mes collègues de l'équipe Actuariat Circulation : Roxane TRIFFAULT, Sabrina HERDT-TRAORE, Rose BELAUD, Patrice MAUGENDRE et Arthur CORTAIS. Merci pour votre bonne humeur contagieuse qui crée une ambiance de travail très agréable et propice aux échanges et à la bonne réalisation des travaux. Travailler avec vous est un véritable plaisir.

Je remercie également le corps enseignant de l'ISFA pour leur pédagogie et leurs enseignements qui m'ont permis d'acquérir de solides connaissances et compétences techniques qui ont rendu possible la bonne réalisation de cette étude. Mes remerciements s'adressent tout particulièrement à Nicolas LEBOISNE, mon tuteur pédagogique pour son intérêt et son engagement au regard de mon sujet, ainsi que pour son accompagnement et ses conseils judicieux.

Mes dernières pensées vont à ma famille, mes parents, ma soeur, mon oncle et mes grands-parents. Je vous remercie pour votre soutien infaillible et vos encouragements qui m'ont accompagnée au quotidien jusqu'à cette dernière année d'étude et qui continueront de me porter dans la suite. Merci pour votre optimisme inconditionnel.

Synthèse

Chapitre 1 - Contexte de l'étude

Le premier chapitre de ce mémoire permet de placer le contexte de l'étude et ainsi de définir le périmètre concerné par la refonte du tarif. L'étude se place sur le marché des flottes automobiles. Une flotte est l'ensemble des véhicules dont dispose une entreprise. Les véhicules sont regroupés en différents ensembles qui dépendent de leurs caractéristiques. Au sein de chaque ensemble, le tarif proposé est identique pour tous les véhicules.

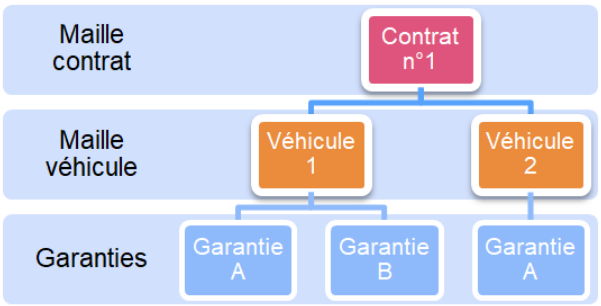

FIGURE 1 – Description de l'emboîtement des mailles

Le marché des flottes est fragmenté en cinq segments pour faciliter sa gestion. Les segments ainsi créés dépendent du nombre de véhicules au sein de la flotte et la gestion diffère selon le segment. Sur les flottes de petite taille la tarification appliquée est un tarif forfaitaire, qui dépend donc de critères tarifaires communs à tous. À l'inverse sur les flottes de taille plus importante, le tarif choisi est un tarif sur-mesure, complètement individualisé selon l'historique de la flotte. Le segment sur lequel porte l'étude est celui des flottes de taille intermédiaire, entre une tarification forfaitaire et sur-mesure. L'idée est donc de partir d'un tarif forfaitaire et de l'individualiser progressivement selon le poids de la flotte.

Parmi la gamme de produits disponibles sur le marché des flottes, celui qui fait l'objet de la révision tarifaire est le produit AutoFleet. Ce produit est aussi appelé TPPC pour Transport Privé pour Propre Compte. Comme son nom l'indique, ce produit assure les véhicules qui sont utilisés dans le cadre de l'activité de l'entreprise uniquement, ne sont pas concernés les véhicules alloués au transport de marchandises ou de personnes pour le compte d'autrui.

Sur ce produit, différentes garanties peuvent être souscrites. Celles qui vont faire l'objet des premières études sont les garanties Responsabilité Civile (RC), Dommages Tous Accidents (DTA) et Bris de Glace (BdG), du fait de l'enjeu qu'elles représentent sur le périmètre présenté. Et plus particulièrement la garantie DTA qui est choisie pour illustrer les résultats tout au long de ce mémoire et qui fait l'objet des premiers modèles de tarification.

Le tarif actuellement mis en place sur les flottes de taille intermédiaire pour le produit TPPC part du principe de crédibilité qui s'applique sur la partie fréquence de la prime pure. La prime pure est calculée avec la méthode fréquence × coût moyen et s'applique à la maille contrat. La fréquence crédibilisée s'écrit ainsi :

$$Freq^{cred}_{contrat} = Freq^{obs}_{contrat} \times z + Freq^{th}_{contrat} \times (1 - z)$$

Le coefficient de crédibilité z est une fonction déterministe de paramètre le nombre de véhicules. La fonction est différente pour chaque garantie. Ainsi la fréquence théorique forfaitaire est progressivement individualisée en fonction du nombre de véhicules de la flotte. Cette fréquence théorique est déterminée selon le croisement de deux critères tarifaires : la zone d'activité de l'entreprise et son activité principale. Les zones sont numérotées de 1 à 6, par risque croissant, et les activités de 1 à 8, sans ordre de risque.

L'exposition au risque sur les différentes catégories est hétérogène. En effet, sur le critère de zones par exemple, la zone 6 est largement sous représentée avec une part dans portefeuille de 2% seulement.

L'objectif de l'étude présentée dans ce mémoire est de refondre le tarif présenté ci-avant. Une première refonte a été étudiée. Elle avait pour ambition de revoir le tarif entièrement et à la maille véhicule, notamment pour y intégrer de nouveaux critères tarifaires. Cependant cette maille véhicule s'est avérée inexploitable du fait de problématiques sur les données (données manquantes, mal renseignées). La refonte tarifaire prend alors un nouvel axe et l'objectif est maintenant de mettre à jour les éléments composants le tarif. Les travaux s'orientent donc sur la mise à jour de la fréquence théorique à la maille contrat sur le croisement de la zone et de l'activité de l'entreprise et la refonte du coefficient de crédibilité z, fonction de la taille de la flotte.

Chapitre 2 - Mise en place d'une base de données d'étude

Dans un premier temps, il est nécessaire de mettre en place une base de données pour l'étude de la refonte tarifaire. Les caractéristiques des contrats et les informations des garanties et des sinistres qui y sont liés doivent apparaître dans cette base. La maille d'intérêt est la maille contrat.

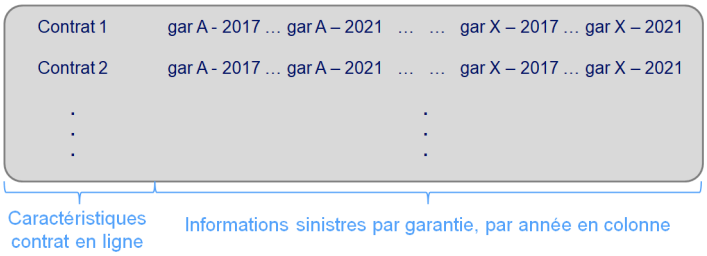

FIGURE 2 – Format de la base de données

Le choix a été fait de conserver un historique sinistre de 4 ans pour correspondre aux relevés d'informations disponibles. L'année 2020 étant particulièrement atypique du fait de la crise Covid-19 et montrant un fort impact à la baisse sur la fréquence sinistre, cette année est écartée de la base. Les données conservées sont donc celles de janvier 2017 à décembre 2021, hors année 2020. Sur cet historique, les informations dans le périmètre d'étude sont conservées : le segment 3, le produit TPPC et les 3 garanties DTA, RC et BdG. Afin d'étoffer la base de données le segment 2 est aussi considéré du fait de sa similarité avec le segment 3 en terme de sinistralité.

Les données nécessaires, contrats et sinistres, sont en grande partie disponibles dans les bases de données de l'entreprise. Celles qui n'y sont pas pourront être créées à partir des existantes.

Certaines hypothèses doivent être émises du fait des données ou de la gestion particulière du segment 3. Concernant les sinistres, la gestion des sinistres atypiques et attritionnels est la suivante :

- Gestion par écrêtement :
  - Charge > seuil de grave : charge hors grave = seuil et charge grave = charge – seuil
  - Charge > seuil de grave : charge hors grave = charge et charge grave = 0
- Seuils de grave définis par garantie

La charge sinistre est gérée ainsi :

- Charge nette de recours évalué
- Sauf sinistres convention IDA : Charge brute de recours évalué

Les sinistres sont comptabilisés selon la responsabilité de l'assuré :

Garantie	Fréquence	Coût
DTA	Responsables uniquement	Tous
RC Matériel	Responsables uniquement	Considéré en DTA
RC Corporel	Responsables uniquement	Considéré en RC
Sans suite	Tous	Non

TABLE 1 – Gestion de la responsabilité

L'exposition au risque est calculée comme suit :

$$\text{Taux de moteurs année} = \frac{\text{Nombre de véhicules moteurs}}{\text{Taux de présence sur l'année}}$$

Puisque les coûts des sinistres ont vocation à être étudiés dans un second temps, mais que le périmètre s'étend jusqu'à 2017, il est nécessaire de retraiter les charges pour qu'elles puissent être considérées en "euros 2021", la dernière année de vision. En effet, dû à des facteurs comme l'inflation et l'augmentation des coûts de réparation automobile, un sinistre survenu en 2017 n'a pas le même coût que s'il était survenu en 2021. D'autant plus que les coûts de réparations ont largement augmenté ces derniers temps, avec aujourd'hui en plus la problématique de disponibilité de pièces automobiles. Le premier indicateur utilisé pour évaluer cette inflation est l'indice SRA (Sécurité et Réparation Automobiles), qui recense l'évolution trimestrielle des principaux éléments constituant le coût de la réparation des véhicules. Le second est l'historique d'évolution du forfait IDA (Indemnisation Directe de l'Assuré), issu de la convention IRSA. Ce forfait d'indemnisation forfaitaire est calculé en fonction du coût moyen des réparations ainsi que du prix de la main d'œuvre et des pièces détachées. Pour la mise en *as-if* des coûts à 2021 la statistique utilisée est la moyenne de ces deux indicateurs d'évolution. En effet ces indicateurs sont assez proches et leur moyenne permet de ne pas surestimer ni sous-estimer l'inflation.

Les seuils de graves utilisés pour déterminer les sinistres atypiques sont les seuils définis par l'entreprise. Ce mémoire n'a pas pour ambition de remettre en cause les seuils définis mais plutôt de les valider, de vérifier leur cohérence avec les données sélectionnées. Pour ce faire trois méthodes ont été implémentées : la méthode des quantiles, un *Mean Excess Plot* et un *Hill Plot*. Ces techniques de détermination de seuils permettent de conclure sur la justesse des seuils définis par l'entreprise sur les données.

Certains sinistres sont survenus sur le périmètre d'étude, mais ne sont pas encore déclarés, ou leur coût ultime n'est pas encore connu du fait que leur liquidation ne soit pas terminée. Ces sinistres *IBNR* (*Incurred But Not Reported*) sont à estimer pour obtenir le coût ultime des sinistres sur le périmètre. La méthode de provisionnement utilisée est celle de *Chain Ladder*, basée sur l'historique des données et qui estime des facteurs de développement à partir de triangles de liquidation. Les facteurs de développement de *Chain Ladder* f_j se calculent de la manière suivante, avec i l'année de survenance sinistre et j l'année d'ouverture :

$$f_j = \frac{\sum_{i=1}^{n-j} X_{i,j+1}}{\sum_{i=1}^{n-j} X_{i,j}}$$

Et les charges provisionnées :

$$\widehat{X}_{i,j+1} = X_{i,j} \times f_j \Leftrightarrow \widehat{X}_{i,j+k} = X_{i,j} \times \prod_{l=0}^{k-1} f_{j+l}$$

Cette méthode repose sur une hypothèse forte : Les facteurs de développement individuels (ou coefficients de passage) doivent être indépendants de l'année de survenance. Il est également nécessaire de vérifier que la profondeur du triangle est suffisante pour représenter les délais d'ouverture des sinistres. La méthode de *Chain Ladder* peut alors être appliquée.

	DTA				
Annee ouverture	N	N+1	N+2	N+3	N+4
Annee survenance					Ultime
2017	36 418 829	44 491 847	44 709 636	44 726 685	**44 738 334**
2018	39 337 043	47 238 691	47 474 363	47 501 925	47 514 297
2019	41 784 323	50 627 629	50 890 202	50 914 830	50 928 090
2020	31 832 933	37 783 964	37 974 010	37 992 387	38 002 282
2021	37 853 850	45 651 271	45 880 888	45 903 092	45 915 047
		1,2060	1,0050	1,0005	1,0003

FIGURE 3 – Complétion du triangle inférieur par *Chain Ladder*

Suite à la mise en place de la base de données, les statistiques des variables d'intérêt peuvent être exploitées. Dans un premier temps il est intéressant de se pencher sur l'exposition au risque et la fréquence sinistre qui fait l'objet de la refonte.

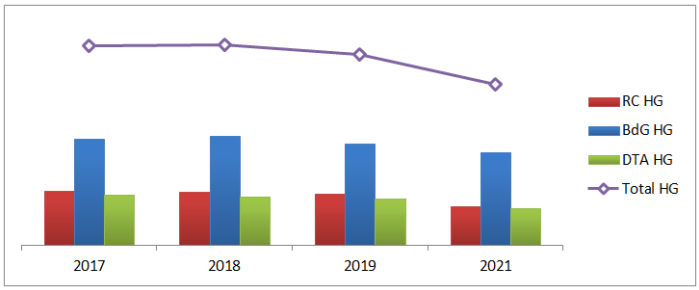

FIGURE 4 – Évolution de la fréquence totale et par garantie

La fréquence sinistre sur le portefeuille montre une tendance à la baisse, tendance portée par les garanties RC et DTA principalement.

Cette tendance se retrouve aussi sur les critères tarifaires de zone et d'activité, sur lesquels il est possible de dégager des tendances entre les modalités. Ce qui sera mis en avant lors des modélisations.

Chapitre 3 - Révision de la fréquence collective

Le troisième chapitre est dédié à la refonte de la fréquence collective :

$$\text{Fréquence} = \frac{\text{Nombre de sinistres}}{\text{Exposition}}$$

La première étape, avant de pouvoir commencer les modélisations est d'étudier la loi des fréquences. Les deux lois classiques d'adéquation sont la loi de Poisson et la loi Négative Binomiale. Après la mise en place de tests d'adéquation, la loi retenue est la loi Négative Binomiale de paramètres $n = 0.66$ et $p = 0.21$.

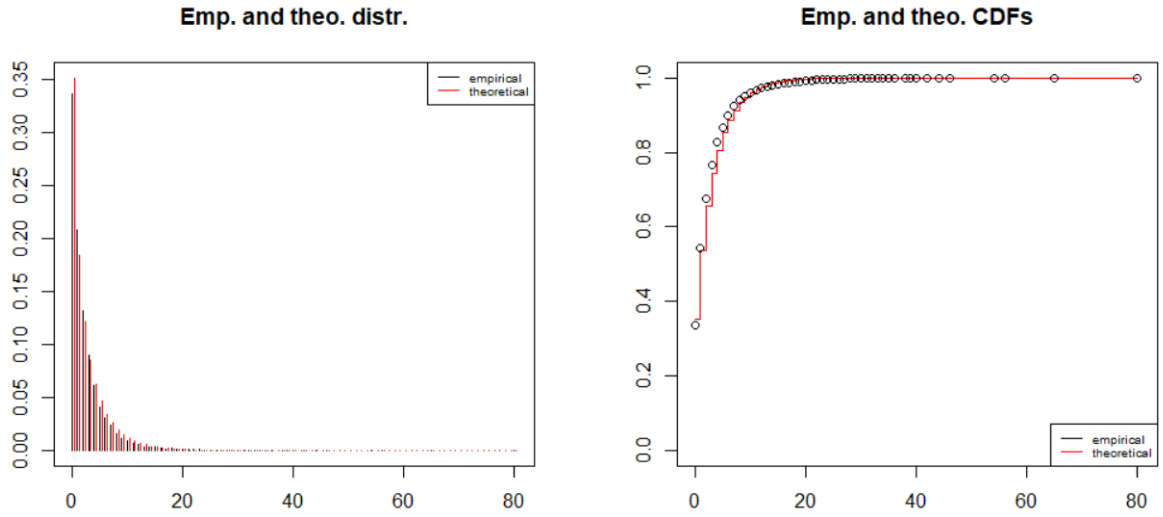

FIGURE 5 – Histogrammes et fonctions de répartition empiriques et estimés par une loi Négative Binomiale

	Poisson	Négative Binomiale
AIC	84 024.59	57 384.30
Maximum de vraisemblance	42 011.29	28 690.15

TABLE 2 – Comparaison des indicateurs de performance des approximations en loi

Les modèles de fréquences théoriques sont entraînés par Validation Croisée. Les premiers modèles testés sont des modèles MLG, avec comme loi sous-jacente les deux lois d'adéquation Négative Binomiale et Poisson. Sans surprise le modèle le plus performant est le modèle avec loi Négative Binomiale.

D'autres modèles sont aussi testés, et notamment des modèles dits Zéro-Inflaté. Ces modèles ont été considérés du fait de l'allure de la distribution des fréquences qui fait apparaître un fort poids en 0. Les modèles Zéro-Inflaté permettent de prendre en compte ce poids en se basant sur deux lois :

— une distribution binomiale qui estime la masse en 0
— une distribution de comptage (Poisson ou Négative Binomiale par exemple) non tronquée, qui peut aussi prendre la valeur 0

Ainsi, les modèles Zéro-Inflaté considèrent deux sources de 0 : les non-observés et les observés avec la valeur 0.

Dans le même esprit, les modèles *Hurdle* sont aussi testés. Ces modèles comptent également deux composantes qui sont :

— une distribution binomiale qui estime la masse en 0
— une distribution de comptage tronquée en 0

Une seule source de 0 est prise en compte ici : les 0 effectivement observés.

En sortie de MLG Zéro-Inflaté et *Hurdle*, il y a donc 2 prédictions : celle du modèle Binomial et celle de la loi de comptage, donc 2 fois plus de coefficients.

Un modèle de prédiction de *Machine Learning* est également testé car réputé pour ses performances : la forêt aléatoire. Une optimisation du paramètre du nombre d'arbres de la forêt est mise en place en amont.

Au regard des sorties et des indicateurs de performance, le modèle finalement sélectionné est le MLG Négatif Binomial classique.

	MLG		Zéro-Inflaté		*Hurdle*		Forêt
	Poisson	Négatif Binomial	Poisson	Négatif Binomial	Poisson	Négatif Binomial	/
MSE	7.774	7.702	7.936	7.824	7.777	7.803	7.784
écart au total	711	291	910	1044	945	798	735

TABLE 3 – Comparaison des indicateurs de performance des modèles par Validation Croisée

De plus, une étude de la sensibilité des modèles a été mise en place. Grâce aux Validations Croisées sur l'entraînement des données, il a été possible de récupérer les écarts quadratiques sur chaque échantillon de validation pour étudier la volatilité des modèles. Sur le modèle Négatif Binomial par exemple, 75% des *MSE* calculés sur les différents échantillons sont en dehors de l'intervalle à 95% autour de la moyenne des *MSE*. Ce taux est très élevé et indique une importante volatilité sur ce modèle. En étudiant de plus près les *MSE* les plus élevés, il s'agit des échantillons dans lesquels se trouvent les nombres de sinistres les plus élevés. En effet un contrat peut compter jusqu'à 80 sinistres, mais ces valeurs élevées ont du mal à être captées par le modèle et viennent biaiser les indicateurs de performance. Ces sinistres exceptionnels pourront notamment être captés grâce à l'intervention de la crédibilité dans un second temps.

Un avantage des MLG est qu'ils proposent en sortie de modèle des coefficients explicites pour chaque modalité. Ce qui permet de les retravailler au besoin. En sortie des modèles prédictifs, les coefficients des zones sont bien croissants, reflétant ainsi correctement la sinistralité croissante, mais jusqu'à la zone 5 seulement. En effet, du fait de la faible représentation de la zone 6, le coefficient estimé ne représente pas forcément la réalité. Il est néanmoins possible de le retraiter. Pour obtenir un coefficient 6 représentatif de la sinistralité du zonier, plusieurs options sont possibles. Les options considérées ici sont : une moyenne de la croissance entre les zones 1 à 5, une estimation du coefficient via une régression linéaire et une régression polynomiale de degré 2. Les résultats obtenus par ces méthodes sont assez proches et le coefficient finalement retenu est la moyenne des trois résultats.

Chapitre 4 - Révision des coefficients de crédibilité

Le modèle actuel construit pour déterminer les coefficients de crédibilité se présente sous la forme de fonctions explicites qui dépendent du nombre de moteurs dans la flotte.

$$z_{DTA} = \sqrt{\frac{n_{DTA} - c}{d - c}}$$

Ce coefficient est forcé à 0 et à 1 aux bornes qui définissent le segment 3, les bornes sont déterminées par c et d dans la formule précédente. La croissance de z est plus forte sur les flottes de petits moteurs que sur les plus

importantes qui sont plus individualisées.

Dans cette formule, n'est utilisée que l'information de l'exposition de la dernière année. Cependant, intuitivement un contrat avec 4 ans d'expérience et un contrat avec le même nombre de moteurs mais avec moins d'expérience ne peuvent être individualisés de la même manière. C'est pourquoi dans la suite, en plus de revoir la formule de crédibilité l'objectif est d'intégrer la dimension de l'expérience, en plus du nombre de moteurs, via l'exposition au risque. Deux théories de la crédibilité vont donc être utilisées afin de réviser le coefficient de crédibilité, à savoir les modèles de Bühlmann-Straub et de Jewell, qui ont l'avantage de prendre en compte l'exposition au risque grâce à l'introduction de poids. Ces modèles permettent de déterminer l'estimateur homogène de crédibilité donné par :

$$\widehat{\widehat{\mu(\Theta_i)}} = z \times \bar{X}_i + (1-z) \times \widehat{\widehat{\mu_0}}$$

Avec :

- $\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$: l'estimateur de la fréquence individuelle
- $\mu_0 = \mathbf{E}[\mu(\Theta_i)] = \mathbf{E}[X_{ij}]$: la fréquence collective et $\widehat{\widehat{\mu_0}}$ son estimateur

Sous hypothèses, le coefficient de crédibilité de Bühlmann-Straub s'écrit de la manière suivante :

$$z_i = \frac{\omega_{i\bullet}}{\omega_{i\bullet} + \frac{\sigma^2}{\tau^2}}$$

Avec ω_{ij} l'exposition du contrat i sur l'année j, σ^2 la variabilité interne du risque et τ^2 l'hétérogénéité du portefeuille, dont les estimateurs sans biais et convergents sont les suivants :

- $\widehat{\sigma^2} = \frac{1}{I(n-1)}\sum_{i=1}^{I}\sum_{j=i}^{n}\left(X_{ij} - \bar{X}_i\right)^2$
- $\widehat{\tau^2} = \frac{1}{I-1}\sum_{i=1}^{I}\left(\bar{X}_i - \bar{X}\right)^2 - \frac{\widehat{\sigma^2}}{n}$

Le modèle proposé par Bühlmann et Straub suppose une certaine homogénéité du risque dans le portefeuille, ce qui en pratique n'est pas forcément le cas. Jewell revient donc sur ce modèle en 1975 pour introduire une crédibilité hiérarchique qui permet de subdiviser un portefeuille en différentes classes de risques. Ici, le modèle présente 3 niveaux : les observations, les individus et un regroupement en classes de risques homogènes.

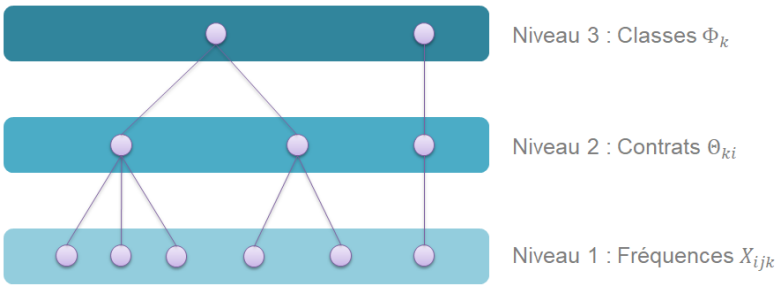

FIGURE 6 – Schéma de la hiérarchie du classement pour la mise en place de la crédibilité

Sur le principe, la crédibilité de Jewell fonctionne de la même manière que Bühlmann-Straub, sur chacun des niveaux. Son estimateur de crédibilité est le suivant :

$$\widehat{\widehat{\mu(\Theta_{ki}, \Phi_k)}} = z_{ki}\bar{X}_{ki} + (1-z_{ki})\widehat{\widehat{\mu(\Phi_k)}}$$
$$\widehat{\widehat{\mu(\Phi_k)}} = z_k\bar{X}_k + (1-z_k)\widehat{\widehat{\mu_0}}$$

Avec :

- $\bar{X}_{ki} = \sum_{j=1}^{n_{ki}} \frac{\omega_{kij}}{\omega_{ki\bullet}} X_{kij}$: moyenne individuelle pondérée des poids
- $\bar{X}_k = \sum_{i=1}^{I_k} \frac{z_{ki}}{z_{k\bullet}} \bar{X}_{ki}$: moyenne de la classe pondérée des coefficients de crédibilité
- $\widehat{\widehat{\mu_0}} = \sum_{k=1}^{K} \frac{z_k}{z_\bullet} \bar{X}_k$: moyenne collective pondérée

Les coefficients de crédibilité sont de la forme suivante :

$$z_{ki} = \frac{\omega_{ki\bullet}}{\omega_{ki\bullet} + \frac{\sigma^2}{a}} \qquad \omega_{ki\bullet} = \sum_{j=1}^{n_{ki}} \omega_{kij}$$
$$z_k = \frac{z_{k\bullet}}{z_{k\bullet} + \frac{a}{b}} \qquad z_{k\bullet} = \sum_{i=1}^{I_k} z_{ki}$$
$$z_\bullet = \sum_{k=1}^{K} z_k \qquad z_{\bullet\bullet} = \sum_{k=1}^{K} z_{k\bullet}$$

Avec a et b les estimateurs des variances intra et inter-classes.

Cette méthode requiert une classification des contrats en classes de risques homogènes. Ces classes vont dépendre du croisement zone × activité. Pour réduire le nombre de croisements une Analyse des Correspondances Multiples est mise en place en amont de la classification. En sortie d'ACM, 5 axes principaux contiennent les informations des 48 croisements. La méthode de classification utilisée est une Classification Hiérarchique sur Composantes Principales pour créer des classes effectives, fonction de la zone et de l'activité du contrat. Cette étape est facilitée par la mise en place de l'ACM qui réduit le nombre de croisements à étudier.

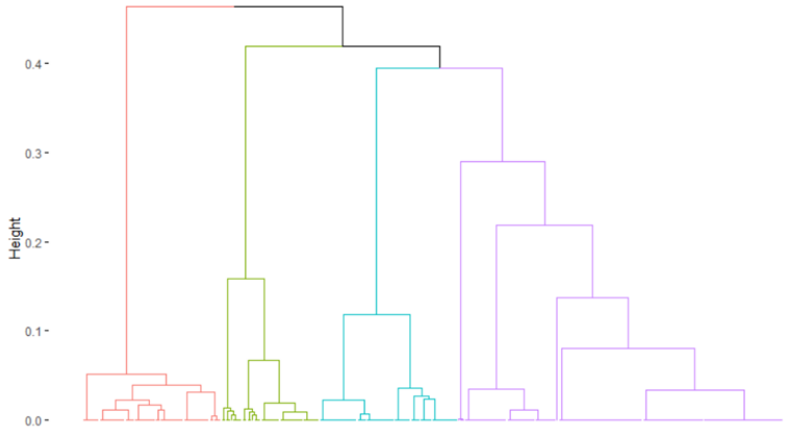

FIGURE 7 – Dendogramme de l'*HCPC*

4 classes de risques sont ainsi créées.

La crédibilité peut alors être mise en place. Les deux méthodes donnent finalement des résultats très proches. La crédibilité de Jewell n'apporte pas de plus-value, mais de l'incertitude avec la création des classes. La méthode finalement choisie est donc celle de Bühlmann-Straub.

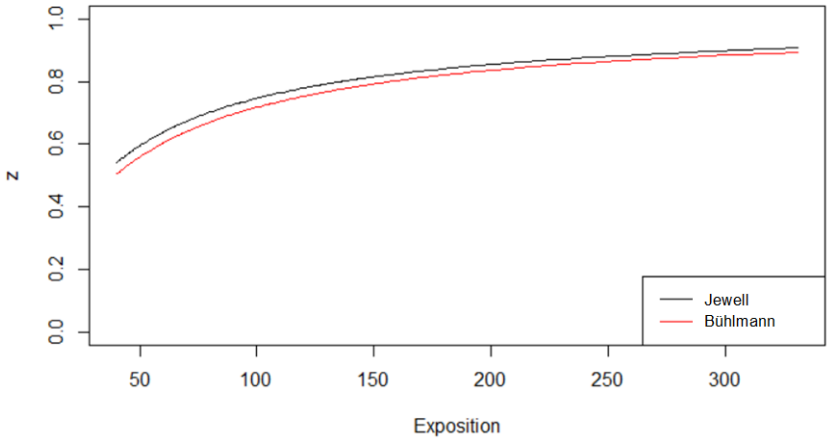

FIGURE 8 – Facteurs de crédibilité estimés par méthode de Bühlmann-Straub et de Jewell en fonction de l'exposition sur les 4 dernières années

Les coefficients de crédibilités semblent plutôt élevés sur les expositions les plus faibles, là où il était forcé proche de 0 avec la formule du tarif actuel. Cette observation peut s'avérer problématique, notamment vis à vis de la continuité avec le segment 2, si les fréquences collectives et individuelles sont éloignées sur les faibles expositions, autrement dit si l'individualisation y est constatable. Ce point sera étudié dans la suite.

À partir de l'estimation du coefficient de crédibilité, l'objectif est d'avoir une fonction explicite qui dépend de l'exposition au risque, pour fonctionner de manière analogue au modèle actuel. En étudiant l'allure de la courbe de z en fonction de l'exposition, un MLG de la forme suivante est mis en place :

$$z = \beta_0 + \beta_1 \times log(\text{Exposition})$$

Les coefficients sont estimés afin d'obtenir la fonction explicite.

Tous les éléments de la fréquence peuvent maintenant être mis à jour et être comparés avec le tarif existant afin d'étudier l'impact de cette refonte tarifaire. En moyenne, la fréquence collective révisée est 56% moins élevée que l'actuelle, écart qui n'est pas homogène sur les critères de zone et d'activité et relativement important. Avec l'individualisation de la fréquence, ces écarts sont voués à s'amoindrir. C'est certes ce qui est observé sur les fréquences crédibilisées, mais en moyenne l'écart tarifaire est tout de même de -36%. L'individualisation creuse les disparités sur les croisements de zone et d'activité. La figure suivante illustre l'écart tarifaire.

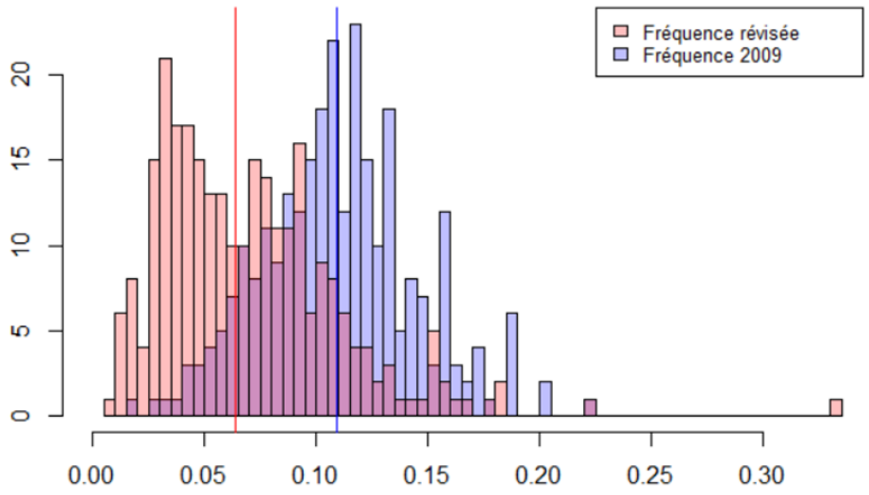

FIGURE 9 – Histogramme des fréquences crédibilisées révisées et actuelles

Aussi, pour évaluer ce nouveau modèle, il va être intéressant de se pencher sur les écarts tarifaires par tranche de risque. Cette étape permet de vérifier que les "bons contrats", les contrats avec peu de sinistralité, ne sont pas pénalisés par cette refonte. En d'autres termes, les fréquences modélisées sur les tranches de risque basses doivent rester relativement basses. Même raisonnement sur les tranches de risque hautes. La fréquence collective pénalise les bons contrats avec une fréquence modélisée en moyenne 4 fois supérieure à la fréquence observée sur les tranches de fréquence les plus basses. La crédibilité permet d'être en meilleure adéquation avec les observations et plus proche de la réalité tout en gardant une part forfaitaire.

Il est également important de se pencher sur la continuité tarifaire entre le segment 3 et ses voisins les segments 2 et 4. Par construction du coefficient de crédibilité qui est proche de 1 sur les grandes expositions, la continuité avec la tarif individualisé du S4 est assurée. Cependant, ce n'est pas si évident avec le segment 2, les coefficients de crédibilité étant particulièrement élevés dès les expositions les plus faibles. En effet, en étudiant de plus près les écarts tarifaires entre la fréquence collective et la fréquence crédibilisée, l'individualisation joue un rôle important avec des écarts d'environ 60% en valeur absolue.

À ce stade, le tarif proposé sur ce périmètre est plus proche de la réalité car plus individualisé. Mais il ne permet pas forcément de tarifer des contrats dont l'historique est trop court pour bien apprécier son risque. En effet, un contrat avec une sinistralité nulle du fait de son manque d'expérience ou sa faible exposition sera statistiquement enclin à subir des sinistres dans les années suivantes. Ce qui peut être pris en compte avec des modèles de fréquence collective, mais pas si le tarif se base uniquement sur l'expérience du contrat.

Dans la continuité de cette étude, des travaux pourraient être mis en place pour la mise en place de plancher de fréquence par exemple, afin d'équilibrer le tarif.

Sommaire

Introduction générale

Ce mémoire a pour ambition de présenter les différentes étapes de la révision du tarif sur les flottes automobiles de taille intermédiaire. Sur ce marché, les méthodes de tarification peuvent différer selon le nombre de véhicules de la flotte. En effet, pour une flotte de taille importante, nous pouvons considérer que son expérience et sa sinistralité passée suffisent à prédire sa sinistralité future, du fait de son importante exposition au risque. À l'inverse, sur des flottes de petite taille, comme pour les véhicules particuliers, nous préférons une tarification collective, du fait que l'exposition au risque de ces flottes n'est pas suffisante pour apprécier sa sinistralité future. Nous parlons alors de tarif forfaitaire, ou *a priori*, car ce type de tarification se base sur des critères tarifaires qui sont identiques pour tous les assurés concernés.

La problématique présentée dans le premier chapitre, porte sur les flottes de taille intermédiaire, qui peuvent prétendre à un tarif partiellement individualisé, mais pas complètement du fait que leur poids n'est pas suffisamment représentatif pour juger leur sinistralité uniquement sur leur expérience. Cette individualisation du tarif permet notamment d'éviter l'antisélection, car elle permet de favoriser les meilleurs risques. Les flottes sur lesquelles porteront l'étude sont celles qui souscrivent au produit phare des flottes MMA : le produit Transport Privé pour Propre Compte (TPPC). Parmi les nombreuses garanties pouvant être proposées sur ce produit, nous nous intéressons dans un premier temps aux garanties Dommages, Responsabilité Civile et Bris de Glace. Ce choix est notamment motivé par le fait que ces trois garanties sont celles principalement souscrites sur le produit.

L'objectif de ce mémoire est alors de déterminer une méthode de tarification adaptée aux flottes de taille intermédiaire. Nous choisissons de nous intéresser aux méthodes de crédibilité. Ces méthodes se basent sur un tarif forfaitaire, commun à tous, qui est ensuite individualisé progressivement en faisant appel à un coefficient, appelé coefficient de crédibilité. Une forme de crédibilité est actuellement mise en place dans le tarif existant, c'est pourquoi nous choisissons d'étudier ces méthodes dans un objectif de refonte du tarif. Cette crédibilité n'est appliquée que sur la composante fréquence sinistre de la prime, et non sur le coût moyen. L'objectif est alors de réviser cette fréquence crédibilisée : son coefficient de crédibilité et sa fréquence collective.

Ce sujet est motivé par la nécessité de réétudier le tarif actuellement mis en place pour les flottes de taille intermédiaire sur le produit TPPC. Le tarif proposé par le modèle actuel n'a pas été revu depuis quelques années et il semble décadré par rapport au marché et est dérogé par les équipes de souscription. Une refonte du tarif permettrait alors d'être plus en adéquation avec le marché et notre portefeuille.

Pour mener à bien cette refonte tarifaire, notre étude s'articule en trois axes principaux.

Dans un premier temps, il est nécessaire de mettre en place une base de données propre à l'étude. Cette base doit permettre d'étudier la sinistralité sur notre périmètre. Elle doit donc prendre en compte la fréquence, notre premier objet d'étude, mais aussi le coût moyen. Cette première étape est chronophage mais primordiale car notre base de données doit correctement refléter notre portefeuille et ses sinistres sur le périmètre d'étude. Ainsi nous devons faire des hypothèses liées aux contraintes métier, comme la gestion de la responsabilité des sinistres ou celle des sinistres graves par exemple. Il va aussi être nécessaire d'actualiser nos sinistres, notamment leur coût pour que l'ensemble de notre base soit à la même vision, la vision actuelle. En effet, le coût d'un sinistre il y a X années n'est pas le même que le coût d'un sinistre aujourd'hui, du fait de paramètres comme l'inflation ou le coût des matériaux et de la réparation par exemple. Nous devrons également prendre en compte et évaluer le montant des sinistres survenus sur notre périmètre mais qui ne sont pas encore déclarés. Pour cela nous utiliserons la méthode de provisionnement de *Chain Ladder*.

Assurés de la fiabilité de la base de données d'étude, les premières modélisations peuvent être mises en place. Nous commençons par étudier la fréquence collective. Grâce à des méthodes d'approximation, il est possible de déterminer les lois théoriques sous jacentes à la distribution des fréquences ainsi que leurs paramètres. Plusieurs méthodes peuvent alors être mises en place pour prédire une fréquence. Nous choisissons une approche par Modèles Linéaires Généralisés. Ces modèles sont réputés particulièrement robustes et sont largement utilisés en

assurance car permettent d'expliquer facilement une variable par des critères explicites choisis par l'utilisateur. Ces modèles utilisent notamment les lois des distributions. Nous étudions ainsi les lois de Poisson, Négative Binomiale mais aussi des lois Zéro-Inflaté. De nombreuses méthodes de *Machine Learning* se développent également, nous nous penchons ainsi sur l'une d'entre elles : les forêts aléatoires. De tous ces modèles, nous sélectionnons le plus adéquat avec nos données et le plus robuste, pour obtenir une modélisation de notre fréquence collective.

La dernière étape pour obtenir une fréquence crédibilisée est de déterminer le coefficient de crédibilité. Avant de parler des modèles théoriques qui permettent de déterminer ce coefficient, il est important de se pencher sur le modèle actuel. Le coefficient actuellement mis en place pour déterminer le tarif est une fonction explicite qui dépend de l'exposition au risque. Notre objectif est donc d'obtenir une fonction explicite suite à la mise en place de modèles de crédibilité. Le premier modèle testé est le modèle de Bühlmann-Straub. Ce modèle permet d'individualiser la fréquence en fonction de l'exposition au risque de l'individu en considérant son poids au portefeuille. C'est un modèle relativement simple d'application et robuste, ce qui en fait son attrait. Le second modèle testé est celui de Jewell. Ce modèle part du principe de Bühlmann-Straub mais y intègre l'hétérogénéité du risque au sein du portefeuille. Cette hétérogénéité se mesure en constituant des classes de risque. Les coefficients de crédibilité des individus dépendent alors de la crédibilité sur les différentes classes de risque qui lui sont associées. C'est pourquoi ce modèle est dit hiérarchique. Le modèle finalement choisi entre Bühlmann-Straub et Jewell sera approché pour obtenir une fonction explicite.

Chapitre 1

Contexte de l'étude

1.1 Le marché des flottes automobiles chez MMA

1.1.1 Qu'est-ce qu'une flotte automobile ?

Dans le cadre de ce mémoire, nous nous intéressons à un segment du marché des flottes automobiles.

De manière générale une flotte de véhicules est définie comme l'ensemble des moyens de transports dont dispose une entreprise ou une association. Nous retrouvons notamment : voitures de fonctions, camionnettes, deux-roues, engins de chanter, cars, etc. Chez MMA, une flotte est constituée d'au moins 5 véhicules. Pour information, ces véhicules ne sont pas soumis au bonus-malus mis en place sur le marché des particuliers. Ils sont utilisés dans le cadre de l'activité de l'entreprise et les produits d'assurance ainsi proposés dépendent de l'activité en question : transport privé pour propre compte (TPPC), transport public de marchandises (TRM), transport routier de voyageurs (TRV), location courte ou longue durée (LCD ou LLD). L'assurance des flottes automobiles s'inscrit donc en assurance Non-Vie, en assurance des professionnels et entreprises.

Une flotte peut être caractérisée par plusieurs éléments et notamment des caractéristiques propres à MMA : si elle est composée de véhicules moteurs ou non, la catégorie des véhicules qui la compose, ou encore leur genre par exemple. Ces notions sont définies de la manière suivante :

- Véhicule moteur : le véhicule est dit "moteur" dès lors qu'il peut rouler sur la voie publique. La liste des véhicules pouvant être amenés à circuler sur la voie publique est décrite dans l'Article R311-1 du Code de la Route. Cette distinction est surtout faite sur les engins de chantier : un transpalette n'est pas amené à circuler sur la voie publique par exemple, un tracteur oui.
- Catégorie : les véhicules sont regroupés selon cinq catégories distinctes chez MMA, qui sont les suivantes :

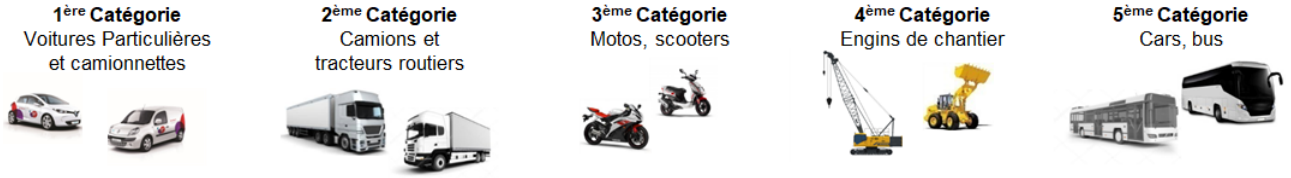

FIGURE 1.1 – Les cinq catégories de véhicules

- Genre : cette notion va dépendre de la catégorie dans laquelle se situe le véhicule, elle permet de préciser le type du véhicule. Pour la catégorie 1, la distinction est faite entre les voitures particulières et les camionnettes par exemple.

Catégorie	Description	Exemples de genres	Véhicule moteur
1	Véhicules à moteur de PTAC (Poids Total Autorisé en Charge) inférieur ou égal à 3.5 tonnes	Voiture particulière	x
2	Véhicules non agricoles de PTAC ou de PTRA (Poids Total Roulant Autorisé) supérieur à 3.5 tonnes	Camion	x
		Tracteur routier	x
3	Véhicules à deux roues, tricycles et quadricycles à moteur, voiturettes et voiturettes électriques, voitures sans permis	Véhicules ≤ 50 cm3	x
4	Véhicules utilisés en tant qu'outil et remorques d'un poids inférieur ou égal à 3.5 tonnes	Matériels de Levage	x
		Appareil Terrestre Attelé	
		Matériel de Jardinage	
		Autres Engins de chantier	x
		Engins de Manutention	
5	Véhicules de plus de 9 places, affectés au transport de personnes	Autocar	x

FIGURE 1.2 – Résumé des notions de catégorie et de genre

Ces flottes sont découpées en ensembles. Un ensemble va regrouper des véhicules possédant les mêmes caractéristiques et mêmes critères tarifaires (catégorie, genre, âge, puissance, etc.), mais aussi les mêmes garanties, clauses et franchises. Les véhicules d'un même ensemble vont donc avoir le même tarif. Les critères tarifaires, et donc de segmentation d'ensemble vont dépendre de la catégorie des véhicules. Par exemple pour la catégorie 1 ces critères peuvent être la tranche d'âge ou le type (le type dépend de la puissance du moteur et de sa valeur), pour la catégorie 2 la tranche de poids total autorisé en charge (PTAC) et pour la catégorie 3 le cylindre.

Pour illustrer la notion d'ensemble, voici l'exemple d'un contrat avec deux ensembles. L'ensemble 1 contient deux véhicules de catégorie 1, de genre voiture particulière (VP), de même type et de même classe d'âge. L'ensemble 2 lui comporte trois véhicules de catégorie 2, de genre tracteur routier (TRR) avec un PTAC $>$ 19T, les mêmes clauses et formules de garanties sont souscrites pour tous les véhicules au sein de l'ensemble.

Véhicule	Catégorie	Entrée	Sortie
Ensemble 1	VP Voiture Particulière / 09E... 2 veh		
Véhicule 1	cat 1	14/06/2019	
Véhicule 2	cat 1	08/02/2021	
Ensemble 2	TRR Tracteur routier / >19.0T 3 veh		
Véhicule 3	cat 2	12/04/2018	20/05/2018
Véhicule 4	cat 2	15/06/2019	
Véhicule 5	cat 2	02/12/2020	
Véhicule 6	cat 2	03/04/2021	

FIGURE 1.3 – Exemple d'ensembles dans un contrat

La gestion des flottes se fait à la maille de l'ensemble et non du véhicule.

Le contrat d'assurance automobile pour le marché des flottes couvre l'entreprise assurée sur les mêmes risques de circulation que le marché des particuliers : responsabilité civile (RC), dommages tout accidents (DTA), bris de glace (BDG), vol, incendie, etc. Des garanties supplémentaires peuvent également être souscrites selon le type d'assurance souhaité.

1.1.2 La segmentation du marché des flottes

Pour des raisons de gestion le marché des flottes est partitionné en cinq segments. Cette segmentation dépend de la taille de la flotte en question et elle est différente selon le produit. Ainsi suivant le segment, la méthode de tarification choisie sera différente.

Il existe différents modes de gestion qui peuvent s'appliquer selon les différents segments :

— Au mouvement : Tous les mouvements d'entrée et de sortie de véhicules sont enregistrés au fur et à mesure et facturés au client. Il y a donc une très bonne connaissance du parc en tout temps. C'est le mode de gestion majoritairement utilisé.

— Périodique : Ici aussi tous les mouvements sont enregistrés au fur et à mesure et la connaissance du parc est donc bonne. En revanche, les mouvements ne sont pas facturés à chaque mouvement. Une régularisation (un bilan) est donc nécessaire. Elle peut être annuelle, semestrielle ou trimestrielle selon le fractionnement choisi.
— Par demi-différence : Toutes les entrées et sorties de véhicules ne sont pas enregistrées au fur et à mesure, mais à chaque fin d'exercice. L'état de parc est annuel. La connaissance du parc est donc partielle puisque si dans la même année un véhicule rentre et sort, ce véhicule ne sera pas connu.

De ces modes de gestion vont se créer les notions d'ensemble désigné et non désigné. Un ensemble désigné est un ensemble dont les informations sur les véhicules qui le composent, ainsi que ses caractéristiques (immatriculation, etc.), sont connues. À l'inverse, dans un ensemble non désigné, ce degré d'information n'est pas connu par l'entreprise, qui ne détient que les informations à la maille de l'ensemble et du nombre de véhicules qui le compose. Les contrats gérés par demi-différence comportent ainsi des ensembles non-désignés, puisque le détail des véhicules n'est pas fourni par l'assuré, seulement leur nombre à chaque état de parc.

Taille de la flotte	Segment	Méthode de tarification	Mode de gestion
Petite et moyenne taille	S0 à S2	Forfaitaire	Mouvement ou Périodique
Taille intermédiaire	S3	Intermédiaire (personnalisation graduelle)	Périodique
Taille importante	S4	Individuelle	Demi-différence

Table 1.1 – Synthèse de la segmentation du marché des flottes

Les segment d'intérêt dans le cadre de ce mémoire est le segment 3, les flottes de taille intermédiaire. Ce segment doit faire le lien entre les segments 2 et 4. Dans la réalité, de grandes différences entre les segments sont identifiées, autant en cotisations qu'en sinistres. Par exemple la prime moyenne par contrat du S3 est 3 fois supérieure à celle du S2 et la prime moyenne par contrat du S4 est 4 fois supérieure à celle du S3. De manière générale, cet ordre de grandeur se retrouve entre les différentes mesures sur ces trois segments. L'idée est donc de lisser ces segments, afin de ne pas avoir une telle disparité.

Le segment 3 est un enjeu particulier sur les flottes automobiles notamment au regard de son taux de concrétisation d'affaires nouvelles. En effet, près de 1 devis sur 3 est concrétisé et rentre au portefeuille. Dans le portefeuille total, le segment 3 a un poids moindre par rapport aux autres segments, il représente en effet environ 9% des véhicules au global soit presque 2% des contrats flottes. Sur l'ensemble des cotisations du portefeuille flottes, le segment 3 en représente près de 10%. Également, les résultats du segment 3 sont légèrement décadrés par rapport au niveau d'équilibre attendu, ce qui nous amène à revoir sa rentabilité et notamment son tarif.

1.2 Le produit d'assurance des flottes automobiles pour propre compte MMA

1.2.1 Présentation du produit

Dans le cadre de cette étude, nous nous intéressons au produit AutoFleet, qui est le produit phare de MMA pour l'assurance en TPPC (Transport Privé pour Propre compte). En assurance TPPC, les véhicules sont utilisés dans le cadre de l'activité de l'entreprise. Les activités concernées sont par exemple les suivantes : artisans et entreprises du bâtiment et des travaux publics, commerçants, industriels, entreprises de service, pompes funèbres, collectivités publiques et associations.

AutoFleet a un poids particulièrement important dans les flottes automobiles : il représente plus de 70% des contrats du portefeuille S3 en cours. Au regard de sa sinistralité, ce produit représente également un enjeu. En effet, AutoFleet c'est 65% des sinistres et 50% de la charge sur le segment 3, poids non négligeable. Dans les mêmes ordres de grandeur, ce produit représente près de 50% des cotisations totales S3. AutoFleet est un produit dont les résultats sont légèrement dégradés par rapport à l'équilibre ; il est donc intéressant de revoir les cotisations sur ce produit.

1.2.2 Les garanties mises en œuvres

Pour l'assurance des flottes en TPPC, les garanties de base sont proposées, comme par exemple :

— Responsabilité civile (RC) : couvre les dommages causés à autrui lors de l'utilisation d'un engin ou un véhicule. Cette garantie est imposée par l'article L 211.1. du Code des Assurances.
— Dommages tous accidents (DTA) : permet l'indemnisation des dommages du véhicule assuré.
— Bris de glace (BDG) : assure les bris, quelle qu'en soit l'origine, du pare-brise, des vitres, des blocs optiques de phares avant, des clignotants.
— Vol et tentative de vol : couvre les dommages matériels du véhicule en cas de détérioration ou disparition du fait d'un vol.
— Incendie : indemnisation des dommages du véhicule ou de sa destruction dû au feu.
— Dommages corporels du conducteur : garantit le conducteur responsable pour les dommages subis par un accident de la circulation et permet de l'indemniser comme une victime non-responsable des conséquences de l'accident (frais médicaux, pertes de salaires, invalidité, frais d'obsèques, etc.).
— Défense pénale et recours suite à accident (DPRSA) : prise en charge par l'assureur des frais de procès lorsque la responsabilité de l'assuré est engagée. L'assureur s'engage également à exercer un recours en cas d'accident non-responsable subi par l'assuré.
— Protection juridique (PJ) : assiste et assure la défense de l'assuré en cas d'infraction au Code de la Route.
— Formules d'assistance et de dépannage : en cas de panne du véhicule, propose des services d'assistance, le remorquage et le dépannage du véhicule.

Mais également, des garanties propres aux activités des entreprises peuvent être souscrites, comme par exemple :
— Responsabilité civile atteintes à l'environnement : couvre les frais de réparation des dommages à l'environnement causés par l'assuré (dommages aux sols, aux eaux, aux espèces ou habitats naturels protégés et à tout tiers concerné). Cette garantie répond à la loi du 1er août 2008 et à son principe de "pollueur-payeur".
— Bagages et effets personnels : couvre les dommages, l'incendie ou le vol des bagages et effets personnels non fixés au véhicule, qui appartiennent aux préposés, à l'entreprise ou aux voyageurs (téléphone, GPS, etc.).
— Marchandises transportées : assure les dommages aux matériels et marchandises transportés dans le véhicule en cas d'accident, de vol ou d'incendie.
— Pertes financières : en cas de perte totale du véhicule assuré, c'est-à-dire si le montant des réparations est supérieur à la valeur du véhicule, cette garantie couvre l'assuré contre tout risque financier lié à l'écart entre les indemnités perçues en cas de sinistre et l'encours financier résiduel sur le véhicule. Elle vient donc en complément des garanties DTA, vol, incendie.

Les garanties pouvant être souscrites sur le produit d'assurance des flottes en TPPC dépendent de la catégorie des véhicules assurés. Au sein d'un même contrat flotte, les véhicules assurés peuvent avoir des couvertures différentes selon les besoins de l'entreprise, mais au sein d'un même ensemble, les garanties souscrites doivent être les mêmes.

Dans un premier temps nous allons nous intéresser aux trois garanties principales : RC, DTA, BDG.

Garantie	S/C	Poids véhicules	Poids cotisations
DTA	90%	60%	27%
RC	103%	95%	36%
BdG	80%	50%	8%

TABLE 1.2 – Répartition des garanties principales sur le portefeuille

Le ratio S/C, pour $\frac{\text{sinistres}}{\text{cotisations}}$, est une mesure de rentabilité d'un périmètre. À l'équilibre, le ratio est égal à 1. Cet indicateur est à comparer avec le S/C d'équilibre, qui prend en compte les frais généraux, les taux de commissions et la réassurance. Le S/C d'équilibre est le S/C qui permet d'atteindre un ratio combiné de 1, c'est à dire la rentabilité en prenant en compte les frais liés à l'activité d'assurance. Le ratio combiné s'écrit de la manière suivante : $\frac{\text{sinistres + commissions + frais généraux + réassurance}}{\text{cotisations}}$.

La garantie Responsabilité Civile demande une vigilance particulière car présente sur la quasi-totalité des véhicules AutoFleet en S3 et concerne à elle seule 20% des sinistres. Sinistres qui sont généralement assez coûteux. La garantie RC représente 35% de la charge sinistre totale sur le produit AutoFleet en S3.

La garantie Dommages Tous Accidents est également importante pour les mêmes raisons : son poids au portefeuille, et sa sinistralité. À moindre mesure toutefois, la garantie DTA étant présente sur les 2/3 des véhicules AutoFleet en S3. Pour ce qui est de la sinistralité, nous sommes sur les mêmes ordres de grandeur, à savoir 20% et 35% des sinistres en nombre et en charges sur le portefeuille AutoFleet S3. Aussi, les coûts moyens par sinistre

et par véhicule sont particulièrement élevés en DTA, avoisinant ceux en RC.

Au regard des indicateurs de rentabilité, à savoir le ratio S/C, la garantie DTA est plus dégradée que la RC par rapport au reste du périmètre. Ces deux garanties étant dégradées par rapport à leur S/C d'équilibre.

Une attention particulière est à apporter à la garantie Bris de Glace. Le coût moyen d'un sinistre BdG est faible, cependant c'est la garantie qui se trouve être la plus dégradée en considérant son écart entre le S/C et le S/C d'équilibre.

1.3 Les flottes MMA en chiffres

Dans cette partie, nous allons nous intéresser à la structure du portefeuille flottes MMA et à sa sinistralité. Les données présentées ci-dessous sont issues des bases de données MMA et des tableaux de bords de suivi de l'activité du marché. Sauf précision les chiffres présentés dans cette partie sont à vision 2021.

1.3.1 Le portefeuille flotte

Covéa est leader sur le marché de l'assurance en France et plus particulièrement en assurance des flottes automobile où elle détient 16,3% des parts de marché. Elle se positionne n°2 sur ce marché (source : FFA, Fédération Française de l'Assurance). Près de la moitié du chiffre d'affaire de MMA se fait sur le marché des entreprises. Ainsi, sur les 2.3 Milliards d'euros de chiffre d'affaire réalisés sur ce marché, 463 Millions se font sur le marché des flottes, soit près de 20%, ce qui en fait une part non négligeable.

Le portefeuille flotte MMA c'est près de 900 000 véhicules assurés, soit une croissance de 8.4% par rapport à 2020, et une prime moyenne par véhicule de 500€. Ces véhicules se répartissent en 64 000 contrats, ce qui représente une croissance de 1.7% depuis 2020, avec une prime moyenne par contrat de 7 000€. Une hausse de la prime moyenne par contrat de l'ordre de 7.5% peut également s'observer.

Depuis les données de nos outils et rapports internes ainsi que des données de France Assureurs (anciennement FFA, Fédération Française de l'Assurance), nous pouvons observer l'évolution du portefeuille sur le marché des flottes MMA.

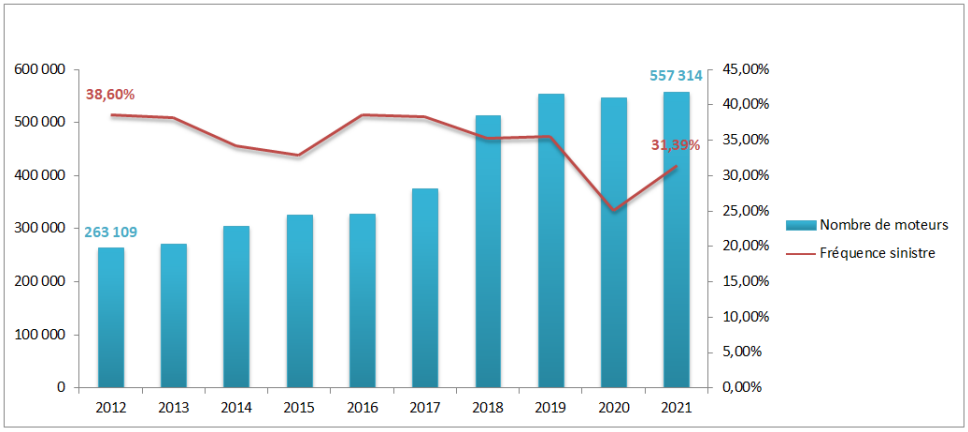

FIGURE 1.4 – Évolution du portefeuille flottes MMA, fréquence et exposition

Nous pouvons remarquer une forte croissance du portefeuille de +112% entre 2012 et 2021. Cette croissance est motivée par une croissance générale du marché des entreprises de l'ordre de +120%, mais également par des politiques de développement de MMA, notamment entre 2016 et 2019. En parallèle une baisse de la fréquence significative de l'ordre de -19.1% est observée, malgré la forte hausse de fréquence entre 2015 et 2016. L'année 2020 créé un biais sur la courbe de la fréquence, du fait de la crise sanitaire. Cependant en parallèle de cette baisse de la fréquence sinistre, une importante hausse des coûts est observée. Le graphique ci-dessous, issu des statistiques SRA, l'illustre bien :

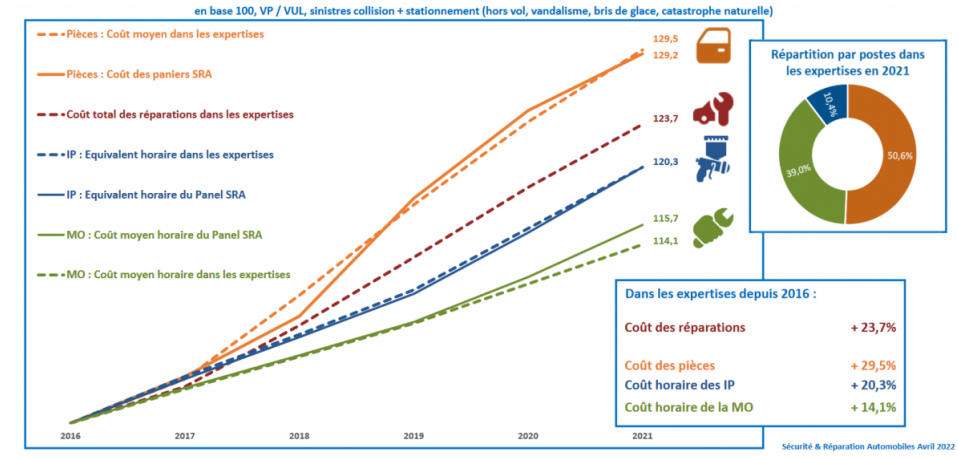

FIGURE 1.5 – Évolution des coûts depuis 2016

La structure du portefeuille ainsi que celle du marché ont bien évoluées ces 10 dernières années. Cette évolution motive donc la révision du tarif actuel, qui a été mis en place en 2009, il y a 13 ans.

1.3.2 Le portefeuille TPPC et le segment 3

Le produit TPPC est le produit principal en assurance des flottes chez MMA. En effet, il représente près de 80% des véhicules assurés en flottes. Il est donc l'acteur majoritaire des évolutions et de la structure du portefeuille. Avec notamment une évolution croissante du chiffre d'affaire du portefeuille, freinée par la crise Covid-19 sur 2019-2020.

Dans le portefeuille TPPC, le segment 3 a une part relativement faible par rapport aux autres segments, en effet il compte 8% du portefeuille en nombre de véhicules, soit près de 63 000 véhicules, et 860 contrats. Ses cotisations s'élèvent à 36 Millions d'euros.

Nous pouvons regarder la répartition de notre parc selon différents critères, comme la zone géographique d'exercice de l'entreprise assurée, son activité, ou la catégorie des véhicules assurés. Ces critères seront étudiés plus en détail dans la suite.

Zone	Part en nombre de véhicules (%)	Part en cotisations (%)
1	21	20
2	16	14
3	37	36
4	10	11
5	14	17
6	2	2

TABLE 1.3 – Répartition du portefeuille TPPC S3 selon les zones

Le premier constat ici est que certaines zones sont sous-représentées par rapport à d'autres. Elles feront donc l'objet d'une attention particulière dans la suite, c'est le cas de la zone 6 notamment. La zone 3 quant à elle représente plus du tiers du portefeuille, elle est sur-représentée.

Activité	Part en nombre de véhicules (%)	Part en cotisations (%)
1	4	4
2	1	1
3	28	28
4	1	1
5	30	30
6	7	5
7	9	10
8	20	21

TABLE 1.4 – Répartition du portefeuille TPPC S3 selon l'activité

Ici encore une répartition hétérogène du portefeuille selon l'activité est à noter et devra être prise en compte lors des étapes de modélisation.

Catégorie	Part en nombre de véhicules (%)	Part en cotisations (%)
1	58	67
2	11	15
3	4	1
4	27	17
5	0.1	0.2

TABLE 1.5 – Répartition du portefeuille TPPC S3 selon la catégorie

Pour rappel, le détail des catégories peut se trouver dans ce tableau. À première vue, le portefeuille est composé majoritairement de véhicules de catégorie 1 (voitures particulières, camionnettes) et de catégorie 4 (engins de chantier). À l'inverse, les véhicules de catégorie 3 (motos, scooters) et 5 (cars, bus) sont en très faible proportion au portefeuille.

1.4 La tarification actuelle du segment concerné : principe de la crédibilité

Dans cette section, l'objectif est de présenter la méthode de tarification utilisée actuellement, telle qu'elle a été mise en place en 2009. Aujourd'hui, les devis et tarifs des contrats sur le segment 3 sont mis en place depuis ce qui est appelé le "tarificateur S3". Ce tarificateur est un outil spécifique à ce segment, mis en place sous Excel et utilisé par les équipes de souscriptions afin de tarifer les affaires nouvelles et de mettre en place des devis.

La méthode de tarification actuelle peut donc être retrouvée dans le code VBA du tarificateur et dans une note explicative. Des formules explicites et des coefficients exposés en dur peuvent notamment y être retrouvés, mais il n'y a pas d'information concernant leur origine. En effet, la théorie ou les méthodes appliquées qui ont mené au choix définitif de ces formules explicites ne sont pas exposées et restent alors inconnues aujourd'hui, malgré une première étape de recherche méthodologique dans la littérature.

1.4.1 Méthodologie

La tarification du segment 3 doit faire le lien entre celle des segments 0 à 2, forfaitaire et celle du segment 4, individualisée. L'objectif est donc de personnaliser progressivement la prime, personnalisation plus forte que l'approche pour les flottes de petite taille mais plus faible que celle pour les flottes plus importantes.

Pour ce faire, le principe de tarif crédibilisé est utilisé, pour les garanties DTA, RC et BDG. La prime pure est calculée au niveau du véhicule pour chaque garantie de la manière suivante :

$$PP_{veh} = CM^{th}_{veh} \times Freq^{cred}_{contrat}$$

Avec :
- PP_{veh} : Prime Pure du véhicule
- CM^{th}_{veh} : Coût Moyen théorique d'un véhicule
- $Freq^{cred}_{contrat}$: Fréquence crédibilisée de la flotte

La prime pure peut se réécrire sous la forme :

$$PP_{veh} = CM^{th}_{veh} \times Freq^{th}_{contrat} \times \underbrace{\frac{Freq^{cred}_{contrat}}{Freq^{th}_{contrat}}}_{CPT}$$

Avec :
— $Freq^{th}_{contrat}$: Fréquence théorique de la flotte
— CPT : Coefficient de Personnalisation Tarifaire

Cette forme permet de faire apparaître le CPT et la prime pure théorique ($PP^{th} = Freq^{th} \times CM^{th}$).

Nous pouvons ensuite calculer la prime pure d'un ensemble (PP_{ens}) de la manière suivante :

$$PP_{ens} = n \times PP_{veh}$$

Avec :
— n : nombre de véhicules dans l'ensemble

Et enfin la prime pure totale du contrat ($PP_{contrat}$) :

$$PP_{contrat} = \sum PP_{ens}$$

L'enchaînement des différentes mailles est donc décruit comme suit :

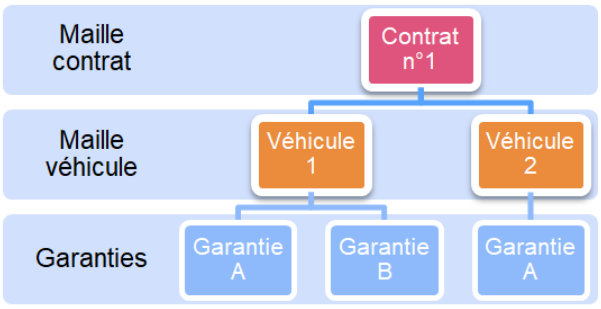

FIGURE 1.6 – Description de l'emboîtement des mailles

Pour le calcul de la prime pure, nous avons donc besoin du coût moyen théorique, calculé à la maille véhicule, et de la fréquence crédibilisée, à la maille contrat, qui se calcule de la manière suivante :

$$Freq^{cred}_{contrat} = Freq^{obs}_{contrat} \times z + Freq^{th}_{contrat} \times (1 - z)$$

Nous retrouvons là la formule de la théorie de la crédibilité avec $z \in [0; 1]$ le coefficient de crédibilité.
— $z = 0$: pas de personnalisation tarifaire, $Freq^{cred}_{contrat} = Freq^{th}_{contrat}$
— $z = 1$: tarif complètement individualisé, $Freq^{cred}_{contrat} = Freq^{obs}_{contrat}$

Il existe différentes méthodes de calcul du coefficient z, nous pensons notamment à la méthode classique de Bühlmann-Straub ou à la crédibilité hiérarchique de Jensen par exemple. Dans le tarif implémenté, le coefficient z a une expression explicite différente selon la garantie souscrite :

$$z_{DTA} = \sqrt{\frac{n_{DTA} - c}{d - c}}$$

Avec :
— n_{DTA} : nombre de moteurs (année N) de la flotte
— c et d : les nombres de moteurs qui définissent les bornes du segment 3

Les fréquences théoriques sont calculées pour chaque année en fonction de leur taux d'exposition. L'exposition étant le nombre de moteurs facteur de leur taux de présence sur l'année, appelés dans la suite "taux de moteurs année".

$$Freq^{th} = \frac{Nombre\ de\ sinistres}{Exposition}$$

Par exemple pour un véhicule moteur présent de janvier à septembre de l'année considérée, son exposition sera de $\frac{9\text{mois}}{12}$. À la maille d'un ensemble de trois véhicules moteurs cela donne :

Véhicule	Exposition
1	$\frac{1}{12}$
2	1
3	$\frac{5}{12}$
Total ensemble (Σ)	1.5

L'exposition de cet ensemble ne sera donc pas de 3 mais de 1.5.

Les fréquences théoriques sont modélisées à la maille du contrat selon le croisement zone x activité.

Le coût moyen théorique est modélisé à la maille véhicule avec des critères tarifaires qui varient selon la catégorie du véhicule :

- Catégorie 1 : Coût moyen par groupe/classe (notion propre à MMA de catégorisation des véhicules)
- Catégorie 2 : Coût moyen par tranche de PTAC (Poids Total Autorisé en Charge)
- Catégorie 3 : Coût moyen par cylindrée
- Catégorie 4 : Coût moyen par genre

1.4.2 Les variables utilisées

Nous distinguons la maille contrat, utilisée pour la fréquence, et la maille véhicule, pour le coût moyen.

À la maille contrat, les deux variables qui vont nous intéresser sont :

- le type d'activité de l'entreprise
- la zone géographique d'activité de l'entreprise

Les activités sont réparties en 8 catégories sur les flottes automobiles. Elles ont été introduites dans le tableau 1.4. À noter également, parmi les activités au portefeuille, celles des ambulances et des taxis peuvent y être retrouvées, en plus des 8 étudiées. Cependant, les taxis sont fermés à la souscription depuis 2019 et ne seront donc pas pris en compte pour modéliser le tarif sur notre périmètre. Aussi, les ambulances ne sont pas un marché cible pour MMA et la politique de souscription est donc très sélective ; la souscription ne s'effectue pas à partir du tarificateur classique, cette activité n'est donc pas prise en compte dans le cadre de ce mémoire. Leur poids au portefeuille est également négligeable : ces deux activités représentent conjointement moins de 1% du portefeuille.

Pour ce qui est des zones, le zonier utilisé en compte 6. Les zones sont construites de manière à ce que leur risque soit croissant ; la zone 1 est moins risquée que la zone 2, qui est moins risquée que la 3, etc. Ce zonier ainsi créé date de 2012. Dans le cadre de cette étude, nous ne le remettons pas en cause, une refonte du zonier pourrait être un sujet traité à part. De plus, d'après les études menées, la pertinence du zonier semble toujours d'actualité malgré la problématique d'hétérogénéité de la répartition du portefeuille dans ces zones.

Comme nous avons pu le voir dans la partie précédente, le coût moyen théorique prend en compte des variables tarifaires différentes selon la catégorie du véhicule. Les notions de groupe/classe, de PTAC, de cylindrée et de genre vont donc nous intéresser en particulier.

Dans le cadre d'une refonte tarifaire complète, l'exhaustivité de cette liste de variables peut être remise en cause.

1.4.3 Problématiques soulevées

La particularité principale des flottes automobiles, contrairement à l'assurance auto des particuliers, outre les garanties spécifiques à l'activité de l'entreprise, est que nous ne disposons pas des informations relatives au conducteur. Ces informations sont pourtant centrales dans la tarification d'un contrat automobile pour particuliers. Parfois, nous ne disposons pas non plus des informations relatives aux véhicules assurés. Il va donc être

nécessaire de contourner ce manque d'informations, notamment en prenant en compte l'expérience de la flotte avec la théorie de la crédibilité par exemple.

À ce jour, le tarif en sortie du tarificateur S3 AutoFleet semble décadré. En effet, en consultant le pôle de souscription, les souscripteurs disent déroger au tarif sur toutes leurs affaires nouvelles à hauteur de 40 à 60 % du tarif en moyenne. Pour déroger ce tarif ils utilisent le tarificateur du segment 4, qui est adapté aux flottes de taille plus importante, pour obtenir un tarif. Le tarificateur S4 a notamment la particularité de prendre en compte l'expérience du coût moyen de la flotte et non seulement de la fréquence. Cette démarche n'est pas optimale car très chronophage pour les souscripteurs et du fait que le tarificateur S4 n'est pas adapté pour les plus petites flottes. Le tarif S4 est calculé sur mesure sur la seule base des antécédents de charges sinistre de l'assuré et sur l'expertise des souscripteurs.

Aussi, nous pouvons remettre en cause la pérennité des fréquences théoriques et autres coefficients utilisés dans le tarificateur S3. En effet comme nous avons pu le voir, la structure du portefeuille et sa sinistralité ont bien évolué depuis la construction du tarificateur il y a 13 ans. L'écart de tarif entre la sortie du tarificateur S3 et l'expertise des souscripteurs couplée avec l'étude du marché peut donc s'expliquer en partie par ces coefficients non à jour.

Pour la refonte du tarif sur le segment 3, nous souhaitons nous baser sur les travaux menés lors de la construction du tarificateur. Cependant nous sommes confrontés à un manque de documentation sur le sujet. En effet, comme cela a pu être évoqué dans la partie précédente, nous n'avons pas le détail ni la démarche derrière les formules utilisées dans le tarificateur, ce qui pose une problématique de justification de la méthode qu'il est souhaitable de revoir.

1.5 Les travaux déjà menées sur le sujet et leurs limites

Le sujet de la révision du tarif sur le segment 3 du produit TPPC a été abordé dans un premier temps avec une idée de refonte complète du tarif à la maille véhicule. Cette problématique a été étudiée dans le cadre de ma première alternance et sur le début de mon alternance actuelle. Les travaux mentionnés dans cette partie sont détaillés dans le rapport d'alternance consacré entre autres à ce sujet [16].

1.5.1 Construction d'une première base de données

Dans un premier temps, nous cherchons à déterminer une prime pure en fonction de différents critères qui se trouvent au niveau véhicule. Nous devons donc construire une base de données d'étude de la fréquence et du coût moyen et nous aurons besoin conjointement des informations sinistres (nombre, coût, garanties sinistrées, etc.) et des critères tarifaires sur notre périmètre (âge, genre, puissance du véhicule par exemple).

La difficulté de mise en place d'une telle base est qu'il n'en existe pas à ce jour dans les bases de données de l'entreprise. Nous avons cependant à notre disposition deux bases distinctes : une base contenant les informations sinistres, une base contenant les informations portefeuille. L'idée ici est donc de joindre ces deux bases à la maille du véhicule, donc sur le croisement de données contrat-ensemble-véhicule. Le contrat est identifié par un numéro de contrat unique défini à la souscription, l'ensemble par un chiffre et le véhicule par son numéro d'immatriculation. Cependant, cette jointure n'est pas évidente. En effet, les données de jointure peuvent être mal renseignées ; nous allons par exemple avoir des données manquantes, des erreurs de saisie, des problèmes de format, etc. Certaines données manquantes s'expliquent par le mode de gestion particulier des véhicules gérés en demi-différence (les informations des véhicules au sein de l'ensemble ne sont pas disponibles, se référer à la partie 1.1.2). Aussi, et c'est le problème principal auquel nous sommes confrontés, les différentes bases de données à notre disposition sont construites à partir de données saisies manuellement. Les gestionnaires ne sont pas dans l'obligation de saisir tous les éléments pour l'enregistrement d'un contrat ou d'un sinistre et les saisies ne sont pas automatisées, ce qui laisse place à des erreurs. Il va donc falloir effectuer des rapprochements en prenant en compte ces éléments. Toutefois, cette problématique de données erronées ne se pose pas sur la maille du contrat. En effet, à cette maille les saisies sont automatiques et les problématiques de jointures ne se posent donc plus. Cela permet de construire une base de vérification afin de contrôler notamment le nombre et la charge de sinistres au total sur le périmètre.

Dans un premier temps, différentes techniques ont été mises en place pour uniformiser les formats, des numéros de contrats et des immatriculations notamment. Aussi, les données de la base contenant les informations portefeuilles étant codifiées, certaines ont été décodifées pour permettre une meilleure lecture des informations. De là nous pouvons commencer à joindre les sinistres au portefeuille à la maille du véhicule. Cette étape se fait en deux fois : d'abord sur les ensembles désignés (sur l'immatriculation), ensuite sur les ensembles non-désignées

(sur le numéro d'ensemble). Sur ces premières jointures, seulement 61% des sinistres sont raccordés en nombre et en charge, c'est-à-dire que 61% des sinistres du périmètre seulement se trouvent dans la base de données. Pour raccorder les 39% restants, sont mises en places différentes méthodes : en joignant à la maille supérieure, en s'assurant de la cohérence avec la période de garantie des véhicules et de la date de survenance sinistre, ou en cas de non cohérence, en raccordant à la date de situation la plus proche. Après ces raccordements nous vérifions la cohérence de la base avec la base de vérification. Par exemple, la sinistralité totale sur la base construite est comparée avec celle de la base de vérifications, les sinistres les plus importants sont identifiés pour s'assurer qu'ils apparaissent bien dans la base construite, ou encore une recherche de doublons est effectuée pour vérifier que des doubles n'aient pas été créés lors des jointures. La base finale contient ainsi 89% des sinistres en nombre et en charge. Les sinistres qui restent non-raccordés sont notamment ceux dont les véhicules renseignés ne sont pas sur le bon contrat et dont le contrat auquel il appartient ne peut être identifié.

1.5.2 Premières modélisations

En amont des modélisations, toute une étape de préparation et d'étude des données a été effectuée. Lors de cette étape, nous sommes à nouveau confrontés à la problématique des données manquantes, sur des données considérées comme des critères tarifaires essentiels car utilisés pour former des ensembles lors de la tarification (par exemple l'âge, ou le genre du véhicule). Ici encore des travaux ont été effectués afin de retrouver un maximum de données. Nous pouvons citer par exemple l'utilisation de la base SIV (Système d'Immatriculation des Véhicules), qui est une base gouvernementale qui lie les immatriculations aux véhicules et aux certificats d'immatriculation. Également des croisements de variables ont été effectués pour en retrouver d'autres. Suite à ces divers retraitements près de 80% de données restent manquantes sur certains critères tarifaires. De plus, nous pouvons aussi nous questionner sur la fiabilité des données "correctement renseignées", en effet, certains champs sont renseignés avec des informations par défaut par les gestionnaires.

Pour tenter de palier à ces données manquantes, des modèles de *Machine Learning* ont été implémentés afin de capter ces informations. Des modèles de Random Forest, Gradient Boosting ont été *tunés* et implémentés. Cependant, les résultats en sortie de modèles ne sont pas concluants, comme nous pouvions nous y attendre du fait que la base de données n'est pas suffisamment fiable.

Du fait des problèmes de fiabilité de données à la maille du véhicule, la problématique de la refonte tarifaire doit être revisitée. C'est pourquoi les travaux présentés dans ce mémoire s'orientent maintenant vers une révision du tarif à la maille contrat. Plutôt que de revoir le tarif dans son ensemble, nous optons donc plutôt pour une révision des coefficients utilisés. Ainsi, la maille contrat est suffisante. En effet, nous savons que les informations de niveau contrat sont bien renseignées et ne poseront donc pas de problème de fiabilité. Ce nouveau choix de maille nous a également poussé à reconsidérer notre approche de la refonte tarifaire.

Chapitre 2

Mise en place d'une base de données d'étude

2.1 Format de la base

Avant de pouvoir mettre en place une base de données quelconque, il est nécessaire de définir son objectifs ainsi que les informations qui doivent en sortir, où trouver ces informations, sous quelle forme les présenter. Aussi différentes hypothèses sont adoptées lors de la création de la base de données, du fait des contraintes métiers ou de la forme des données brutes.

Dans un premier temps nous nous consacrerons à l'étude de la fréquence, mais nous mettons aussi en place la base de données d'étude pour le coût moyen.

2.1.1 Idée de construction de la base

Nous souhaitons, avec cette base de données, étudier la fréquence et le coût moyen. Cette étude se fait à la maille du contrat, du fait des problématiques évoquées dans la partie 1.5. L'objectif est ainsi de joindre les critères tarifaires de niveau contrat avec les informations sinistres correspondantes, avec en orange la représentation des lignes dans la base de données.

Pour cette étude, la base de données souhaitée contient les informations relatives au contrat en ligne, avec une ligne par contrat, et les informations liées aux sinistres en colonne par garantie, par année. Ainsi l'historique de sinistralité sur chacune des garanties est retranscrit dans la base de données et il est aisé d'agréger les colonnes par garantie ou par année par la suite. Le schéma suivant illustre le format souhaité de la base.

FIGURE 2.1 – Format de la base de données

Pour mettre en place cette base, les informations portefeuilles nécessaires à sa création sont les suivantes :

- Numéro de contrat
- Segment de gestion
- Années de situation
- Exposition au risque : le nombre de véhicules moteurs dans la flotte pour le risque (la garantie)
- Garanties souscrites
- Critères tarifaires : zone et activité
- Cotisations sur les différentes garanties

Et les informations sinistres :

- Numéro de sinistre
- Année de survenance sinistre
- Nombre de sinistres
- Garantie sinistrée
- Responsabilité de l'assuré
- Coût sinistre : charge sinistre, montant de recours

Certaines de ces données peuvent être directement récupérées depuis les bases de données dont à disposition. D'autres devront faire l'objet de retraitements, évoqués dans les parties suivantes. L'implémentation de la base de données se fait uniquement sous SAS et se décompose en 3 étapes :

1. Extraction du portefeuille
2. Extraction des sinistres
3. Jointure portefeuille - sinistres

La suite de ce chapitre est donc dédié à la mise en place des bases de données de la forme décrite précédemment.

2.1.2 Les bases à disposition

Deux des bases de données MMA à disposition sont utilisées : l'une contenant les informations portefeuille, l'autre contenant les informations des sinistres.

La base de données portefeuille recense les différentes informations contrat, client, véhicules, de l'ensemble des contrats ayant souscrit le produit AutoFleet. Cette base est à la maille de la situation par véhicule. En d'autres termes, nous avons une nouvelle ligne pour toute nouvelle situation, tout avenant, concernant un contrat. Ces situations sont recalculées pour prendre en compte les années civiles et les dates d'entrée et de sortie d'un véhicule. Le schéma suivant résume la granularité de la base portefeuille.

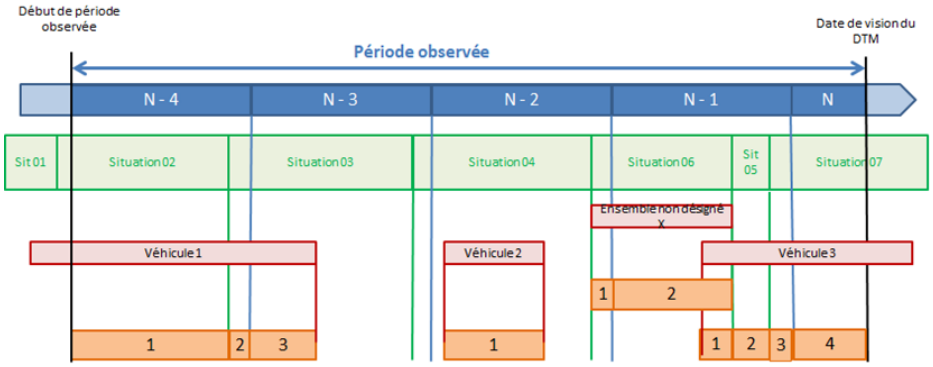

FIGURE 2.2 – Principe de construction de la base portefeuille

La base de données sinistres contient, pour l'ensemble du portefeuille MMA, les informations relatives à un sinistre : garantie engagée, assuré ayant subit le sinistre, responsabilité, charges, etc. Cette base est simplement construite avec la granularité d'une ligne par garantie sinistrée. Si un sinistre X engage à la fois la garantie Responsabilité Civile et Dommages Tous Accidents par exemple, 2 lignes seront consacrées à ce sinistre dans la base.

Nous verrons donc dans la suite comment gérer ces tables pour construire notre base de données d'étude.

2.1.3 Le périmètre

Afin de sélectionner notre périmètre d'étude, les bases sont filtrées sur :

- le produit : AutoFleet
- la période : janvier 2017 - décembre 2021 (hors 2020)
- les segments : S2 et S3

— les garanties : RC, BdG, DTA
— les individus : les assurés ayant un contrat sur le périmètre

Le choix du produit AutoFleet est naturel du fait de l'enjeu qu'il représente pour le marché des flottes automobiles MMA, comme nous avons pu le voir dans la partie 1.2 notamment.

Pour ce qui est de la profondeur de la période, nous décidons de nous arrêter à décembre 2021 afin de n'avoir que des années complètes pour faciliter les traitements. Nous remontons jusqu'à 2017 pour avoir un historique de 4 ans, qui correspond à l'historique des relevés d'informations de la sinistralité passée. L'historique considéré est de 4 ans et non 5 (entre 2017 et 2021), puisque l'année 2020 est entièrement exclue de l'étude. En effet, pour cause de la crise sanitaire, cette année est très atypique en sinistralité et viendrait biaiser nos résultats. Des études ont été menées sur l'impact de la crise covid sur la fréquence sinistre selon les différentes garanties (le coût moyen d'un sinistre reste relativement semblable vis à vis de la crise sanitaire). L'idée derrière ces études était d'éventuellement conserver les mois de l'année les moins touchés et qui auraient pu être considérés comme "normaux". Nous pouvons par exemple évaluer les écarts de fréquence entre la sinistralité relevée en 2020 et la tendance moyenne des années 2019 et 2021.

	BdG	DTA	RC	
janv-20	15,4%	34,1%	4,2%	
févr-20	18,5%	42,8%	6,0%	
mars-20	-35,2%	-24,6%	-48,3%	
avr-20	-60,3%	-59,0%	-69,3%	Confinement 1
mai-20	-7,8%	-12,8%	-37,4%	
juin-20	17,1%	3,9%	-17,0%	
juil-20	14,8%	22,3%	-9,8%	
août-20	22,7%	45,2%	10,8%	
sept-20	21,9%	39,8%	4,1%	
oct-20	7,8%	31,5%	-9,4%	
nov-20	8,6%	-4,6%	-28,5%	Confinement 2
déc-20	10,7%	16,6%	-12,1%	

FIGURE 2.3 – Ecart de la fréquence sinistre par rapport à la tendance attendue

Mais finalement l'ensemble de l'année est largement impacté et pour simplifier la construction de la base de données le choix a été de simplement écarter l'année 2020 de notre base.

Nous choisissons naturellement le segment 3 dans notre périmètre d'étude. Mais nous nous intéressons également au segment 2. En effet la sinistralité sur ce segment est assez similaire à celle du segment 3. Les différences entre les segments que nous exploitons portent sur la volumétrie et la qualité des données. En effet, le segment 2 est bien plus important en volume que le segment 3, ce qui nous permet de gonfler notre base de données d'étude pour fiabiliser nos modèles. Aussi, les données sur le segment 2 sont renseignées automatiquement lors de la création d'un contrat et donc plus fiables.

Les garanties qui nous intéressent dans le cadre de ce mémoire sont les garanties Responsabilité Civile, Dommages Tous Accidents et Bris de Glace. Comme nous avons pu le voir, ces trois garanties sont celles qui pèsent le plus au portefeuille, en nombre de véhicules et en cotisations et représentent donc un enjeu majeur. L'ensemble des garanties souscriptibles en S3 seraient à revoir, mais ne sont pas l'objet de ce mémoire.

L'ensemble des contrats sur le périmètre sont considérés, qu'ils soient en cours à la date de vision ou résiliés. La sinistralité et la structure du portefeuille étant relativement similaires entre 2017 et 2021, nous pouvons également considérer les contrats résiliés jusqu'en 2017.

2.1.4 Les hypothèses sur les variables

Avant de construire la base de données, un certain nombre d'hypothèses doivent être mises en place. Elles proviennent des modes de gestion actuels du segment 3. Cette gestion peut être remise en cause, mais dans une première intention de mise à jour du tarif, nous conservons ces hypothèses.

Concernant les sinistres, la gestion des sinistres atypiques et attritionnels est la suivante :
— Gestion par écrêtement
— Seuils de grave définis par garantie

La notion d'écrêtement est telle que lorsque le coût du sinistre dépasse un certain seuil, seule la charge sinistre supérieure au seuil est comptabilisée comme grave. Par exemple avec un seuil de 10 000€, un sinistre de 12 000€ sera considéré comme grave. Mais nous comptabiliserons en charge hors graves 10 000€, et en charge grave les 2 000€ restant supérieurs au seuil. Cette méthode de gestion permet de mutualiser la sinistralité grave. Pour ce qui est des seuils de grave, aussi appelés seuils d'écrêtement dans ce cas, ils sont définis dans nos outils suite à des études internes qui ont été effectuées. Dans le cadre de ce mémoire, nous vérifierons la cohérence de ces seuils dans la partie 2.3.3.

La charge sinistre est gérée ainsi :
- Charge nette de recours évalué
- Sauf sinistres convention IDA : Charge brute de recours évalué

Nous nous intéressons effectivement au coût pour l'entreprise du sinistre, et non au coût total du dit sinistre. C'est pourquoi nous considérons les charges sinistres auxquelles sont soustraits les recours. Exception faite toutefois des sinistres issus de la convention IDA (Indemnisation directe de l'assuré).

Sont concernés par cette convention les accidents de la circulation en France, impliquant au moins deux véhicules assurés auprès de compagnies adhérentes à la convention. Seuls les sinistres matériels sont concernés, pas les dommages corporels. Le principe de base est le suivant : à la survenance d'un tel sinistre, les sociétés adhérentes se doivent d'indemniser leurs assurés respectifs avant même de mettre en place des recours. Ainsi si un assuré est victime d'un accident de la circulation et qu'il n'est pas responsable, son assurance lui versera l'indemnité à laquelle il a le droit pour les dommages matériels subis. Ensuite seulement, l'assureur pourra faire un recours à l'assurance du responsable de l'accident en question. Cet accord entre les compagnies d'assurances a été mis en place pour permettre d'accélérer l'indemnisation des assurés. Le règlement des sinistres se base sur un barème qui regroupe les différentes situations possibles et est donc forfaitaire. Tous les ans, la convention détermine le nouveau montant des forfaits de recours. Si le taux de responsabilité est de 100%, les montants de recours seront appliqués tels quels. Si le taux est de 50%, le montant du recours est de 50% $\times$ forfait (ils sont appelés des demi-IDA). C'est du fait de cette particularité d'indemnisation et de leur gestion dans nos bases de données que les sinistres IDA ne peuvent être considérés nets de recours, auquel cas nous aurions des charges négatives ou nulles.

Également nous allons comptabiliser les sinistres selon la responsabilité de l'assuré comme suit :

Garantie	Fréquence	Coût
DTA	Responsables uniquement	Tous
RC Matériel	Responsables uniquement	Considéré en DTA
RC Corporel	Responsables uniquement	Considéré en RC
Sans suite	Tous	Non

TABLE 2.1 – Gestion de la responsabilité

Nous ne comptabilisons en fréquence que les sinistres responsables. Ainsi n'est imputé à l'assuré que les sinistres dont il est à l'origine. Pour l'étude du coût, tous les sinistres sont comptabilisés, qu'ils soient responsables ou non. Puisque la charge sinistre considérée est la charge nette de recours évalué, nous prenons donc en compte l'ensemble des coûts payés par l'entreprise au bénéfice de son assuré. Mais pour la garantie RC, il y a distinction entre la RC Corporelle et la RC Matérielle, qui sont alors considérées respectivement en DTA et en RC. En effet, les coûts liés à la RC Corporelle peuvent être particulièrement élevés et sont moins fréquents que les coûts liés à la RC Matérielle qui sont plus proches des sinistres DTA. Les sinistres sans suite sont bien considérés en fréquence, mais pas en coût. En effet, un sinistre sans suite a bien eu lieu, même s'il n'a pas de charge pour l'entreprise d'assurance.

Nous devons également faire une hypothèse sur le portefeuille : le contrat est rattaché sur l'année au segment auquel il est rattaché au 31/12 de l'année considérée

En d'autres termes, nous devons forcer l'affiliation d'un contrat à un segment sur l'ensemble de l'année. En effet, un contrat peut passer d'un segment à l'autre en cours d'année selon les variations de taille de sa flotte. Seulement, la base de données portefeuille dont nous disposons, décrite dans la partie 2.1.2, comptabilise tous les changements de situation d'un contrat. Ainsi, si nous récupérions uniquement les lignes indexées sur notre périmètre (segment = "S3" ou segment="S2") la partie de l'année hors périmètre ne serait pas prise en compte, l'exposition du contrat sur cette partie non plus donc. Un biais dans le calcul de la fréquence serait donc induit. Par exemple, prenons un contrat présent sur notre périmètre du 1/03 au 31/06 de l'année N, mais passant dans un autre segment du 1/07 jusqu'au 31/12. Son exposition sur l'année est de $\frac{10}{12}$ (10 mois sur l'année), mais elle

est seulement de $\frac{4}{12}$ (4 mois) sur notre périmètre. En parallèle, nous ne pouvons pas faire la distinction entre les segments sur les sinistres et récupérons donc la sinistralité sur l'année complète. Nous surestimons donc la fréquence en ne considérant que les lignes présentes sur le périmètre car : $\frac{\text{n sinistres}}{10/12} < \frac{\text{n sinistres}}{4/12}$. C'est pourquoi pour considérer toute l'exposition du contrat, nous forçons le même segment sur l'ensemble de l'année, et prenons en compte le segment d'affiliation au 31 décembre, soit la dernière situation de l'année. À noter que les contrats ne sont pas amenés à changer de segment régulièrement, les segments de gestion sont construits de sorte à éviter qu'un assuré ait à changer de segment et donc d'interlocuteur trop régulièrement.

Enfin, comme nous avons pu le voir dans la partie 1.4.1, nous prenons l'exposition au risque suivante :

$$\text{Taux de moteurs année} = \frac{\text{Nombre de véhicules moteurs}}{\text{Taux de présence sur l'année}}$$

Nous aurons donc besoin pour cette donnée des informations : nombre de véhicules moteurs, qui n'existe pas telle quelle dans la base de données portefeuille, et de la donnée de présence sur l'année, présente dans les bases.

2.2 Extraction du portefeuille de contrats

Dans cette première partie, l'objectif est de récupérer le portefeuille au niveau contrat x année et à le préparer pour la jointure avec les sinistres. Nous partons donc de la base portefeuille de maille véhicule x situation, présentée plus tôt dans la partie 2.1.2.

De cette base nous commençons par extraire le périmètre qui nous concerne, en portant une attention particulière à l'information du segment, comme vu dans la partie précédente. À cette étape, nous sommes toujours à la maille véhicule x situation. Nous pouvons alors récupérer la plupart des variables dont nous avons besoin. En revanche, les nombres et taux de moteurs année ne peuvent être récupérés directement. Nous les calculons de la manière suivante :

1. Calcul du nombre de véhicules
   - Si l'ensemble est désigné : nombre de véhicule = 1 car Une ligne par véhicule est créée dans ce cas
   - Si l'ensemble est non désigné : nombre de véhicule = nombre de véhicule de l'ensemble (donnée dans la base) car une ligne par ensemble est créée dans ce cas
2. Calcul du nombre de moteurs
   - Création de l'indicatrice moteur ou non en se basant sur le genre du véhicule
   - Nombre de moteurs = nombre de véhicules × indicatrice moteur/non moteur
3. Calcul du taux de moteurs année
   - taux de moteur année = nombre de moteurs × exposition de la situation sur l'année

L'exposition est directement récupérée dans la base portefeuille et est calculée sur chacune des situations. Ainsi en agrégeant les situations par année, l'exposition se sommera naturellement, et avec, le taux de moteurs année.

Dans un second temps, nous cherchons à obtenir le dernier taux de moteurs année d'un contrat, puisque c'est la donnée utilisée pour calculer le CPT de la garantie DTA. Pour ce faire nous devons récupérer la dernière situation de chaque véhicule/ensemble et ne comptabiliser que les véhicules en cours à la date de fin du contrat. En agrégeant cette information à la maille contrat, nous obtenons l'information du dernier taux de moteurs présents à la fin de chaque année d'existence du contrat.

Maintenant que nous avons notre exposition, le taux de moteurs année, nous calculons cette exposition selon la garantie. Sur chacune des situations nous regardons si la garantie est bien souscrite sur le véhicule/l'ensemble et croisons cette information avec le taux de moteurs année. Cette manipulation est faite sur les trois garanties RC, DTA et BdG.

Nous avons toutes les informations dont nous avons besoin pour pouvoir agréger à la maille contrat x année comme nous le souhaitons. C'est donc naturellement la prochaine étape. Nous groupons alors notre base de données à cette maille, en sommant nos taux moteurs année pour compter l'exposition totale d'un contrat, ainsi que les montants des cotisations. En effet ces derniers sont, dans la base de données portefeuille, recalculés pour être proratisés à la situation, ce qui nous permet de les sommer lors de cette étape afin d'obtenir des cotisations annuelles.

Pour finaliser notre extraction portefeuille, quelques retraitements sont effectués. Le format du numéro de contrat est notamment retraité pour qu'il soit identique à celui de la base sinistre, permettant ainsi de faire la jointure.

Le libellé de certaines données peut également être récupéré pour permettre une meilleure compréhension des données. Comme sur le mode de gestion du contrat par exemple, ce qui permet d'identifier et d'exclure les contrats fictifs de notre base.

2.3 Extraction des sinistres

Dans cette partie est présentée l'extraction des sinistres. Cette extraction doit se faire une fois que le portefeuille est extrait puisque l'extraction des sinistres se base sur son périmètre. En effet, pour extraire le périmètre souhaité depuis la table sinistre, les numéros de contrats et les années extraits précédemment sont utilisés. Ainsi ne sont gardées que les lignes sinistres concernées par le croisement contrat x année.

Suite à cette première extraction, de nombreux retraitements sont à effectuer afin d'obtenir le format de base désiré.

2.3.1 Actualisation des coûts des sinistres

Puisque le coût des sinistres est voué à être étudier pour le refonte tarifaire mais que le périmètre s'étend jusqu'à 2017, il est nécessaire de retraiter les charges pour qu'elles puissent être considérées en "euros 2021", l'année de vision. En effet, dû à des facteurs comme l'inflation et l'augmentation des coûts de réparation automobile, un sinistre survenu en 2017 n'a pas le même coût que s'il était survenu en 2021. D'autant plus que les coûts de réparations ont largement augmenté ces derniers temps, avec aujourd'hui en plus la problématique de disponibilité de pièces automobiles, ce qui influe donc sur les garanties d'immobilisation du véhicule.

Pour évaluer l'augmentation des coûts dans le temps, nous pouvons regarder deux indicateurs, qui prennent aussi en compte l'inflation :

— Les statistiques SRA (Sécurité et Réparation Automobiles) de recensement de l'évolution trimestrielle des principaux éléments constituant le coût de la réparation des véhicules [24]
— L'historique d'évolution du forfait IDA

Voici l'évolution de ces deux indicateurs.

2017	2018	2019	2020	2021
1420	1446	1482	1568	1678

Table 2.2 – Évolution du forfait IDA de 2017 à 2021

Les statistiques SRA donnent directement l'évolution annuelle des coûts de réparation et l'évolution annuelle du forfait IDA est déductible du tableau précédent.

Année	SRA	IDA	Moyenne SRA-IDA
2017-2018	4.7%	1.8%	3.3%
2018-2019	5%	2.5%	3.8%
2019-2020	6.7%	5.8%	6.3%
2020-2021	4.2%	7%	5.6%

Table 2.3 – Évolution annuelle des indicateurs de l'évolution des coûts

Pour la variation annuelle 2020-2021 des statistiques SRA, elle est estimée à partir des données de 2019 à 2021 en prenant en compte l'atypie de l'année 2020.

De ces données, la mise en *as-if* à 2021 peut être directement calculée et l'impact sur le coût moyen s'observe sur le graphique ci-après.

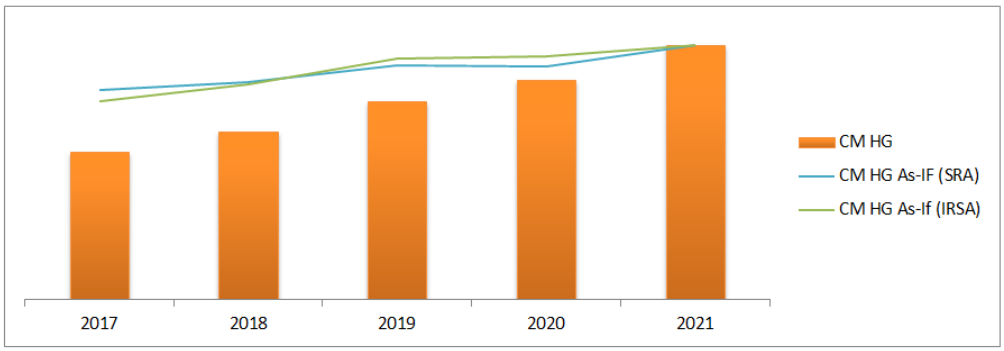

FIGURE 2.4 – Évolution annuelle du coût moyen et du coût moyen *as-if*

Le coût moyen par sinistre, en *as-if* 2021 depuis les statistiques SRA et IRSA, est en augmentation depuis 2017 (+16% environ). Mais cette augmentation est bien lissée par rapport au coût moyen sans retraitement.

L'impact des coefficients SRA et IDA étant relativement similaire, le coefficient finalement choisi est la moyenne des deux. D'autres méthodes et d'autres coefficients auraient pu être implémentés. Nous aurions par exemple pu faire le choix de prendre le maximum de nos statistiques SRA et IDA pour gonfler le poids de l'inflation.

2.3.2 Gestion de la responsabilité

Dans cette partie, l'objectif est de mettre en place la gestion particulière des garanties selon leur responsabilité, vue dans le tableau 2.1. Les taux de sinistres responsables sur les garanties DTA et RC se sont les suivants :

Garantie	Taux de responsables (nb)	Taux de responsables (charge)
DTA	49%	55%
RC	83%	81%

TABLE 2.4 – Répartition de la responsabilité selon la garantie

En DTA, la moitié des sinistres sont des sinistres non-responsables, ils ne seront donc comptés qu'en charge. En RC en revanche, la majeure partie des sinistres sont responsables et seront donc comptabilisés en charge et en nombre. Également, 90% des sinistres RC sont des sinistres RC matériels et seront considérés en charge comme de la DTA.

Aussi certains sinistres de la base de données ne portent pas l'information du taux de responsabilité et il ne nous est pas possible de le retrouver puisqu'il n'est pas encore renseigné par les gestionnaires. Ces sinistres représentent 0.5% de la base et le choix a été fait de les écarter.

2.3.3 Gestion de la sinistralité grave

Dans cette partie, l'objectif n'est pas de redéfinir un seuil de grave pour les garanties, mais plutôt de vérifier si les seuils définis par l'entreprise sont toujours cohérents. Une étude de refonte des seuils de grave pourrait faire l'objet d'un sujet à part entière.

Pour étudier le seuil de grave, plusieurs techniques sont utilisées :
- Méthode des quantiles
- *Mean Excess Plot*
- *Hill Plot*

Méthode des quantiles

Dans un premier temps, nous nous intéressons aux quantiles de la distribution des charges sinistres. Sur la garantie DTA, le seuil actuel correspond au quantile à 99.5% en nombre de sinistres et au quantile à 92.4% en charge. Reste donc à mutualiser entre 7 et 8% de la charge sinistre sur cette garantie. Pour la garantie RC, nous sommes au quantile à 99.8% en nombre et 72.5% en charge. Les sinistres RC sont donc bien plus mutualisés, près de 30%. Ce constat semble logique, leurs montants étant bien plus élevés ; un sinistre RC peut dépasser le million d'euros, là où en DTA il ne dépasse pas les quelques centaines de milliers au maximum. Pour rappel, en charge RC nous

ne considérons que la RC corporelle, la RC matérielle est assimilée à la garantie DTA.

Les méthodes présentées ci-dessous sont notamment décrites dans les cours de Théorie des valeurs extrêmes [12] et [2], ainsi que dans le mémoire de A. PIERRE sur le sujet [3].

Mean Excess Plot

La méthode des excès moyens (*mean excess*) se base sur la fonction des excès moyens qui est l'espérance conditionnelle de l'excédent de charge X par rapport à un seuil donné u, conditionnellement au dépassement de ce seuil.

$$e(u) = \mathbf{E}[X - u | X > u] = \frac{\int_u^{+\infty} (1 - F_X((x))dx}{1 - F_X(u)}$$

À partir de cette fonction, nous pouvons représenter le graphique des excès moyens (*Mean Excess Plot* ou *ME plot*). Cette représentation permet de déterminer graphiquement le seuil à partir duquel nos observations vont suivre une distribution de Pareto Généralisée. En effet, le *ME plot* d'une telle distribution est linéaire en le seuil :

$$e(u) = \frac{\gamma}{1-\gamma} \times u + \frac{\sigma}{1-\gamma}$$

car $F_X(u) = \left(1 + \frac{\gamma u}{\sigma}\right)^{-\frac{1}{\gamma}}$ pour une loi de Pareto Généralisée de paramètres (γ, σ).

Nous cherchons donc le seuil à partir duquel le *ME plot* n'est plus stable. En prenant l'exemple de la garantie DTA nous obtenons le graphique suivant :

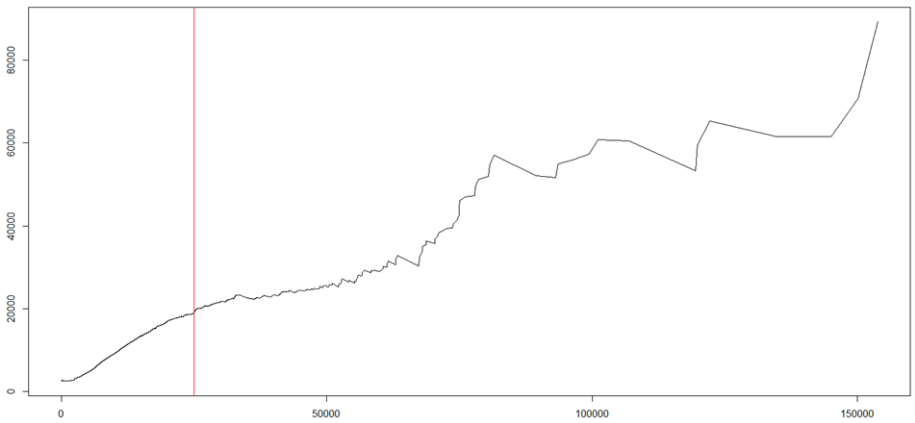

FIGURE 2.5 – *Mean Excess Plot* des charges sur la garantie DTA

La droite verticale rouge représente le seuil actuel. Au vu de ces premiers résultats nous pouvons estimer que le seuil actuel est toujours cohérent avec nos données et permet bien d'éliminer la variance après le seuil. Il serait toutefois possible d'augmenter un peu plus le seuil au vu de ce graphique, mais nous perdrions en mutualisation de la charge grave. Les résultats sur la garantie RC sont équivalents, avec le seuil correspondant.

Hill Plot

Nous nous intéressons également à l'estimateur de Hill et à son graphique associé. Pour cet estimateur, nous devons considérer que la fonction de distribution de notre variable d'intérêt X, la charge sinistre ici, appartient au domaine d'attraction de Fréchet, soit $\gamma > 0$. Dans ce cas l'estimateur de Hill s'écrit en fonction de la moyenne des logarithmes des observations :

$$\widehat{\gamma}_n = \frac{1}{k_n} \sum_{i=1}^{k_n} log\left(X_{n-i+1,n}\right) - log\left(X_{n-k_n,n}\right)$$

Avec k_n le nombre d'excès considéré, à choisir.

Le graphique de Hill consiste à tracer les $\widehat{\gamma_n}$ en fonction de k_n. À partir de ce graphique, nous choisissons le nombre d'excès et la charge correspondante tels que le graphique se stabilise au-dessus de ce point.

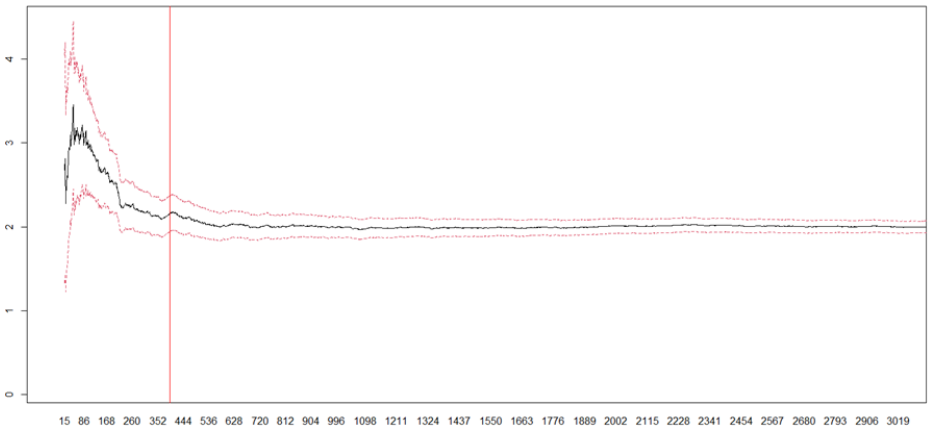

FIGURE 2.6 – *Hill Plot* des charges sur la garantie DTA

En rouge ici aussi le seuil actuel. Ce graphique nous confirme bien que le seuil actuel est toujours cohérent. Même conclusion que sur le *ME plot*, il est possible d'être légèrement plus prudent en baissant le seuil (se déplacer vers la droite sur le graphique).

Ces méthodes de détermination de seuils sont satisfaisantes pour confirmer la justesse des seuils utilisés actuellement.

2.4 Provisionnement

Dans cette partie, nous présenterons la méthode mise en place pour évaluer nos provisions pour sinistres *IBNR* (*Incurred But Not Reported*).

En effet, certains sinistres sont survenus sur notre période d'étude, mais ne sont pas encore déclarés, ils sont appelés *IBNyR* (*Incurred But Not yet Reported*). Nous pouvons également considérer les sinistres survenus mais dont le coût ultime estimé est sous-évalué, appelés *IBNeR* (*Incurred But Not enough Reported*). La somme des *IBNyR* et des *IBNeR* forme les *IBNR* que nous allons chercher à évaluer.

De nombreuses méthodes de provisionnement peuvent être mises en place. La plus connue et utilisée : la méthode de *Chain Ladder*, basée sur l'historique des données et qui estime des facteurs de développement à partir de triangles de liquidation. Une autre méthode envisageable : Bornhuetter-Fergusson, qui vient compléter *Chain Ladder* pour les triangles de liquidation moins stables, particulièrement sur les années les plus récentes en utilisant des informations exogènes au triangle. D'autres méthodes un peu plus complexes et stochastiques permettent également d'estimer nos provisions mais aussi leur variabilité, comme par exemple les méthodes de Mack ou des techniques de Bootstrap ou autres méthodes de *Machine Learning* pour du provisionnement ligne à ligne.

Dans le cadre de ce mémoire, nous nous intéresserons seulement à la méthode de *Chain Ladder*. Nous verrons en effet que, bien que classique, cette méthode est très robuste et suffisante pour provisionner nos données. De nombreux mémoires traitent le sujet du provisionnement et mettent en place des méthodes plus complexes (voir par exemple : [15], [13], [14]). Nous choisissons de provisionner la charge sinistre, ce qui inclus également le nombre de sinistres implicitement.

2.4.1 Principe de la méthode de *Chain Ladder*

La méthode de *Chain Ladder* se base sur un triangle de liquidation de charges. Ce triangle peut être à la maille de l'année, du trimestre, du mois, etc. La profondeur du triangle va dépendre du délai de forclusion des sinistres. Par exemple en assurance automobile, selon la garantie, le délai d'ouverture d'un sinistre peut aller jusqu'à plusieurs années. Dans notre cas, nous choisissons de construire un triangle de liquidation annuel pour chacune de nos garanties.

Dans la suite nous prendrons l'exemple de la garantie DTA, mais le même travail a été mis en place pour les garanties BdG et RC. Nous ne considérons que les charges écrêtées. En effet, les charges atypiques sont bien moins stables, dans leur survenance autant que dans leur délai d'ouverture. Dans un soucis de confidentialité, les données ne représentent pas la réalité, mais le principe s'applique toujours.

La première étape est de constituer notre triangle décumulé.

	DTA				
Annee ouverture	N	N+1	N+2	N+3	N+4
Annee survenance					
2017	36 418 829	8 073 018	217 789	17 049	11 649
2018	39 337 043	7 901 648	235 672	27 562	
2019	41 784 323	8 843 306	262 573		
2020	31 832 933	5 951 031			
2021	37 853 850				

FIGURE 2.7 – Triangle de liquidation décumulé

Ce triangle s'interprète de la manière suivante : sur les sinistres survenus en 2017, nous recensons 36 millions d'euros de charge sur les sinistres ouverts cette même année (année N), 8 millions pour les sinistres ouverts l'année suivante (N+1), de même jusqu'en 2021 où 11 mille euros de charge correspondant aux sinistres de 2017 ont été déclarés.

À partir de ce triangle, nous calculons notre triangle de liquidation cumulé $X_{i,j}^{cum} = X_{i,j-1} + X_{i,j}$, avec :
- i l'année de survenance en ligne
- j l'année d'ouverture en colonne
- $X_{i,j}$ les charges décumulées
- $X_{i,j}^{cum}$ les charges cumulées

Les facteurs de développement de *Chain Ladder* f_j se calculent de la manière suivante :

$$f_j = \frac{\sum_{i=1}^{n-j} X_{i,j+1}}{\sum_{i=1}^{n-j} X_{i,j}}$$

Notre triangle devient donc :

	DTA				
Annee ouverture	N	N+1	N+2	N+3	N+4
Annee survenance					**Ultime**
2017	36 418 829	44 491 847	44 709 636	44 726 685	44 738 334
2018	39 337 043	47 238 691	47 474 363	47 501 925	
2019	41 784 323	50 627 629	50 890 202		
2020	31 832 933	37 783 964			
2021	37 853 850				
		1,2060	1,0050	1,0005	1,0003

FIGURE 2.8 – Triangle de liquidation cumulé et coefficients de *Chain Ladder*

Nous pouvons à présent estimer nos charges provisionnées sur le triangle inférieur :

$$\widehat{X}_{i,j+1} = X_{i,j} \times f_j \Leftrightarrow \widehat{X}_{i,j+k} = X_{i,j} \times \prod_{l=0}^{k-1} f_{j+l}$$

	DTA				
Annee ouverture	N	N+1	N+2	N+3	N+4
Annee survenance					Ultime
2017	36 418 829	44 491 847	44 709 636	44 726 685	**44 738 334**
2018	39 337 043	47 238 691	47 474 363	47 501 925	**47 514 297**
2019	41 784 323	50 627 629	50 890 202	50 914 830	**50 928 090**
2020	31 832 933	37 783 964	37 974 010	37 992 387	**38 002 282**
2021	37 853 850	45 651 271	45 880 888	45 903 092	**45 915 047**
		1,2060	1,0050	1,0005	1,0003

FIGURE 2.9 – Complétion du triangle inférieur par *Chain Ladder*

La charge à l'ultime apparaît donc en N+4. Avec la méthode de *Chain Ladder*, nous provisionnons :

2017	2018	2019	2020	2021	**Total**
0	12 372	37 888	218 318	8 061 197	**8 329 775**

TABLE 2.5 – Provisionnement *Chain Ladder*

Cette méthode a l'avantage d'être très simple à implémenter et d'être particulièrement robuste, lorsque les hypothèses sous-jacentes à la méthode sont respectées. Ces hypothèses sont étudiées dans la partie suivante.

Cependant, un inconvénient principal de *Chain Ladder* est qu'il ne modélise pas toujours bien les années récentes, d'où la mise en place d'autres méthodes en complément comme Bornhuetter-Fergusson. Aussi, en agrégeant nos charges sous forme de triangles, nous perdons de l'information sur les sinistres. Nous ne pouvons plus différencier l'évolution de la sinistralité (hausse ou baisse de la sinistralité selon le contexte portefeuille ou d'autres facteurs exogènes) de la variation de sinistres dûe à l'enregistrement de sinistres tardifs (*IBNR*).

2.4.2 Vérification des hypothèses

Afin de s'assurer de pouvoir utiliser la méthode de provisionnement de *Chain Ladder*, nous devons vérifier quelques hypothèses.

L'hypothèse principale de la méthode *Chain Ladder* est la suivante : Les facteurs de développement individuels (ou coefficients de passage) $f_{i,j} = \frac{X_{i,j+1}}{X_{i,j}}$ doivent être indépendants de l'année de survenance i. Ils doivent donc être linéaires en i. Pour vérifier cela, regardons la linéarité des couples $(X_{i,j}, X_{i,j+1})$, représentés ci-dessous :

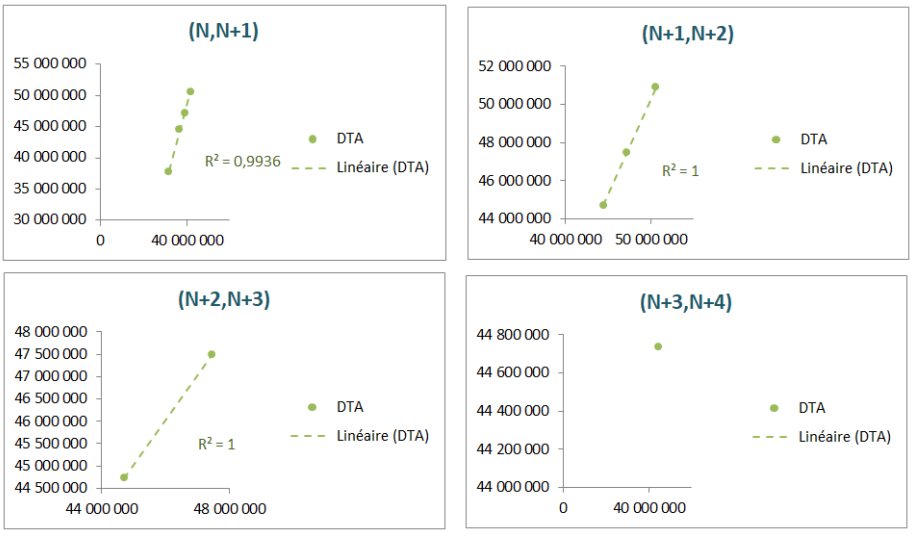

FIGURE 2.10 – Linéarité des couples $(X_{i,j}, X_{i,j+1})$

Sur les points des couples $(X_{i,j}, X_{i,j+1})$, une courbe de tendance linéaire a été implémentée pour chaque graphique. Visuellement, la tendance s'observe bien. Ce qui est appuyé par l'estimation du R^2 sur chaque graphique : de 0.9936 et 1 pour les graphiques comparant les années $(N, N+1)$ et $(N+1, N+2)$. Pour le graphique $(N+2, N+3)$, le R^2 est évidemment de 1 puisqu'il ne se compose que de 2 points. Le graphique $(N+3, N+4)$ ne comporte qu'un unique point, il n'y a pas de tendance.

Nous pouvons également regarder directement la linéarité des coefficients de passage, calculés dans le tableau ci-dessous :

	DTA				
Annee ouverture / Annee survenance	N	N+1	N+2	N+3	N+4
2017	-	1,22	1,00	1,00	1,00
2018	-	1,20	1,00	1,00	
2019	-	1,21	1,01		
2020	-	1,19			
2021	-				

FIGURE 2.11 – Coefficients de passage $f_{i,j}$

Graphiquement nous obtenons :

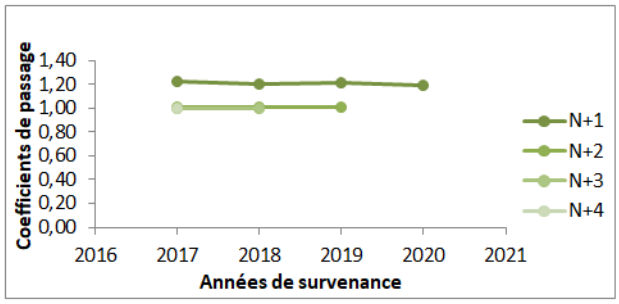

FIGURE 2.12 – Linéarité des coefficients de passage $f_{i,j}$

Encore une fois, nous sommes rassurés quant à la linéarité de nos coefficients. Même imparfaite, elle est suffisamment proche pour permettre de mettre en place la méthode de *Chain Ladder*.

Également, nous devons vérifier que la profondeur de notre triangle de développement est suffisante. Pour cela, intéressons-nous maintenant aux délais d'ouverture des sinistres. Pour ce faire, la distribution du délai d'ouverture d'un sinistre, calculé comme la différence entre la date d'ouverture et la date de survenance du sinistre, est étudiée dans le graphique qui suit.

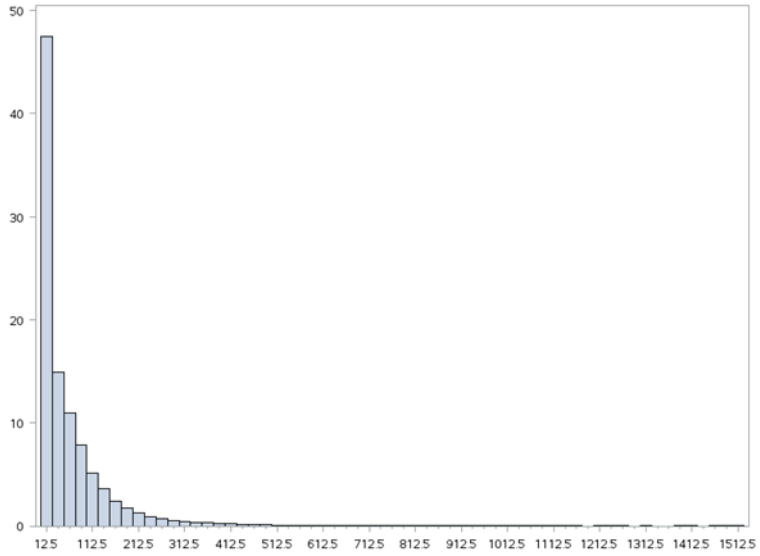

FIGURE 2.13 – Distribution du délai d'ouverture sinistre (jours)

Depuis cette distribution, nous remarquons que le délai d'ouverture maximum est de près de 4 ans, mais 99% des sinistres sont ouverts sous 1 an - 1 an et demi. Ces résultats valent aussi bien pour la garantie DTA que pour les garanties BdG et RC. Notre profondeur d'étude (2017-2021) nous semble donc satisfaisante.

Nous pouvons donc bien appliquer *Chain Ladder*.

2.4.3 Application du provisionnement sous SAS

La mise en place de la base de données se faisant sous SAS, nous y implémentons donc naturellement le provisionnement par méthode de *Chain Ladder*, grâce à une macro SAS qui nous permet de récupérer les coefficients f_j, qui sont ensuite stockés dans des macros variables pour être réutilisés.

La macro fonctionne en 3 étapes :

1. Construction du triangle date de survenance-date d'ouverture
2. Calcul du triangle cumulé
3. Calcul et stockage des coefficients *Chain Ladder*

Nous appliquons ces coefficients à la diagonale des charges des sinistres déjà existants. Ainsi, lorsque nous agrégerons nos lignes dans un deuxième temps, seront sommés nos sinistres déjà survenus et ouverts, avec ceux qu'il reste à ouvrir. Nous aurons donc une approximation de *Chain Ladder*.

	DTA							
Annee ouverture / Annee survenance	N	N+1	N+2	N+3	N+4	Application Ultime	Chain Ladder Ultime	*Différence*
2017	36 418 829	8 073 018	217 789	17 049	11 649	44 738 334	44 738 334	*0,00%*
2018	39 337 043	7 901 648	235 672	27 562	27 569	47 501 932	47 514 297	*-0,03%*
2019	41 784 323	8 843 306	262 573		262 769	50 890 397	50 928 090	*-0,07%*
2020	31 832 933	5 951 031			5 985 417	37 818 349	38 002 282	*-0,48%*
2021	37 853 850				45 915 047	45 915 047	45 915 047	*0,00%*
		1,2060	1,0050	1,0005	1,0003			

FIGURE 2.14 – Principe de l'application de *Chain Ladder*

Par exemple, pour un sinistre survenu en 2019 et ouvert deux ans plus tard (en N+2), nous lui appliquons $f_3 \times f_4$. Nous aurons donc son équivalent en N+4, que nous sommons ensuite avec les sinistres ouverts avant la diagonale pour obtenir notre équivalent en ultime.

En appliquant ces coefficients de la sorte, nous accordons plus de poids sur les sinistres déjà présents au portefeuille. Cependant, dans la réalité nous voudrions « créer de nouveaux sinistres », ce qui est difficile à mettre en place. La question du biais induit par cette méthode peut alors se poser : est-ce que certains profils de risque ne seraient pas lésés par cette méthode ? À priori non, d'après la distribution des fréquences par année par garantie et par critère tarifaire.

Avec cette dernière étape, nous avons maintenant notre base de données sinistre, que nous pouvons joindre avec notre portefeuille. Cette base de données à la maille sinistre peut être utilisée pour étudier le coût moyen.

2.5 Jointure des sinistres à la base des contrats

Avant de pouvoir joindre nos bases portefeuille et sinistres, nous devons nous assurer qu'elles soient à la même maille. Pour l'instant nos bases ont la granularité suivante :

— Base portefeuille : contrat x année
— Base sinistre : garantie sinistrée

Pour rappel, nous souhaitons avoir la granularité contrat x année. Quelques travaux de regroupement sont donc à mettre en place sur la base sinistre.

Dans un premier temps, nous regroupons notre base sinistre à la maille contrat x année x garantie. À partir de cette maille, nous pouvons passer les garanties et les informations sinistres qui y sont liées (nombres, charges, etc.) en colonne. Notre base sinistre est donc bien agrégée à la maille contrat x année.

Nous lions donc nos bases en prenant soin de conserver les lignes portefeuille non sinistrées, pour étudier la fréquence.

Dans un second temps, nous agrégeons notre base à la maille contrat, en passant les informations par année en colonne. Nous aurons donc nos taux moteurs, notre nombre de moteurs, nos sinistres et nos fréquences par année, par garantie, en colonne. Notre base de données d'étude de la fréquence est ainsi construite.

2.6 Statistiques descriptives

Suite à la mise en place de notre base de données, nous pouvons commencer par nous intéresser à notre exposition, le taux de moteurs année. Nous regardons l'exposition totale des 4 ans, soit la somme des taux moteurs année sur la période, mais aussi la dernière exposition connue.

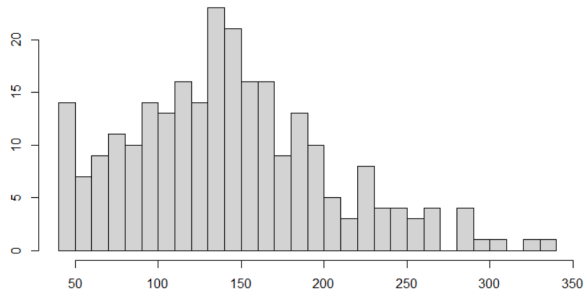

FIGURE 2.15 – Distribution de l'exposition totale sur 4 ans

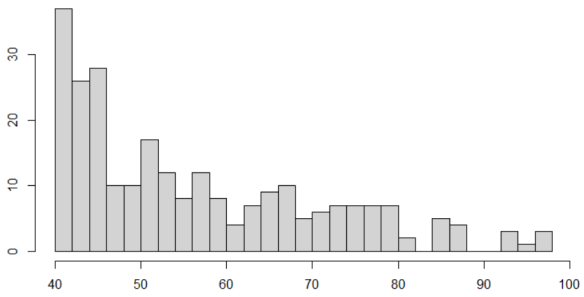

FIGURE 2.16 – Distribution de la dernière exposition connue

Les contrats les plus importants en terme d'exposition comptent au total 337 moteurs, soit une moyenne de 84.25 moteurs par an. Le contrat le plus exposé en compte 98. L'exposition minimale est de 40 moteurs avec un historique de 1 an. Au regard de la distribution, notre portefeuille est principalement composé de contrats avec entre 40 et 46 moteurs et très peu au-dessus de 80. En moyenne, la profondeur d'historique de nos contrats est de 3 ans.

Plus du tiers des contrats sont au portefeuille depuis au moins 4 ans, même ordre de grandeur pour les contrats de 1 an ou moins. Le dernier tiers est réparti de manière assez homogène sur les contrats de 2 à 3 ans.

1 an	2 ans	3 ans	4 ans
28.59%	17.62%	19.41%	34.38%

TABLE 2.6 – Distribution de la profondeur d'exposition

Nous pouvons également étudier nos distributions de fréquences, qui sont étudiées hors-grave.

Nous commençons par regarder l'évolution de notre fréquence, sur les garanties RC, DTA, BdG et sur le total toutes garanties.

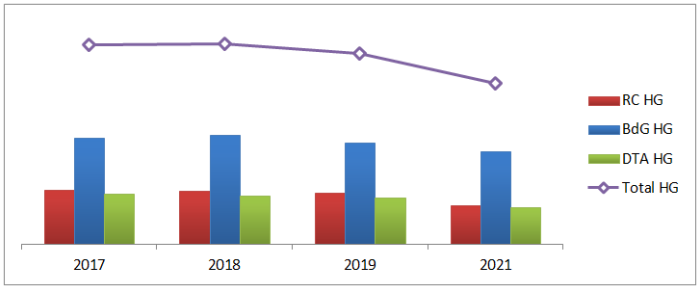

FIGURE 2.17 – Évolution de la fréquence totale et par garantie

Comme nous avons déjà pu le voir dans la partie 1.3.1 la fréquence sinistre est bien en baisse sur notre portefeuille sur les dernières années, de près de 19%. Cette évolution est surtout portée par les garanties RC et DTA, avec des baisses respectives de 27.5% et 26.4%. La garantie BdG enregistre une baisse moins prononcée de 12.6%.

Afin d'alléger la présentation des résultats suivants, seules les fréquences de la garantie DTA seront exposées. Les conclusions apportées sur cette garantie, en termes d'évolution, de répartition sur les critères tarifaires et d'exposition, peuvent être translatées sur les garanties RC et BdG. La différence principale entre les trois garanties étant l'échelle de la fréquence ; la garantie BdG est plus sinistrée que les garanties RC et DTA comme nous avons pu le voir sur le graphique précédent.

Critère de zone

Nous pouvons commencer par regarder l'évolution de la sinistralité selon la zone géographique à laquelle est rattaché le contrat.

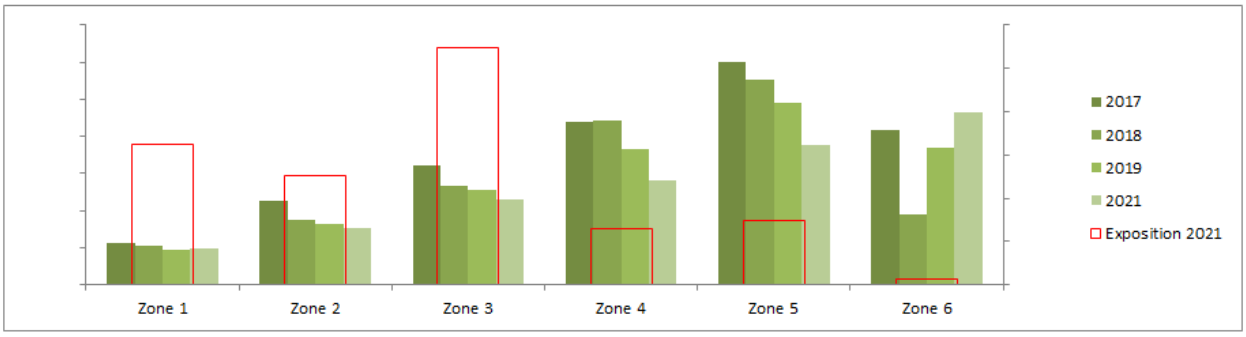

FIGURE 2.18 – Évolution annuelle et exposition de la fréquence DTA par zone

Sur ce premier graphique, nous observons l'évolution annuelle de la sinistralité sur chacune des zones. L'exposition au risque est également rappelée sous forme de rectangles rouges et correspond au taux de moteurs année. L'évolution à la baisse est plutôt bien suivie sur l'ensemble des zones, à l'exception de la zone 6. Certaines zones connaissent une évolution plus ou moins forte, par exemple la zone 5 qui enregistre une baisse de 22.4% alors que la zone 1 évolue seulement de -3%. Les baisses de fréquence les plus prononcées sont sur les zones où la fréquence

sinistre était déjà plus élevée. Pour ce qui est de la zone 6, il nous est difficile de tirer de réelles conclusions du fait de la très faible exposition sur cette zone, entraînant une forte volatilité.

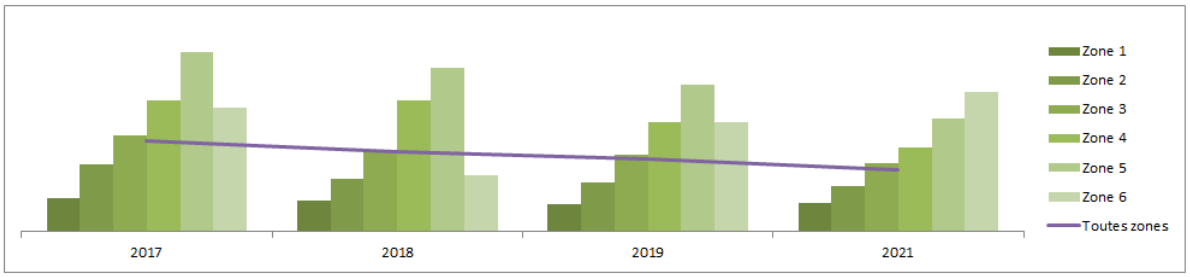

FIGURE 2.19 – Évolution de la fréquence DTA selon la zone par année

Sur ce deuxième graphique, nous observons la sinistralité en fonction de la zone, année par année. Les conclusions sur ce graphique nous rassurent quant à la qualité de notre zonier. En effet, nous observons bien une hausse relativement constante de la sinistralité en fonction de la zone. En considérant la moyenne des 4 années d'observation par zone, nous avons l'évolution suivante entre les zones 1 à 5 :

Zones	Zone 1 à 2	Zone 2 à 3	Zone 3 à 4	Zone 4 à 5
Évolution moyenne	+15.5%	+15.6%	+16.8%	+15.8%
Écart-type	4.8%	0.7%	7.3%	2.9%

L'évolution est bien constante entre les différentes zones, de l'ordre de 16% en moyenne.

Ces conclusions ne s'appliquent pas à la zone 6, toujours du fait de sa forte volatilité. Une amélioration de ce zonier pourrait être de rapprocher cette zone avec une autre pour palier à ce problème d'exposition. Théoriquement, la zone 6 est plus proche en sinistralité de la zone 5, un rapprochement de ces zones pourrait être envisagé. La problématique de la zone 6 dans la modélisation sera traitée dans la suite. Ce traitement se fera directement dans le modèle du fait que le zonier ne sera pas réétudier dans le cadre de ce mémoire.

Critère d'activité

Intéressons nous maintenant au deuxième critère tarifaire : l'activité. Nous pouvons sortir les même graphiques que pour le critère de zone.

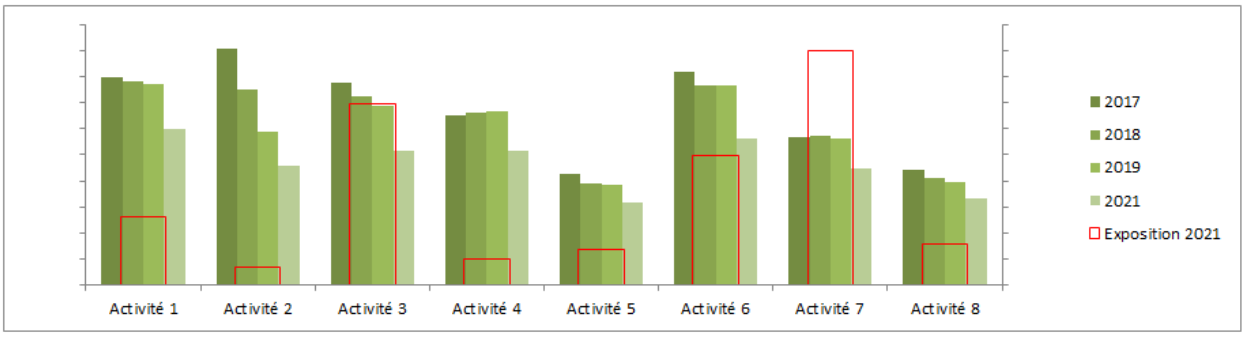

FIGURE 2.20 – Évolution annuelle et exposition de la fréquence DTA par activité

Sur ce graphique mettant en avant l'évolution annuelle sur chacune des activités, nous retrouvons à nouveau notre baisse de sinistralité. Ici aussi l'évolution est plus ou moins prononcée selon le type d'activité. Sur l'activité 2 par exemple, nous enregistrons une baisse de 49.6% de la fréquence sinistre. Mais sur l'activité 7, celle dont la baisse est la moins prononcée, nous sommes plutôt aux alentours de 19.7%, ce qui reste assez élevé toutefois.

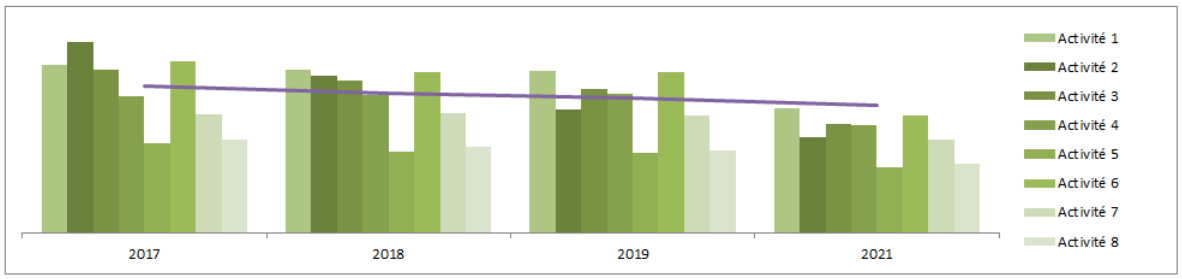

FIGURE 2.21 – Évolution de la fréquence DTA selon l'activité par année

Lorsque les activités sont comparées entre elles sur chacune des années de notre périmètre, la conclusion n'est pas aussi évidente que sur les zones. Certaines activités se démarquent, comme les activités 5, 7 et 8 car elles présentent la sinistralité la plus faible. À l'inverse, les activité 1 et 4 semblent être toujours plus sinistrées que les autres. Mais les différences s'amenuisent avec les années et les écarts ne sont plus si flagrants en 2021. Les regroupements des activités sont effectués dans un objectif de gestion, à l'inverse du zonier où les zones créées sont des groupes de risques distincts. Ces résultats ne nous surprennent donc pas. Nous prêterons une attention particulière à la significativité des modalités de la variable activité comme critère tarifaire pour essayer d'en dégager une tendance.

À l'analyse de ces résultats, il nous semble bien pertinent de conserver la zone et l'activité comme critères tarifaires.

Chapitre 3

Révision de la fréquence collective

3.1 Étude des lois de fréquence

Nous avons maintenant à notre disposition une base de données propre à l'étude de la fréquence de notre périmètre. Pour rappel, nous cherchons à mettre à jour la fréquence suivante :

$$Freq^{cred} = z \times Freq^{obs} + (1 - z) \times Freq^{th}$$

Avec :
- $Freq^{cred}$: la fréquence crédibilisée
- $Freq^{obs}$: la fréquence observée, la fréquence individuelle de l'assuré issue de son historique sinistre
- $Freq^{th}$: la fréquence théorique, la fréquence collective du portefeuille que nous cherchons à modéliser

Pour ce faire, nous procédons en 2 étapes :

1. Mise à jour de la fréquence théorique
2. Mise à jour du coefficient de crédibilité

Cette partie sera consacrée à la première étape de refonte de la fréquence collective.

$$\text{Fréquence} = \frac{\text{Nombre de sinistres}}{\text{Exposition}}$$

Ainsi, avant de commencer tout type de modélisation, intéressons-nous aux lois de probabilité des fréquences sur les garanties RC, DTA et BdG. Plus précisément, nous cherchons une approximation par une loi usuelle de notre nombre de sinistre. En effet, l'exposition sera utilisée en offset des modélisations, mais nous évoquerons ce sujet dans la partie suivante.

Ici encore nous choisirons de ne regarder que la garantie DTA afin d'alléger ce mémoire, mais le travail a été effectué sur les trois garanties conjointement. Les analyses et conclusions s'appliqueront donc aussi bien à la garantie DTA qu'aux garanties RC et BdG.

Dans un premier temps, nous choisissons de modéliser la fréquence des sinistres hors-graves.

L'ensemble des études qui suivent a été traité sous R.

3.1.1 Allure des distributions

Regardons dans un premier temps l'allure de nos distributions de sinistres DTA sur l'ensemble des années d'étude. Pour ce faire, nous pouvons implémenter l'histogramme de la distribution :

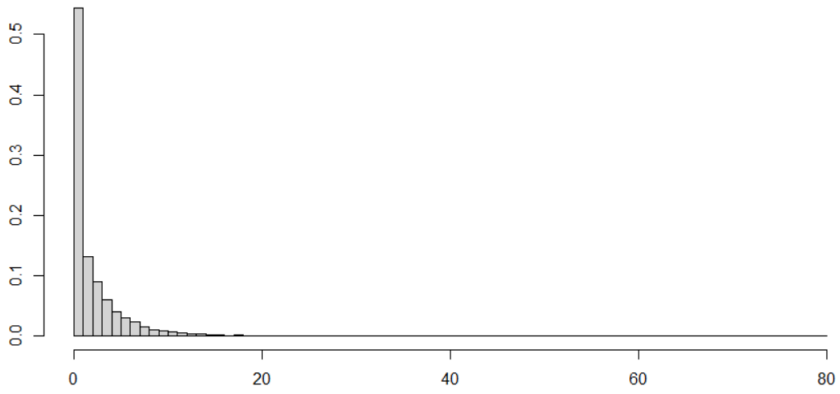

FIGURE 3.1 – Histogramme du nombre de sinistres

Dans nos données, un contrat peut avoir entre 0 et 80 sinistres. Le tiers de nos assurés n'ont déclaré aucun sinistre DTA. Et seulement 5% d'entre eux ont déclaré plus de 10 sinistres. Nous avons donc un lourd poids en 0 et une queue de distribution très étalée.

En moyenne, le nombre de sinistre est de 2.5, avec une variance de 16. La variance étant supérieure à l'espérance, la loi de la distribution est sur-dispersée.

La distribution année par année a l'allure suivante :

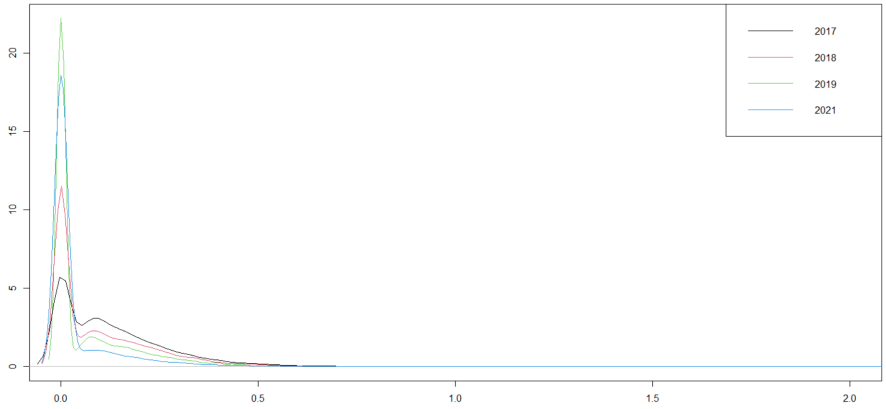

FIGURE 3.2 – Densités à noyaux des distributions de sinistre annuelles, zoom sur $[0, 2]$

Pour mieux apprécier le comportement autour de 0, l'échelle de l'abscisse a volontairement été tronquée sur $[0, 2]$, sans quoi l'échelle écrase complètement l'allure. Nous remarquons qu'avec les années, le poids en 0 s'accentue, la baisse de fréquence observée jusqu'ici se retrouve donc bien. De manière générale, nous retrouvons bien l'allure de l'histogramme précédent.

L'allure décroissante de la distribution empirique nous fait penser à une loi de Poisson ou Négative Binomiale, qui sont toutes deux des lois classiques d'estimation de la fréquence. La sur-dispersion de nos données nous ferait plutôt pencher sur une loi Négative Binomiale. Aussi nous devons porter une attention particulière à ce poids en 0, qui peut aussi nous faire penser à des modèles Zéro-Inflaté.

3.1.2 Approximation par une loi de Poisson

Pour rappel, soit X une variable aléatoire discrète positive, si $X \sim \mathcal{P}(\lambda)$ avec $\lambda > 0$:

$$\mathbf{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \ \forall \ k \geq 0$$

Avec λ l'unique paramètre de la loi de Poisson tel que $\mathbf{E}[X] = \mathbf{V}(X) = \lambda$. La loi de Poisson est donc une loi equi-dispersée.

Malgré la sur-dispersion des données, nous regardons tout de même l'approximation de notre distribution empirique avec une loi de Poisson. Pour cela, nous utilisons la fonction *fitdist* du package*fitdistrplus* de R [22] pour une approximation par maximum de vraisemblance. Cette fonction a l'avantage de produire des sorties graphiques et des critères d'appréciation de l'approximation.

Nous pouvons commencer par regarder les graphiques :

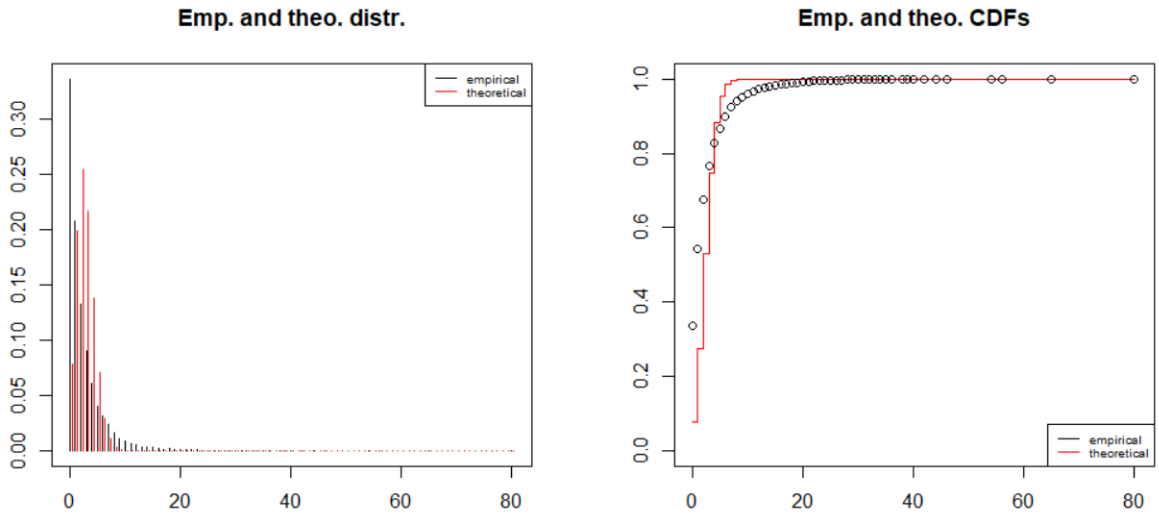

FIGURE 3.3 – Histogrammes et fonctions de répartition empiriques et estimés par une loi de Poisson

Graphiquement, nous remarquons d'emblée que la distribution de Poisson ne capte pas bien le poids en 0 et semble décalée. À première vue, nous ne pouvons pas accepter la distribution de Poisson.

Le paramètre λ est pourtant bien estimé, à environ 2.5 avec un écart-type d'erreur de 0.0136.

3.1.3 Approximation par une loi Négative Binomiale

Regardons maintenant l'approximation par un loi Négative Binomiale dont nous estimerons les paramètres.

Soit X une variable aléatoire discrète positive, si $X \sim \mathcal{NB}(n, p, q)$ avec $n \in \mathbf{N}^*$, $p \in]0, 1]$ et $q = 1 - p$:

$$\mathbf{P}(X = k) = \frac{\Gamma(n+k)}{k!\Gamma(r)} p^n q^k \ \forall k \geq 0$$

Avec la fonction gamma : $\Gamma(x+1) = x!$

Et $\mathbf{E}[X] = \frac{n(1-p)}{p} \leq \mathbf{V}(X) = \frac{n(1-p)}{p^2}$, ainsi la loi Négative Binomiale est sur-dispersée. Nous utilisons toujours la fonction *fitdist* pour approcher notre distribution de données par une loi Négative Binomiale et nous en sortons le graphique d'adéquation.

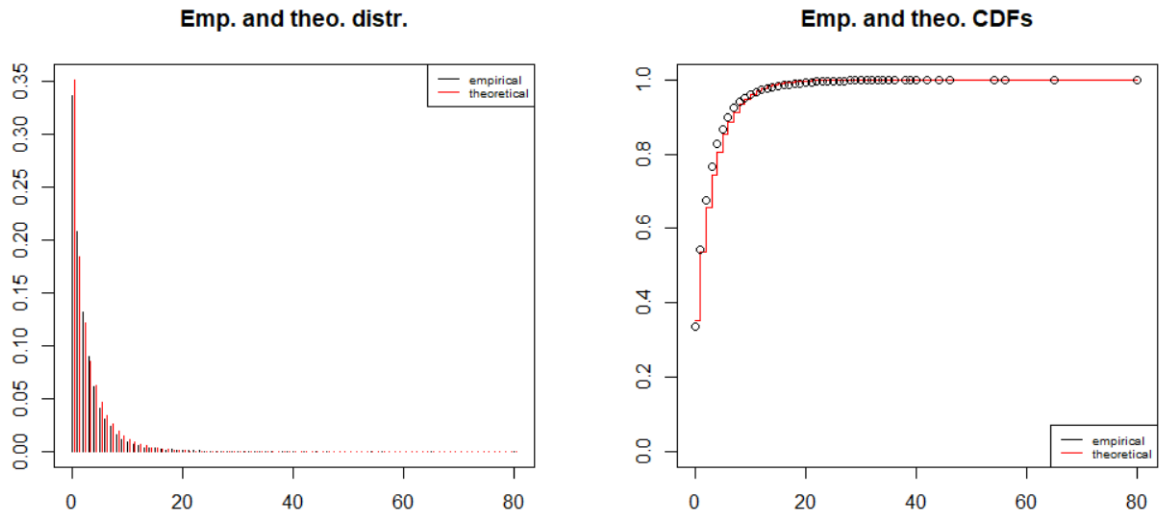

FIGURE 3.4 – Histogrammes et fonctions de répartition empiriques et estimés par une loi Négative Binomiale

Graphiquement les résultats sont déjà plus en adéquation. Le poids en 0 est bien capté et le reste des données le semble également.

Les paramètres estimés par *fitdist* sont les suivants : $n = 0.66$ et $\mathbf{E}[X] = 2.55 \Leftrightarrow p = 0.21$.

3.1.4 Loi choisie

Pour valider notre choix de loi, nous regardons différents critères. Ces critères se trouvent en sortie de la fonction *gofstat* de R qui prend en paramètres nos *fitdist* Poisson et Négatif Binomial précédents.

	Poisson	Négative Binomiale
AIC	84 024.59	57 384.30
Maximum de vraisemblance	42 011.29	28 690.15

TABLE 3.1 – Indicateurs de performance des approximations en loi

Les critères d'*AIC* et de maximum de vraisemblance satisfont notre choix de la loi Négative Binomiale plutôt que la loi de Poisson.

Également, nous regardons la "table du χ^2" en sortie de la fonction *gofstat*, table qui compte le nombre d'observation et le nombre de valeurs théoriques pour le calcul du χ^2, sur chaque modalité.

	obscounts	theo pois	theo nbinom
<= 0	2397	1.939714e+01	2574.85431
<= 1	1723	1.272726e+02	1681.16439
<= 2	1505	4.175440e+02	1305.74972
<= 3	1189	9.132261e+02	1068.04160
<= 4	934	1.498013e+03	895.64081
<= 5	788	1.965816e+03	762.15456
<= 6	675	2.149754e+03	654.85226
<= 7	542	2.015060e+03	566.51665
<= 8	472	1.652705e+03	492.60133
<= 9	418	1.204897e+03	430.02369
<= 10	389	7.905827e+02	376.57842
<= 11	328	4.715762e+02	330.62306
<= 12	269	2.578504e+02	290.89587
<= 13	243	1.301433e+02	256.40386
<= 14	192	6.099455e+01	226.35039
<= 15	194	2.668069e+01	200.08635
<= 17	267	1.516445e+01	333.94957
<= 19	240	2.070993e+00	262.54777
<= 22	240	2.451493e-01	293.69212
<= 26	206	6.320579e-03	261.23583
<= 32	182	2.656206e-05	222.96717
<= 42	190	2.525325e-09	155.50760
> 42	136	0.000000e+00	76.56266

FIGURE 3.5 – Table du χ^2 pour les lois de Poisson et Négative Binomiale

Nous confirmons que la loi Négative Binomiale s'adapte bien mieux aux nos données, sur chaque intervalle. La loi de Poisson sous-estime largement les sinistres au delà de 14 et ne les estime pas du tout au delà de 42. La loi Négative Binomiale sous-estime les sinistres au delà de 42, elle en compte environ 77 au lieu de 136 mais les estime tout de même, mieux que la loi de Poisson.

3.2 Modélisation de la fréquence théorique

Dans cette section, seront présentées les différentes techniques de modélisation utilisées pour estimer la fréquence sinistre collective. Nous nous intéressons tout particulièrement aux Modèles Linéaires Généralisés (MLG) pour leur robustesse et leur simplicité d'interprétation qui en font aujourd'hui l'outil le plus utilisé en tarification actuarielle. Nous nous pencherons également sur un modèle de *Machine Learning*, les forêts aléatoires, qui ont également la réputation d'être très performants. Ce modèle sera comparé aux prédictions en sortie de MLG.

Afin de valider les modèles, une méthode de Validation Croisée (*Cross Validation* ou *CV*) est mise en place. Ces méthodes consistent à faire tourner un même modèle sur plusieurs échantillons d'entraînement différents afin de mesurer la volatilité d'un modèle et réduire le sur-apprentissage. La *Cross Validation K-fold* consiste à créer K échantillons, d'entraîner le modèle sur $K-1$ échantillons et de le tester sur le restant. En itérant autant de fois qu'il y a d'échantillons. Sur le même principe, la *Leave One Out Cross Validation* ou *LOOCV* entraîne le modèle sur un échantillon de taille $N-1$, avec N la taille de la base, et le test sur la ligne restante, N fois. Cette dernière méthode est particulièrement coûteuse en temps de calcul lorsque notre base de données est grande. Pour limiter ce temps de calcul tout en gardant l'efficacité de la Validation Croisée, nous implémentons 149 échantillons aléatoires ; $K = 149$. Ce chiffre a été choisi car multiple du nombre de lignes.

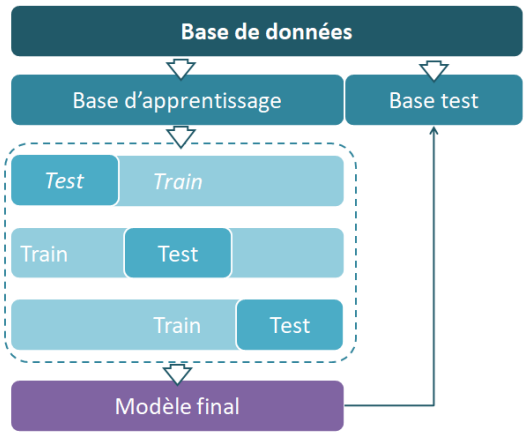

FIGURE 3.6 – Schéma explicatif d'une Validation Croisée K-Fold, K=3

À partir d'ici, nous nous sommes concentrés uniquement sur la garantie DTA, le même travail pourra être mis en place sur les garanties RC et BdG dans un second temps.

3.2.1 Approches MLG

Les modèles linéaires généralisés permettent d'expliquer une variable cible X, de distribution d'une loi de famille exponentielle, en fonction d'un vecteur de variables explicatives Y. Les MLG fonctionnent de manière analogue à des régressions linéaires, à la différence que la dépendance modélisée est non-linéaire puisque les MLG font intervenir une fonction de lien g. Le modèle a la forme suivante :

$$g(\mathbf{E}[X|Y]) = Y^t\beta$$

Les β sont les coefficients de la régression que nous cherchons à estimer, les X nos sinistres et Y les paramètres.

Les modèles MLG n'acceptent en *input* du modèle que des variables décorrélées. En cas de corrélation entre les variables explicatives, des retraitements et regroupements devront donc être envisagés. Ce n'est pas notre cas ici, les variables de zone et d'activité sont bien décorrélées. Pour vérifier cela, la méthode du V de Cramer est utilisée, cette mesure permet d'évaluer la dépendance entre deux variables qualitatives et se calcule de la manière suivante :

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

Avec n le nombre d'observations, p le nombre de modalités de la variable 1 et q celui de la variable 2. χ^2 est la statistique du Chi-2 de contingence. Pour plus de détail sur le calcul du χ^2 de contingence et le V de Cramer, se référer au cours [6]. Pour évaluer cette statistique, deux solutions sont possibles. La première est de regrouper les variables de zone et d'activité dans un tableau de contingence avec la fonction *table()* de R, et d'y appliquer la fonction *cramer.v()* du package *questionr* pour obtenir la statistique. La deuxième est de calculer le V à la main depuis sa formule, en utilisant la fonction *chisq.test()* pour évaluer le χ^2. Dans les deux cas, le V est estimé à 0.0634, statistique très proche de 0. Les variables de zone et d'activité sont donc considérées comme décorrélées. Dans le cas contraire, le V de Cramer aurait été plus proche de 1.

Pour mettre en place un MLG sous R, la fonction *glm()* du package *stats* peut être utilisée. Cette fonction requiert en paramètres la formule de modélisation, la distribution de la donnée cible et sa fonction de lien, ainsi que la base de données utilisée pour l'étude et un éventuel échantillon d'observations à utiliser pour entraîner le modèle.

Dans notre cas, nous allons chercher à modéliser le nombre de sinistres (notre variable cible), dont nous avons précédemment approché la loi par une distribution Négative Binomiale, en fonction de la zone et de l'activité de l'assuré. Notre base de données d'étude est la base précédemment construite dont nous récupérons uniquement les contrats ayant souscrit à la garantie DTA. Cette base d'étude propre à la garantie DTA compte près de 13 000 lignes, ce qui est un volume suffisant de données pour construire une modélisation robuste. À partir de cette base, un échantillon d'apprentissage (*train*) est construit. Tous nos modèles seront implémentés sur cette base *train* par Validation Croisée avant d'être testés sur une base de test pour s'assurer de la solidité des modèles et éviter le sur-apprentissage. L'échantillon *train* est construit aléatoirement et contient 70% de la base d'étude.

Dans un soucis d'élargir nos modèles, nous testerons également des MLG avec une loi de Poisson. Dans les deux cas, loi de Poisson ou Négative Binomiale, la fonction de lien correspondante est la fonction $log()$, nous aurons donc une tarification multiplicative. Notre modèle est donc le suivant :

$$\begin{aligned} log(\text{Nombre de sinistres}) &= \beta_0 + \beta_{activite} + \beta_{zone} + log(\text{Exposition}) \\ \Leftrightarrow \text{Nombre de sinistres} &= e^{\beta_0+\beta_{activite}+\beta_{zone}} \times \text{Exposition} \\ \Leftrightarrow Freq^{th} &= e^{\beta_0+\beta_{activit}+\beta_{zone}} \end{aligned}$$

Avec β les coefficients estimés par le MLG.

Avant de mettre en place nos MLG, nous définissons les catégories de référence, soit les catégories les plus représentées dans notre base :

- Activité : 5
- Zone : 3

En effet, sur notre base *train*, nous avons l'exposition suivante :

	Z1	Z2	Z3	Z4	Z5	Z6
1	178	142	233	48	19	2
2	105	77	166	35	45	5
3	630	500	1 234	324	405	29
4	238	238	357	67	49	7
5	992	741	1 657	408	489	33
6	95	112	152	54	24	3
7	228	182	482	154	176	11
8	409	412	851	246	341	25

FIGURE 3.7 – Exposition sur les croisements zone x activité de la base train

Nous retrouvons bien notre manque de représentation sur la zone 6. Également sur certains croisements nous avons un peu moins de données. Dans un premier temps, nous faisons le choix de conserver notre zone 6 sans la regrouper avec une autre zone. En effet nous sommes dans une démarche de refonte des fréquences sur l'ensemble des croisements.

MLG Poisson

En sortie d'un MLG avec une distribution de Poisson, nous avons les informations suivantes :

```
Coefficients:
                   Estimate Std. Error  z value Pr(>|z|)
(Intercept)        -2.87147    0.01496 -191.926  < 2e-16 ***
CD_ZONE_TARI_GP1   -0.28077    0.01872  -15.002  < 2e-16 ***
CD_ZONE_TARI_GP2   -0.13425    0.01929   -6.959 3.43e-12 ***
CD_ZONE_TARI_GP4    0.14582    0.02121    6.875 6.18e-12 ***
CD_ZONE_TARI_GP5    0.24883    0.01922   12.946  < 2e-16 ***
CD_ZONE_TARI_GP6    0.12711    0.05830    2.180   0.0292 *
CD_FAMI_ATVT_RETE7  0.25602    0.02267   11.292  < 2e-16 ***
CD_FAMI_ATVT_RETE2  0.20492    0.04097    5.001 5.69e-07 ***
CD_FAMI_ATVT_RETE3  0.18200    0.01748   10.411  < 2e-16 ***
CD_FAMI_ATVT_RETE6  0.07574    0.03453    2.194   0.0283 *
CD_FAMI_ATVT_RETE1 -0.27001    0.04051   -6.665 2.65e-11 ***
CD_FAMI_ATVT_RETE8  0.22215    0.01857   11.965  < 2e-16 ***
CD_FAMI_ATVT_RETE4 -0.27314    0.03932   -6.947 3.73e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 18682  on 9386  degrees of freedom
Residual deviance: 17361  on 9374  degrees of freedom
```

FIGURE 3.8 – Sortie R d'un MLG avec loi de Poisson

D'après cette sortie, tous les coefficients sont significatifs et nous sommes plutôt satisfaits de ce modèle au regard de ses coefficients. Gardons néanmoins à l'esprit la volatilité induite par la sous représentation de certaines modalités, qui biaise cette significativité.

De plus, nous avons implémenté une sélection par *AIC* avec méthode *stepwise*, grâce à la fonction *stepAIC* du package *MASS* pour étudier la pertinence de nos modalités. Cette procédure nécessite de retravailler nos données pour voir apparaître toutes les modalités en colonne et ainsi les traiter comme des variables. En se basant sur l'*AIC*, qui doit être minimisé, l'algorithme *stepwise* introduit dans le modèle MLG les modalités les unes après les autres pour estimer la significativité globale du modèle. Mais après chaque introduction de variable, l'algorithme réévalue la significativité des variables déjà introduites et si elles ne s'avèrent plus significatives, la moins significative d'entre elle est écartée. La procédure prend fin lorsqu'aucune des modalités ne peut être réintroduite ou retirée du modèle.

En implémentant cette méthode, la conclusion est qu'aucune modalité n'a d'intérêt à être écartée ou regroupée. Pour la zone 6 notamment, malgré sa faible exposition, elle apporte une information importante pour l'estimation de la sinistralité.

MLG Négatif binomial

Intéressons-nous maintenant à la sortie d'un MLG avec une distribution Négative Binomiale.

```
Coefficients:
                  Estimate Std. Error  t value Pr(>|t|)
(Intercept)       -2.90315    0.02457 -118.162  < 2e-16 ***
CD_ZONE_TARI_GP1  -0.28929    0.03076   -9.403  < 2e-16 ***
CD_ZONE_TARI_GP2  -0.09907    0.03204   -3.092  0.00199 **
CD_ZONE_TARI_GP4   0.16475    0.03815    4.318 1.59e-05 ***
CD_ZONE_TARI_GP5   0.27337    0.03576    7.645 2.30e-14 ***
CD_ZONE_TARI_GP6   0.13300    0.11343    1.173  0.24102
CD_FAMI_ATVT_RETE7 0.25677    0.03968    6.472 1.02e-10 ***
CD_FAMI_ATVT_RETE2 0.17489    0.06545    2.672  0.00755 **
CD_FAMI_ATVT_RETE3 0.20415    0.02985    6.840 8.40e-12 ***
CD_FAMI_ATVT_RETE6 0.08228    0.06162    1.335  0.18182
CD_FAMI_ATVT_RETE1 -0.31326   0.06356   -4.928 8.44e-07 ***
CD_FAMI_ATVT_RETE8 0.25563    0.03225    7.927 2.51e-15 ***
CD_FAMI_ATVT_RETE4 -0.30088   0.05561   -5.411 6.44e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(2) family taken to be 1.038723)

    Null deviance: 9141.1  on 9386  degrees of freedom
Residual deviance: 8602.1  on 9374  degrees of freedom
```

FIGURE 3.9 – Sortie R d'un MLG avec loi Négative Binomiale

Dans ce modèle, la zone 6 et l'activité 6 ne sont pas considérées comme significatives. La suite logique serait de regrouper ces variables avec leur modalité de référence. Cependant, nous avons également effectué un *step AIC* pour ce MLG et la conclusion de cette procédure est qu'aucune des modalités n'est à écarter car elles apportent toutes de l'information.

Lorsque les déviances ("*Null*" et "*Residual*") des deux MLG sont comparées, celles de la loi Négative Binomiale sont plus faibles. Cela signifie que le MLG avec loi Négative Binomiale prédit mieux la sinistralité avec un modèle où il n'y aurait que l'intercept (*Null deviance*), de même avec un modèle qui contient les variables prédictives (*Residual deviance*). Les coefficients sont toutefois assez proches entre les deux modèles. La différence la plus importante est sur l'estimation des coefficients de l'activité 3, la plus faible sur l'activité 6. En effet, sur l'activité 3, la fréquence estimée double entre le modèle Négatif Binomial et le Poisson. Elle passe de 6.8% à 12.9% par exemple sur le croisement activité 3 x zone 3. Mais sur le croisement activité 6 x zone 3, les fréquences estimées sont très proches, elles passent de 5.9% à 6.1%. En moyenne, le modèle de Poisson estime des fréquences légèrement plus élevées de 11% par rapport au modèle Négatif Binomial.

3.2.2 Modèles Zéro-Inflaté

Du fait du poids en 0 des distributions sinistres nous testons également des modèles de régression Zéro-Inflaté. Ces modèles sont construits de manière à capter le fort poids en 0. En effet, les modèles Zéro-Inflaté considèrent deux sources de 0 : les non-observés et les observés avec la valeur 0. Dans le cas d'étude de fréquence sinistre, la valeur 0 peut correspondre aux contrats non-sinistrés, des 0 observés donc. Mais elle peut également correspondre aux assurés sinistrés qui n'ont pas déclaré leur sinistre à l'assureur, ce cas est plus rare. De manière générale les non-observés correspondent à des données de censure.

Les modèles Zéro-Inflaté voient le jour en assurance au début des années 2000 avec notamment les études de LEE et *al.* (2002) et MELGAR et *al.* (2005), mais étaient déjà utilisés dans les années 90 dans d'autres secteurs, dans l'industrie notamment avec les travux de Diane LAMBERT (1992). Dans le cadre de ce mémoire nous nous sommes principalement penchés sur l'article de VASECHKO, GRUN-RÉHOMME et BENLAGHA (2009) [25] qui introduit ce type de modèle de manière empirique pour la modélisation de la fréquence sinistre en assurance.

Les modèles Zéro-Inflaté fonctionnent en deux parties :
- une distribution binomiale qui estime la masse en 0
- une distribution de comptage (Poisson ou Négative Binomiale par exemple) non tronquée, qui peut aussi prendre la valeur 0

La probabilité $\mathbf{P}(X = k)$ s'écrit donc comme suit, avec $X \geq 0$ une variable aléatoire discrète :

$$\mathbf{P}(X_i = k) = \begin{cases} p + (1-p) f_{count}(0) & pour\ k = 0 \\ (1-p) f_{count}(k) & pour\ k > 0 \end{cases}$$

Avec $f_{count}(k)$ la fonction de distribution de la loi de comptage considérée.

Sous R, la fonction *zeroinfl* du package *pscl* permet de mettre en place des modèles de régression Zéro-Inflaté. Nous l'implémentons en utilisant les lois de Poisson puis Négative Binomiale comme loi de comptage.

Zéro-Inflaté Poisson

En sortie de la fonction *zeroinfl*, deux modèles sont présentés : l'estimation de la loi de comptage et l'estimation de la masse en 0.

```
Count model coefficients (poisson with log link):
                   Estimate Std. Error  z value Pr(>|z|)
(Intercept)        -2.85625    0.01587 -179.996  < 2e-16 ***
CD_ZONE_TARI_GP1   -0.25080    0.02008  -12.492  < 2e-16 ***
CD_ZONE_TARI_GP2   -0.10884    0.02040   -5.336 9.52e-08 ***
CD_ZONE_TARI_GP4    0.12889    0.02204    5.848 4.99e-09 ***
CD_ZONE_TARI_GP5    0.24201    0.01980   12.225  < 2e-16 ***
CD_ZONE_TARI_GP6    0.12086    0.05907    2.046   0.0407 *
CD_FAMI_ATVT_RETE7  0.24100    0.02364   10.193  < 2e-16 ***
CD_FAMI_ATVT_RETE2  0.24834    0.04269    5.817 5.99e-09 ***
CD_FAMI_ATVT_RETE3  0.19105    0.01851   10.321  < 2e-16 ***
CD_FAMI_ATVT_RETE6  0.23992    0.03592    6.680 2.39e-11 ***
CD_FAMI_ATVT_RETE1 -0.22160    0.04530   -4.892 9.99e-07 ***
CD_FAMI_ATVT_RETE8  0.23023    0.01933   11.912  < 2e-16 ***
CD_FAMI_ATVT_RETE4 -0.26470    0.04349   -6.087 1.15e-09 ***

Zero-inflation model coefficients (binomial with logit link):
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -6.93696    0.21467 -32.315  < 2e-16 ***
CD_ZONE_TARI_GP1   -0.02194    0.23922  -0.092 0.926931
CD_ZONE_TARI_GP2   -0.01309    0.25593  -0.051 0.959211
CD_ZONE_TARI_GP4   -0.70776    0.41584  -1.702 0.088754 .
CD_ZONE_TARI_GP5   -1.33538    0.50523  -2.643 0.008215 **
CD_ZONE_TARI_GP6   -1.00693    1.02366  -0.984 0.325286
CD_FAMI_ATVT_RETE7 -0.59143    0.45177  -1.309 0.190486
CD_FAMI_ATVT_RETE2  1.29860    0.35908   3.616 0.000299 ***
CD_FAMI_ATVT_RETE3 -0.25755    0.27694  -0.930 0.352377
CD_FAMI_ATVT_RETE6  0.73963    0.38243   1.934 0.053108 .
CD_FAMI_ATVT_RETE1  1.05860    0.34452   3.073 0.002122 **
CD_FAMI_ATVT_RETE8 -0.26711    0.31395  -0.851 0.394877
CD_FAMI_ATVT_RETE4  0.65865    0.43723   1.506 0.131962
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 41
Log-likelihood: -1.783e+04 on 26 Df
```

FIGURE 3.10 – Sortie R d'un modèle Zéro-Inflaté avec une distribution de Poisson

Sur le modèle de comptage, les résultats de significativité des modalités sont assez satisfaisants. En revanche sur la modélisation de la masse en 0, ce n'est pas le cas. Le modèle binomial ne permet pas de capter la masse en 0. Pourtant nous avons un nombre de données a priori suffisant sur le poids en 0 et sur chacune des modalités. Les coefficients du modèle de comptage sont toutefois assez proches des modèles précédents.

Zéro-Inflaté Négatif Binomial

Les résultats du modèle Zéro-Inflaté avec une loi de comptage Négative Binomiale sont les suivants.

```
Count model coefficients (negbin with log link):
                   Estimate Std. Error  z value Pr(>|z|)
(Intercept)        -2.90108    0.02227 -130.287  < 2e-16 ***
CD_ZONE_TARI_GP1   -0.27243    0.02790   -9.764  < 2e-16 ***
CD_ZONE_TARI_GP2   -0.10965    0.02936   -3.734 0.000188 ***
CD_ZONE_TARI_GP4    0.15813    0.03422    4.621 3.81e-06 ***
CD_ZONE_TARI_GP5    0.26980    0.03186    8.469  < 2e-16 ***
CD_ZONE_TARI_GP6    0.13828    0.09937    1.392 0.164062
CD_FAMI_ATVT_RETE7  0.24823    0.03562    6.970 3.18e-12 ***
CD_FAMI_ATVT_RETE2  0.16387    0.06437    2.546 0.010902 *
CD_FAMI_ATVT_RETE3  0.21304    0.02699    7.894 2.93e-15 ***
CD_FAMI_ATVT_RETE6  0.12466    0.05630    2.214 0.026811 *
CD_FAMI_ATVT_RETE1 -0.26039    0.05954   -4.374 1.22e-05 ***
CD_FAMI_ATVT_RETE8  0.22805    0.02878    7.924 2.31e-15 ***
CD_FAMI_ATVT_RETE4 -0.26662    0.05061   -5.268 1.38e-07 ***
Log(theta)          1.01093    0.03566   28.350  < 2e-16 ***

Zero-inflation model coefficients (binomial with logit link):
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.079e+01  5.839e+02  -0.036    0.972
CD_ZONE_TARI_GP1   -1.358e+01  6.731e+02  -0.020    0.984
CD_ZONE_TARI_GP2    3.589e-02  1.422e+00   0.025    0.980
CD_ZONE_TARI_GP4   -1.713e+01  4.271e+03  -0.004    0.997
CD_ZONE_TARI_GP5   -1.606e+01  2.243e+03  -0.007    0.994
CD_ZONE_TARI_GP6   -2.059e+01  8.305e+04   0.000    1.000
CD_FAMI_ATVT_RETE7 -7.063e+00  1.193e+04  -0.001    1.000
CD_FAMI_ATVT_RETE2  1.441e+01  5.839e+02   0.025    0.980
CD_FAMI_ATVT_RETE3 -2.172e+00  1.155e+03  -0.002    0.998
CD_FAMI_ATVT_RETE6  3.977e+00  6.066e+02   0.007    0.995
CD_FAMI_ATVT_RETE1  1.393e+01  5.839e+02   0.024    0.981
CD_FAMI_ATVT_RETE8 -6.230e+00  7.212e+03  -0.001    0.999
CD_FAMI_ATVT_RETE4 -4.617e-01  3.887e+03   0.000    1.000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 2.7481
Number of iterations in BFGS optimization: 51
Log-likelihood: -1.644e+04 on 27 Df
```

FIGURE 3.11 – Sortie R d'un modèle Zéro-Inflaté avec une distribution Négative Binomiale

Ici aussi les résultats de la régression sur la distribution de comptage sont assez similaires aux sorties MLG, ce qui est rassurant. Cependant, le modèle binomial ne sort aucune modalité comme significative, pas même l'intercept.

Le maximum de vraisemblance est plus faible sur le modèle Négatif Binomial que sur le Poisson, ce qui nous indique qu'il estime tout de même mieux notre sinistralité que le modèle poissonnien.

De manière générale, nous ne sommes pas convaincus par les sorties de ces modèles, qui n'apportent a priori pas beaucoup plus d'information que les MLG avec des lois classiques.

3.2.3 Modèles *Hurdle*

Intéressons-nous maintenant à un deuxième type de modèle : les modèles *Hurdle*. Comme les modèles Zéro-Inflaté, ces modèles permettent d'estimer une masse en 0. Cependant, une seule source de 0 est considérée ici : les 0 effectivement observés. L'absence de sinistre et la non déclaration d'un sinistre sont donc considérés de la même manière.

Ces modèles sont notamment l'objet des travaux de John MULLAHY (1986) [11] et de ZEILIS, KLEIBER et JACKMAN (2008) [27] pour leur application sous R.

Ces modèles comptent également deux composantes qui sont :
- une distribution binomiale qui estime la masse en 0
- une distribution de comptage tronquée en 0

en reprenant les mêmes notations que précédemment :

$$\mathbf{P}(X = k) = \begin{cases} p & pour\ k = 0 \\ (1-p)\frac{f_{count}(k)}{1-f_{count}(0)} & pour\ k > 0 \end{cases}$$

Pour l'implémentation d'un tel modèle sous R, nous utilisons la fonction *Hurdle* du package *pscl* qui fonctionne de manière identique à la fonction *zeroinfl* vue plus tôt.

Hurdle Poisson

Ci-dessous, les résultats d'un modèle *Hurdle* avec une loi de Poisson comme loi de comptage :

```
Count model coefficients (truncated poisson with log link):
                   Estimate Std. Error  z value Pr(>|z|)
(Intercept)        -2.78415    0.01653 -168.430  < 2e-16 ***
CD_ZONE_TARI_GP1   -0.29713    0.02099  -14.154  < 2e-16 ***
CD_ZONE_TARI_GP2   -0.15835    0.02113   -7.494 6.67e-14 ***
CD_ZONE_TARI_GP4    0.09057    0.02229    4.064 4.82e-05 ***
CD_ZONE_TARI_GP5    0.18368    0.02021    9.090  < 2e-16 ***
CD_ZONE_TARI_GP6    0.09476    0.06530    1.451    0.147
CD_FAMI_ATVT_RETE7  0.26117    0.02441   10.698  < 2e-16 ***
CD_FAMI_ATVT_RETE2  0.24629    0.04546    5.418 6.03e-08 ***
CD_FAMI_ATVT_RETE3  0.20494    0.01922   10.663  < 2e-16 ***
CD_FAMI_ATVT_RETE6  0.16409    0.03661    4.482 7.38e-06 ***
CD_FAMI_ATVT_RETE1 -0.24006    0.04908   -4.891 1.00e-06 ***
CD_FAMI_ATVT_RETE8  0.23040    0.01995   11.549  < 2e-16 ***
CD_FAMI_ATVT_RETE4 -0.28418    0.04880   -5.824 5.76e-09 ***
Zero hurdle model coefficients (binomial with logit link):
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.49850    0.05252 -47.569  < 2e-16 ***
CD_ZONE_TARI_GP1   -0.22303    0.06555  -3.402 0.000668 ***
CD_ZONE_TARI_GP2   -0.17132    0.06937  -2.470 0.013523 *
CD_ZONE_TARI_GP4    0.01850    0.08823   0.210 0.833905
CD_ZONE_TARI_GP5    0.47357    0.08891   5.326 1.00e-07 ***
CD_ZONE_TARI_GP6    0.20868    0.28834   0.724 0.469245
CD_FAMI_ATVT_RETE7  0.40553    0.09715   4.174 2.99e-05 ***
CD_FAMI_ATVT_RETE2 -0.09134    0.13556  -0.674 0.500442
CD_FAMI_ATVT_RETE3  0.30412    0.06771   4.491 7.08e-06 ***
CD_FAMI_ATVT_RETE6 -0.06272    0.14060  -0.446 0.655508
CD_FAMI_ATVT_RETE1 -0.52514    0.11588  -4.532 5.85e-06 ***
CD_FAMI_ATVT_RETE8  0.27117    0.07593   3.571 0.000356 ***
CD_FAMI_ATVT_RETE4 -0.39077    0.09428  -4.145 3.40e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 19
Log-likelihood: -1.771e+04 on 26 Df
```

FIGURE 3.12 – Sortie R d'un modèle *Hurdle* avec une distribution de Poisson

Les résultats sur la partie comptage sont relativement similaires aux résultats précédents des modèles linéaires généralisés et Zéro-Inflaté. La masse en 0 cependant semble bien mieux modélisée ; bien plus de modalités sont significatives et le maximum de vraisemblance est également plus faible.

Hurdle Négatif Binomial

Sur le modèle *Hurdle* avec une loi Négative Binomiale, nous pouvons tirer les mêmes conclusions d'après la sortie suivante.

```
Count model coefficients (truncated negbin with log link):
                   Estimate Std. Error  z value Pr(>|z|)
(Intercept)        -2.86653    0.02731 -104.969  < 2e-16 ***
CD_ZONE_TARI_GP1   -0.32823    0.03380   -9.712  < 2e-16 ***
CD_ZONE_TARI_GP2   -0.17589    0.03503   -5.021 5.15e-07 ***
CD_ZONE_TARI_GP4    0.12006    0.03977    3.019 0.002539 **
CD_ZONE_TARI_GP5    0.19866    0.03610    5.503 3.73e-08 ***
CD_ZONE_TARI_GP6    0.17608    0.11753    1.498 0.134075
CD_FAMI_ATVT_RETE7  0.25462    0.04171    6.104 1.03e-09 ***
CD_FAMI_ATVT_RETE2  0.22785    0.07381    3.087 0.002022 **
CD_FAMI_ATVT_RETE3  0.23381    0.03183    7.346 2.04e-13 ***
CD_FAMI_ATVT_RETE6  0.12595    0.06561    1.920 0.054918 .
CD_FAMI_ATVT_RETE1 -0.26393    0.07459   -3.539 0.000402 ***
CD_FAMI_ATVT_RETE8  0.25589    0.03405    7.516 5.67e-14 ***
CD_FAMI_ATVT_RETE4 -0.33615    0.06829   -4.922 8.57e-07 ***
Log(theta)          1.00212    0.04785   20.943  < 2e-16 ***
Zero hurdle model coefficients (binomial with logit link):
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.49850    0.05252 -47.569  < 2e-16 ***
CD_ZONE_TARI_GP1   -0.22303    0.06555  -3.402 0.000668 ***
CD_ZONE_TARI_GP2   -0.17132    0.06937  -2.470 0.013523 *
CD_ZONE_TARI_GP4    0.01850    0.08823   0.210 0.833905
CD_ZONE_TARI_GP5    0.47357    0.08891   5.326 1.00e-07 ***
CD_ZONE_TARI_GP6    0.20868    0.28834   0.724 0.469245
CD_FAMI_ATVT_RETE7  0.40553    0.09715   4.174 2.99e-05 ***
CD_FAMI_ATVT_RETE2 -0.09134    0.13556  -0.674 0.500442
CD_FAMI_ATVT_RETE3  0.30412    0.06771   4.491 7.08e-06 ***
CD_FAMI_ATVT_RETE6 -0.06272    0.14060  -0.446 0.655508
CD_FAMI_ATVT_RETE1 -0.52514    0.11588  -4.532 5.85e-06 ***
CD_FAMI_ATVT_RETE8  0.27117    0.07593   3.571 0.000356 ***
CD_FAMI_ATVT_RETE4 -0.39077    0.09428  -4.145 3.40e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta: count = 2.7241
Number of iterations in BFGS optimization: 21
Log-likelihood: -1.642e+04 on 27 Df
```

FIGURE 3.13 – Sortie R d'un modèle *Hurdle* avec une distribution Négative Binomiale

Ici encore le modèle avec une distribution Négative Binomiale semble plus adapté que la loi de Poisson.

Des modèles avec prédiction de la masse en 0, notre modèle qui paraît le plus adapté, d'après les premières sorties R est le modèle *Hurdle* avec loi de comptage Négative Binomiale. Cette conclusion est en accord avec nos précédentes analyses : la distribution des sinistres est Négative Binomiale. La masse en 0 semble mieux s'expliquer par une unique source de 0 Pour valider ces résultats et les comparer avec les modèles linéaires généralisés, nous étudierons d'autres indicateurs de validation de modèles dans la partie 3.2.5.

3.2.4 Approche *Machine Learning* : les forêts aléatoires

Les forêts aléatoires sont des méthodes de *Machine Learning* dont le principe est d'agréger plusieurs arbres décisionnels. Les arbres CART sont des algorithmes d'apprentissage qui permettent de construire une classification hiérarchique, d'où leur nomination d'arbres décisionnels. L'arbre de décision part d'un premier noeud racine et se divise pour créer des points de décision, des noeuds, qui sont des questions auxquelles les branches répondent. Quand l'arbre ne peut plus créer de noeuds, un noeud terminal, appelé aussi feuille, est atteint. Il représente la décision finale. Par exemple dans notre cas, un noeud pourrait-être "Le contrat appartient-il à la zone 1 ?" si oui, nous continuons sur la branche de gauche, sinon sur celle de droite. En descendant l'arbre nous arriverons à une feuille qui prédira le nombre de sinistre selon les critères se trouvant sur les différents noeuds précédents. La figure suivante schématise la construction d'un arbre.

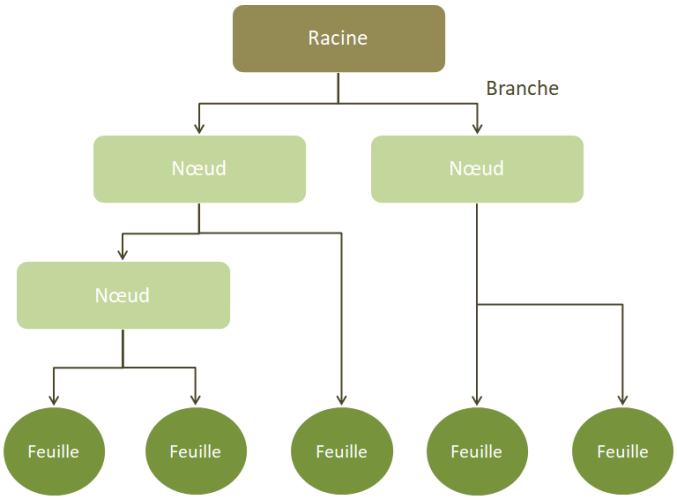

FIGURE 3.14 – Schéma d'un arbre de décision

Ces arbres sont des modèles non-linéaires particulièrement efficaces, notamment du fait qu'ils peuvent capter des interactions entre les variables, à l'inverse des MLG par exemple qui n'acceptent que des variables décorrélées. Cependant, l'efficacité de ces arbres est aussi suivie par une instabilité et du sur-apprentissage. En effet, les modèles prédits peuvent fortement varier selon la base *train* sur laquelle se base l'algorithme.

C'est là qu'interviennent les forêts aléatoires. Ces méthodes consistent en l'agrégation de plusieurs arbres en intégrant de l'aléa, notamment un système de ré-échantillonage avec remise (*bootstrap*). Les arbres alors construits sont décorrélés entre eux et leurs sorties sont utilisées pour créer un modèle unique, plus stable. Ces forêts aléatoires sont largement utilisées en *Machine Learning* pour diverses études et sont réputées très efficaces mais également particulièrement robustes en comparaison d'autres méthodes, c'est pourquoi c'est le modèle de *Machine Learning* que nous privilégions ici pour prédire nos sinistres et pour comparer ces prédictions aux modèles précédents.

Pour implémenter une forêt sous R, nous avons à notre disposition la fonction *randomForest* du package du même nom. Comme pour les modèles précédents, il est nécessaire de renseigner la variable cible, les variables explicatives, le jeu de données et aussi les indices de la base d'apprentissage. La fonction *randomForest* présente également d'autres paramètres qu'il est possible de tuner afin d'optimiser la forêt aléatoire. Nous pouvons notamment regarder les paramètres *ntree*, le nombre d'arbres dans la forêt, et *mtry*, le nombre de co-variables tirées aléatoirement à chaque *split* (noeud), c'est ce qui permet de décorréler les arbres. Nous avons dans notre jeu de données 2 variables, la zone et l'activité. C'est pourquoi nous fixons le paramètre *mtry* à 1. Les paramètres *nodesize*, qui détermine le nombre minimal d'individus dans une feuille, et *maxnodes*, le nombre maximal d'individus dans une feuille, déterminent la profondeur des arbres et nous fixons ces paramètres par défaut. Ainsi nous avons au moins 5 individus, mais pas de maximum. Le *tuning* du paramètre *ntree* est mis en place par Validation Croisée. Nous cherchons la valeur du paramètre qui minimise l'erreur de prédiction.

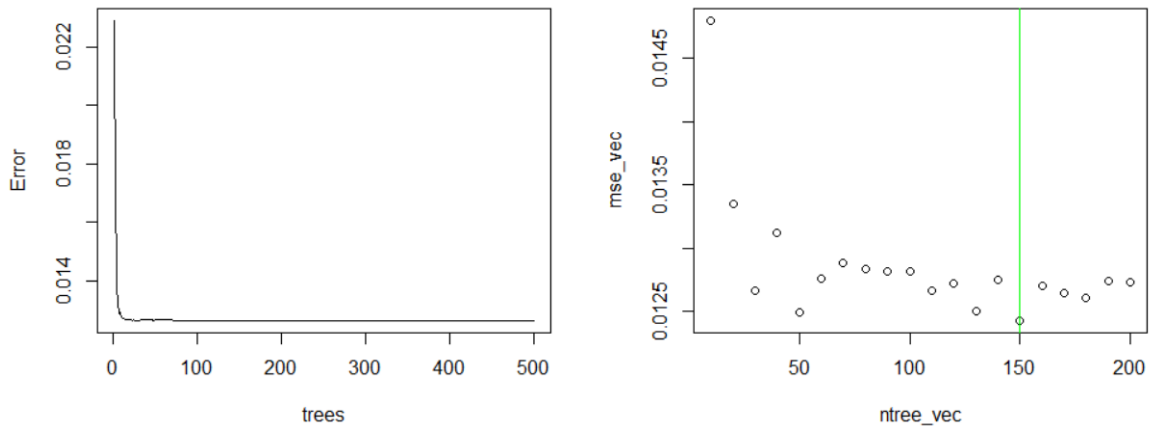

FIGURE 3.15 – Optimisation du paramètre *ntree* par étude de l'erreur *OOB* (gauche) et du *MSE* (droite)

Le nombre d'arbre par défaut est de 500, nous choisissons dans un premier temps d'étudier un premier *plot* de notre *randomForest*, qui indique l'évolution du nombre d'erreur *OOB (Out Of Bag)* en fonction du nombre d'arbres (figure de gauche). L'erreur *OOB* est pour nous un moyen de valider un modèle de forêt aléatoire. Lors de l'échantillonnage *bootstrap* pour mettre en place les arbres, les lignes qui ne sont pas utilisées pour entraîner un arbre sont alors *out of bag* (hors échantillon). Elles sont ensuite utilisées comme un échantillon de test. Chaque arbre non entraîné sur une ligne va prédire cette ligne en fonction de son modèle construit et la prédiction finale de la ligne sera la moyenne de l'ensemble des prédictions. L'erreur *OOB* est donc la différence entre ces prédictions et les données et nous cherchons à la minimiser. Dans notre cas, l'erreur *OOB* a plutôt tendance à être minimisée et à se stabiliser entre 0 et 100. L'erreur *OOB* décroît très vite et l'échelle ne nous permet pas de tirer de bonnes conclusions visuellement. Pour préciser le nombre d'arbres optimal, nous nous restreignons donc à l'intervalle [0,200]. Sur cet intervalle, l'objectif est de sélectionner le nombre d'arbre qui minimise le *MSE* (figure de droite). Le *MSE* (*Mean Squared Error*) ou erreur quadratique moyenne en français est une autre mesure de l'erreur de prédiction et se calcule simplement de la manière suivante :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(X_i^{pred} - X_i^{obs})^2$$

Le *ntree* optimal pour ce modèle est de 150, représenté par la droite verticale verte.

En appelant notre modèle avec les paramètres choisis plus tôt, nous avons la sortie suivante :

```
Call:
 randomForest(formula = freq_hg_DTA ~ CD_ZONE_TARI_GP + CD_FAMI_ATVT_RETE,
   data = base_mlg_DTA, ntree = ntree_opt, subset = indice_train)
               Type of random forest: regression
                     Number of trees: 150
No. of variables tried at each split: 1

          Mean of squared residuals: 0.01265724
                    % Var explained: 1.67
```

FIGURE 3.16 – Sortie d'un modèle *randomForest*

D'après cette sortie, le modèle de forêt aléatoire n'explique que 1.67% de la variance cible. Ce pourcentage est très faible et indique que ce modèle n'est peut-être pas bien *fitté*. Pour s'en assurer, nous le comparerons à nos autres modèles de prédiction.

3.2.5 Comparaison et choix des modèles

Afin de déterminer le modèle qui prédit le mieux notre sinistralité, nous allons comparer différents indicateurs en sortie des modèles qui ont pu être implémentés plus tôt.

En notant k le nombre de paramètres de notre modèle et n le nombre d'observations, nous retrouvons les indicateurs :

— la log vraisemblance :

$$log\left(\mathcal{L}\left(\beta,(x,y)_i\right)\right) = \prod_{i=1}^{n}\mathbf{P}\left(X = x_i | Y = y_i\right)$$

— l'*AIC* (*Akaike's Information Criterion*) :

$$AIC = 2k - 2log\left(\mathcal{L}\right)$$

— le *BIC* (*Bayesian Information Criterion*) :

$$BIC = -2log\left(\mathcal{L}\right) + k \times log(n)$$

— la déviance :

$$D = -2\left(log(\mathcal{L}) - log(\mathcal{L}_{sat})\right)$$

Avec $\mathcal{L}_{sat}$ la vraisemblance du modèle saturé, soit le modèle qui prédit parfaitement la base d'apprentissage, qui pour chaque observation utilise un paramètre.

L'ensemble de ces indicateurs doit être minimisé pour obtenir un meilleur modèle. Nous cherchons donc le modèle qui a les indicateurs les plus faibles.

En sortie des modèles de régression, il est simple de récupérer un certain nombre d'indicateurs, comme ci-dessous.

	MLG		Zéro-Inflaté		*Hurdle*	
	Poisson	Négatif Binomial	Poisson	Négatif Binomial	Poisson	Négatif Binomial
loglik	17 667	16 334	17 525	16 361	17 447	16 276
AIC	35 360	32 694	35 103	32 776	34 946	32 605
BIC	35 453	32 787	35 289	32 969	35 132	32 798
deviance	17 327	8 665	/	/	/	/

TABLE 3.2 – Indicateurs de performance des modèles de régression

Pour tous les types de modèles implémentés, les meilleurs modèles sont ceux dont la distribution sous-jacente est la distribution Négative Binomiale. Cette conclusion va de paire avec le fait que nous avons approché nos sinistres par une loi Négative Binomiale. Parmi les trois types de modèles, le moins performant est le modèle Zéro-Inflaté. Entre les modèles MLG et *Hurdle*, le choix est moins évident, les statistiques sont assez proches, relativement plus faible sur le modèle *Hurdle*, mais gardons en tête que ces statistiques peuvent varier légèrement selon la base d'apprentissage choisie.

Penchons-nous maintenant sur les indicateurs que nous avons calculés par Validation Croisée :

	MLG		Zéro-Inflaté		*Hurdle*		Forêt
	Poisson	Négatif Binomial	Poisson	Négatif Binomial	Poisson	Négatif Binomial	/
MSE	7.774	7.702	7.936	7.824	7.777	7.803	7.784
écart au total	711	291	910	1044	945	798	735

TABLE 3.3 – Indicateurs de performance des modèles par Validation Croisée

Nous nous intéressons cette fois aux critères :
- le *MSE* (*Mean Squared Error*), l'erreur quadratique moyenne
- l'écart au total, la différence entre le nombre de sinistres total observé et le nombre de sinistres total prédit, pour référence le total moyen sur la base test est de 8 758 :

$$abs(\sum_{i=1}^{n} X_i^{pred} - \sum_{i=1}^{n} X_i^{obs})$$

Au regard de ces critères, le meilleur modèle semble être le MLG Négatif Binomial. Les *MSE* sont toutefois assez similaires. La différence la plus flagrante est sur l'écart au total, ou le MLG Négatif Binomial se démarque. Aussi, nous introduisons ici à la comparaison les forêts aléatoires. Les résultats apportés par ce modèle sont plutôt proches des résultats des autres modèles implémentés.

Pour distinguer encore nos modèles, nous pouvons étudier leur adéquation à la distribution de nos sinistres observés. À ce stade, les modèles qui vont nous intéresser et que nous allons comparer sont les modèles : MLG Négatif Binomial, *Hurdle* Négatif Binomial et la forêt aléatoire. Pour les comparer, nous pouvons étudier la dispersion des fréquences observées et modélisées comme sur l'histogramme ci-dessous.

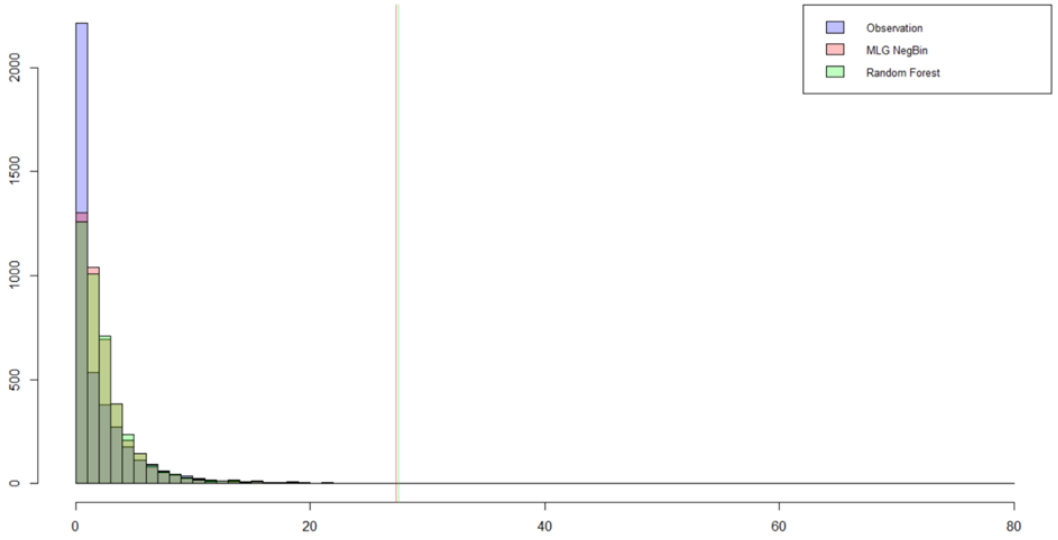

FIGURE 3.17 – Histogramme de comparaison entre les observations, le MLG et la forêt aléatoire

D'après cet histogramme, les MLG (en rouge) et forêts aléatoires (en vert) sont très similaires. Les droites verticales à environ 30 sinistres représentent le maximum des sinistres prédits par les deux méthodes. La masse en 0 est un peu mieux prédite par le modèle MLG et son maximum est très légèrement supérieur. Bien qu'il est difficile de faire un choix entre les deux modèles via ce graphique, nous remarquons toutefois que la masse en 0 n'est pas si bien capté, de même pour les sinistres au dessus de 30, ils ne sont pas modélisés.

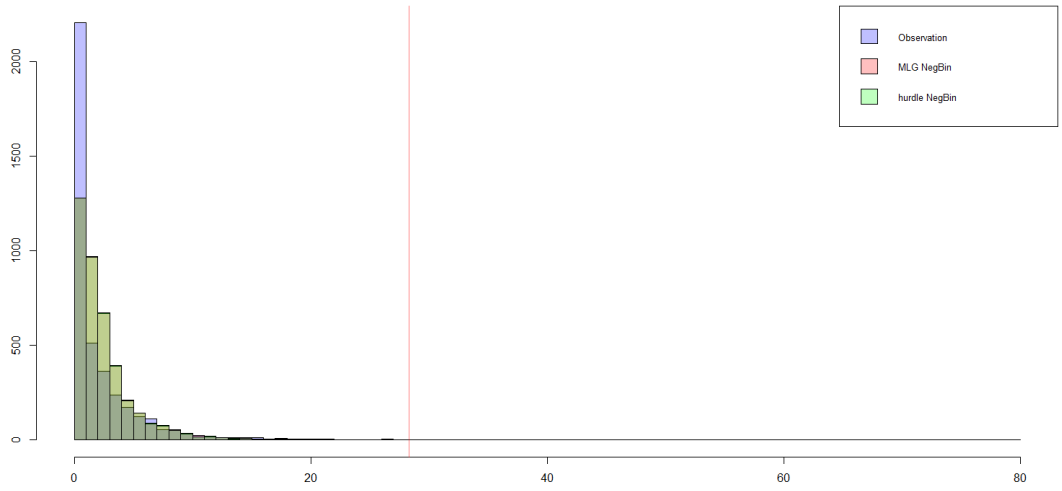

FIGURE 3.18 – Histogramme de comparaison entre les observations, le MLG et le modèle *Hurdle*

Sur ce graphique, les histogrammes se superposent quasi-parfaitement, même en se penchant plus en détail sur les queues de distribution. Il est difficile de départager les modèles à partir de ce simple graphique. Notons toutefois que le modèle *Hurdle* ne capte pas mieux la masse en 0 que le MLG simple.

Après analyse de ces différents critères, nous faisons le choix de modéliser notre sinistralité via un MLG Négatif Binomial. En effet, ce modèle est parmi nos meilleurs modèles implémentés. Il a également l'avantage non négligeable d'être très simple d'implémentation et de compréhension. En effet, nous pourrons notamment être amenés à retoucher des coefficients pour intégrer certaines contraintes, comme nous le verrons par la suite dans la partie 3.4. Le modèle *Hurdle* Négatif Binomial pourrait également être considéré, mais n'étant pas plus efficace que le MLG, notamment sur la prédiction de la masse en 0 et étant plus complexe car faisant apparaître deux fois plus de coefficients, nous décidons de ne pas le conserver.

Nous avons donc le modèle suivant :

Critère	β	p-value
Intercept	-2.90315	$<$ 2e-16
Zone 1	-0.28929	$<$ 2e-16
Zone 2	-0.09907	0.00199
Zone 4	0.16475	1.59e-05
Zone 5	0.27337	2.30e-14
Zone 6	0.13300	0.24102
Activité 1	-0.31326	8.44e-07
Activité 2	0.17489	0.00755
Activité 3	0.20415	8.40e-12
Activité 4	-0.30088	6.44e-08
Activité 6	0.08228	0.18182
Activité 7	0.25677	1.02e-10
Activité 8	0.25563	2.51e-15

TABLE 3.4 – Modèle choisi : MLG Négatif Binomial

3.3 Sensibilité des modèles

En validant nos modèles par méthode de Validation Croisée, nous nous assurons de sélectionner le plus adapté à notre problématique. Cependant, nous devons aussi prêter attention à la volatilité de ces modèles. Si en moyenne un modèle de prédiction peut paraître performant, il se peut que sur un nouvel échantillon de notre base de donnée il soit très décalé par rapport à cette moyenne. Grâce à la Validation Croisée nous pouvons donc évaluer la volatilité des modèles, par exemple en étudiant la volatilité de l'erreur quadratique.

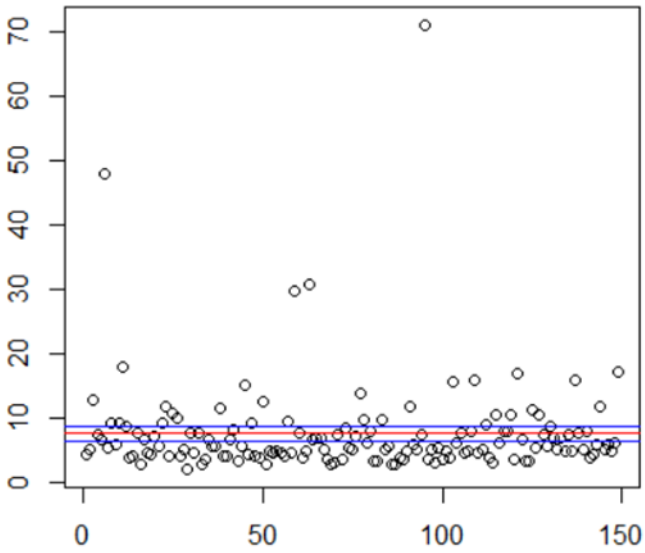

FIGURE 3.19 – *MSE* sur 149 échantillons de Validation Croisée et intervalle de confiance à 95%

Avec un *MSE* moyen (μ_{MSE}) de 7.72 et un écart type autour de cette moyenne (σ_{MSE}) de 7.53 nous pouvons construire l'intervalle de confiance autour de la moyenne :

$$IC_{1-\alpha} = \left[\mu_{MSE} \pm t \times \frac{\sigma_{MSE}}{\sqrt{K}}\right]$$

Avec $K = 149$, $\alpha = 0.05$ et t tel que $\Phi(t) = 1 - \frac{\alpha}{2}$, Φ la fonction de répartition de la loi normale centrée réduite.

Nous comptons 75% des *MSE* calculés sur les différents échantillons en dehors de l'intervalle à 95% autour de la moyenne des *MSE*. Ce taux est très élevé et indique une certaine volatilité sur ce modèle. En étudiant de plus près les *MSE* les plus élevés, il s'agit des échantillons dans lesquels se trouvent les nombres de sinistres les plus élevés. En effet, nous comptons jusqu'à 80 sinistres pour un seul contrat, mais ces valeurs élevées ont du mal à être captées par le modèle et viennent biaiser nos indicateurs de performance. Comme nous avons pu le constater sur l'histogramme 3.18 par exemple, nos modèles ne prédisent pas plus d'une trentaine de sinistres.

Malgré ces défauts, nous choisissons de conserver notre modèle pour prédire la fréquence collective, ces sinistres importants étant exceptionnels et pouvant être gérés particulièrement. Notamment grâce à l'intervention de la crédibilité, qui va permettre d'ajuster la fréquence sinistre en prenant en compte l'historique plus ou moins important des différents contrats.

3.4 Retraitement du modèle final

Comme nous pouvons le voir dans le tableau 3.4, les coefficients β des zones 1 à 5 prédits sont bien croissants. Ce constat est bien en accord avec nos observations du marché que nous pouvons retrouver sur les graphiques 2.18 et 2.19. Nous ne retrouvons cependant pas cette croissance entre les zones 5 et 6 dans notre modèle. Du fait de cette contrainte métier, nous souhaitons rehausser manuellement le coefficient β lié à la zone 6. Cette correction permet d'assurer la modélisation de l'augmentation de sinistralité dans cette zone et ainsi de conserver la cohérence souhaitée dans les zones.

Afin d'évaluer ce coefficient β_{zone6}, nous nous basons sur la croissance des coefficients des zones en sortie de notre modèle MLG.

Zone	β	Croissance
Zone 1	-0.28929	
Zone 2	-0.09907	0.19
Zone 3	0	0.099
Zone 4	0.16475	0.165
Zone 5	0.27337	0.109
Zone 6	0.13300	

TABLE 3.5 – Évolution des coefficients des zones

La croissance est assez stable d'une zone à l'autre, de moyenne 0.14067. Nous pourrions appliquer cette moyenne pour obtenir un coefficient de zone 6 de 0.41404, mais regardons dans un premier temps d'autres appréciations de cette croissance. À l'aide d'une régression linéaire et d'une régression polynomiale de degré 2, de R^2 égal à 0.9902 et 0.9937 respectivement, le β_{zone6} est prédit à 0.4266 et à 0.3766. La moyenne de ces trois coefficients est utilisée pour estimer la valeur finale du coefficient que nous cherchons : 0.40575.

Notre modèle devient :

Critère	β
Intercept	-2.90315
Zone 1	-0.28929
Zone 2	-0.09907
Zone 3	0
Zone 4	0.16475
Zone 5	0.27337
Zone 6	0.40575
Activité 1	-0.31326
Activité 2	0.17489
Activité 3	0.20415
Activité 4	-0.30088
Activité 5	0
Activité 6	0.08228
Activité 7	0.25677
Activité 8	0.25563

TABLE 3.6 – Modèle final de prédiction de la fréquence collective DTA

À partir de ces coefficients, nous pouvons retrouver la fréquence collective prédite sur chacun des croisements zone x activité avec la formule :

$$Freq^{th} = e^{\beta_0 + \beta_{activit} + \beta zone}$$

Chapitre 4

Révision des coefficients de crédibilité

4.1 Objectifs

La deuxième composante de la fréquence est le coefficient de crédibilité z qu'il reste à déterminer maintenant que nous avons un modèle pour la fréquence théorique.

$$Freq^{cred} = z \times Freq^{obs} + (1 - z) \times Freq^{th}$$

Dans le tarif actuellement mis en place, nous avons une expression explicite de ce z, sous forme de formule. Pour plus de détails, se référer à la partie 1.4. C'est pourquoi nous souhaitons à la fin de cette étude avoir une expression explicite du coefficient de crédibilité, afin de simplifier l'implémentation dans le tarificateur Excel. Pour obtenir cette formule, nous souhaitons tout d'abord partir des théories de la crédibilité classiques afin de les étudier.

Dans un soucis de fluidité, les démonstrations des théorèmes présentés ci-dessous ne seront pas décrites dans le corps de ce document. Elles peuvent néanmoins être retrouvées dans le chapitre 4 de *A Course in Credibility Theory and its Applications* (2005) de Bühlmann et Gisler[4].

Nous nous concentrons ici aussi uniquement sur la garantie DTA.

4.2 Le modèle actuel

Pour rappel, le coefficient de crédibilité actuellement mis en place a la forme suivante :

$$z_{DTA} = \sqrt{\frac{n_{DTA} - c}{d - c}}$$

Sous forme graphique avec le dernier taux de moteurs DTA en exposition, cela prend la forme :

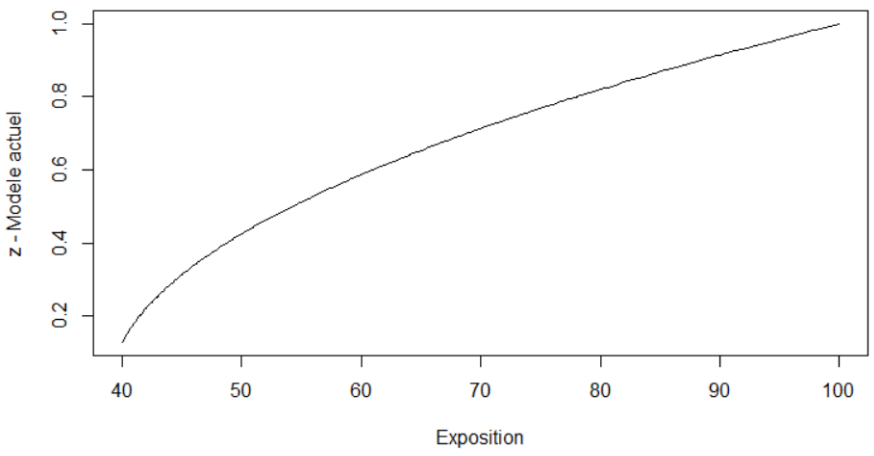

FIGURE 4.1 – Facteur de crédibilité DTA actuellement mis en place en fonction du dernier taux de moteurs année

Ce coefficient est forcé à 0 et à 1 aux bornes qui définissent le segment 3. La croissance de z est plus forte sur les petits moteurs que sur les grands qui sont plus individualisés.

La première remarque que nous pouvons faire porte sur l'exposition. En effet, n'est utilisée pour déterminer ce coefficient de crédibilité que l'information de la dernière année. Cependant, intuitivement nous ne pouvons pas accorder le même poids à un contrat de 4 ans d'expérience, qu'à un contrat avec le même nombre de moteurs annuel mais qui ne dispose que d'un an d'expérience. C'est pourquoi dans la suite nous nous penchons sur deux modèles de crédibilité qui prennent en compte l'expérience, à savoir les modèles de Bühlmann-Straub et de Jewell.

4.3 Le modèle de Bühlmann-Straub

4.3.1 Rappels de la théorie

La théorie de la crédibilité voit le jour au début du XXe siècle avec Mowbray (1914) et Whitney (1918) qui interprète la prime de risque comme une moyenne pondérée de l'expérience de l'individu et celle d'une classe de risque. Depuis, de nombreuses théories en découlent, à commencer par celle de Bühlmann (1967) [9] qui propose une formule explicite du coefficient de crédibilité, présentée plus bas.

Les notations

Dans la suite, un certain nombre de notations vont être utilisées, elles sont introduites ici.
- $i \in [1, I]$: le nombre d'individus, de contrats dans notre cas
- $j \in [1, n]$: les années d'observations
- Θ_i : le risque de l'individu i, une variable aléatoire
- X_{ij} : la fréquence sinistre pour le contrat i sur l'année j

Ainsi pour un contrat Θ_i, nous associons un historique de sinistres $(X_{i1}, X_{i2}, X_{i3}, X_{i4})$, puisque nous possédons les informations sinistres sur les 4 dernières années.

Le premier modèle de Bühlmann

Avant d'introduire le modèle de Bühlmann-Straub, commençons par un rappel sur le modèle de Bühlmann, modèle qui sert de base à la généralisation de Bühlmann-Straub.

Ce modèle a été introduit en 1967 et suppose deux hypothèses fortes :

(H1) Les variables aléatoires X_{ij} sont, conditionnellement à $\Theta_i = \theta$, indépendantes et identiquement distribuées selon une loi F_θ. Nous notons ses moments conditionnels d'ordre 1 et 2 :

$$\mu(\theta) = \mathbf{E}\left[X_{ij} | \Theta_i = \theta\right]$$
$$\sigma^2(\theta) = \mathbf{V}\left(X_{ij} | \Theta_i = \theta\right)$$

(H2) Les couples (Θ_i, X_i) sont indépendants et identiquement distribués.

Théorème 1 *Modèle de Bühlmann. Sous (H1) et (H2), l'estimateur homogène de crédibilité est donnée par :*

$$\widehat{\widehat{\mu(\Theta_i)}} = z \times \bar{X}_i + (1 - z) \times \widehat{\widehat{\mu_0}}$$

Avec :
- $\bar{X}_i = \frac{1}{n} \sum_{j=1}^{n} X_{ij}$ *: l'estimateur de la fréquence individuelle*
- $\mu_0 = \mathbf{E}[\mu(\Theta_i)] = \mathbf{E}[X_{ij}]$ *: la fréquence collective et* $\widehat{\widehat{\mu_0}}$ *son estimateur*

Le coefficient de crédibilité est le suivant :

$$z = \frac{n}{n + \frac{\sigma^2}{\tau^2}}$$

Avec :
- $\sigma^2 = \mathbf{E}[\sigma^2(\Theta_i)]$ *: la variabilité interne du risque*
- $\tau^2 = \mathbf{V}(\mu(\Theta_i))$ *: l'hétérogénéité du portefeuille*

L'enjeu maintenant est de déterminer $K = \frac{\sigma^2}{\tau^2}$ pour obtenir une expression de z et donc l'estimateur $\widehat{\widehat{\mu(\Theta_i)}}$. Pour cela, nous calculons les estimateurs sans biais et convergents suivants :

— $\widehat{\sigma^2} = \frac{1}{I(n-1)} \sum_{i=1}^{I} \sum_{j=i}^{n} \left(X_{ij} - \bar{X}_i\right)^2$

— $\widehat{\tau^2} = \frac{1}{I-1} \sum_{i=1}^{I} \left(\bar{X}_i - \bar{X}\right)^2 - \frac{\widehat{\sigma^2}}{n}$

Du fait que l'estimateur de crédibilité soit sans biais, la prime crédibilisée assure que les primes perçues couvrent correctement les sinistres de l'assuré.

La première hypothèse de Bühlmann (H1) est très forte et même non observable en pratique. En effet, un même contrat va évoluer dans le temps, son exposition va changer. Par exemple dans le cadre des flottes, des véhicules vont être amenés à entrer ou sortir du contrat, les garanties peuvent changer également, nous avons donc une forte variabilité de l'exposition dans le temps. La sinistralité d'une année à l'autre n'est donc pas identiquement distribuée pour un même contrat.

À partir de ce constat, les travaux de Bühlmann et Straub permettent de se défaire de cette hypothèse.

Le modèle de Bühlmann Straub

Ce modèle permet, grâce à l'introduction de poids, de prendre en compte l'exposition d'un contrat sur chaque année et sa variation. Ces poids sont notés :

— ω_{ij} : l'exposition du contrat i sur l'année j

Les hypothèses du modèle deviennent :

(H1) Les variables aléatoires X_{ij} sont, conditionnellement à Θ_i, indépendantes. Nous notons ses moments conditionnels d'ordre 1 et 2 :

$$\mu(\Theta_i) = \mathbf{E}\left[X_{ij}|\Theta_i\right]$$
$$\sigma^2(\Theta_i) = \omega_{ij}\mathbf{V}\left(X_{ij}|\Theta_i\right)$$

(H2) Les couples (Θ_i, X_i) sont indépendants et identiquement distribués.

Théorème 2 *Modèle de Bühlmann-Straub. Sous (H1) et (H2) l'estimateur non-homogène de crédibilité est donné par :*

$$\widehat{\widehat{\mu(\Theta_i)}} = z_i \times \bar{X}_i + (1 - z_i) \times \widehat{\widehat{\mu_0}}$$

Avec :

— $\bar{X}_i = \sum_{j=1}^{n} \frac{\omega_{ij}}{\omega_{i\bullet}} X_{ij}$

— $\widehat{\widehat{\mu_0}} = \sum_{i=1}^{I} \frac{z_i}{z_\bullet} \bar{X}_i$

— $\omega_{i\bullet} = \sum_{j=1}^{n} \omega_{ij}$

Le coefficient de crédibilité est calculé de la manière suivante :

$$z_i = \frac{\omega_{i\bullet}}{\omega_{i\bullet} + \frac{\sigma^2}{\tau^2}}$$

Les estimateurs pour la détermination de $K = \frac{\sigma^2}{\tau^2}$ sont les suivants :

— $\widehat{\sigma^2} = \frac{1}{I(n-1)} \sum_{i=1}^{I} \sum_{j=i}^{n} \omega_{ij} \left(X_{ij} - \bar{X}_i\right)^2$

— $\widehat{\widehat{\tau^2}} = \frac{\omega_{\bullet\bullet}}{\omega_{\bullet\bullet}^2 - \sum_i \omega_{i\bullet}^2} \left(\sum_{i=1}^{I} \omega_{i\bullet} \left(\bar{X}_i - \bar{X}\right)^2 - (I-1)\widehat{\sigma^2}\right)$

— $\widehat{\tau^2} = max\left(\widehat{\widehat{\tau^2}}; 0\right)$

$\widehat{\sigma^2}$ et $\widehat{\widehat{\tau^2}}$ sont des estimateurs sans biais et convergents, à la condition pour $\widehat{\widehat{\tau^2}}$ que $\sum_i (\frac{\omega_{i\bullet}}{\omega_{\bullet\bullet}})^2 \longrightarrow 0$ quand $I \longrightarrow \infty$.

En d'autres termes, il faut qu'aucun des risque ne soit dominant sur les autres. Aussi, $\widehat{\widehat{\tau^2}}$ peut prendre des valeurs négatives. C'est pourquoi nous utilisons $\widehat{\tau^2}$, qui du fait du $max()$ n'est plus sans biais.

Pour calculer le coefficient de crédibilité, les étapes à suivre sont les suivantes :

1. Calcul des poids ω_{ij}
2. Calcul des ratios X_{ij}
3. Calcul de $\widehat{\sigma^2}$
4. Calcul de $\widehat{\tau^2}$
5. Calcul de $\widehat{z_i}$

Avec ce modèle, plus le poids d'un individu est important, plus son coefficient de crédibilité va être proche de 1, du fait que z_i augmente avec $\omega_{i\bullet}$. En d'autres termes, plus l'exposition d'un contrat est importante, plus sa prime sera individualisée. Ce constat est simple mais intuitif, car c'est ce que nous cherchons à appliquer dans la réalité.

4.3.2 Application

Pour implémenter la crédibilité de Bühlmann-Straub sous R, nous avons à disposition la fonction *cm* du package *actuar*. Les ratios à projeter sont les fréquences sinistres et les poids les taux de moteurs années, ce sur les 4 années d'historique.

En sortie de ce modèle, nous avons les estimateurs suivants :

$\widehat{\widehat{\mu_0}}$	$\widehat{\tau^2}$	$\widehat{\sigma^2}$	$\widehat{K}$
0.0641	0.0017	0.0550	32.0526

TABLE 4.1 – Estimateurs pour le calcul de z, modèle de Bühlmann-Straub

Ce qui va nous intéresser maintenant, c'est l'évolution de ce coefficient en fonction de l'exposition du contrat, qui est le taux de moteurs du dit contrat sur les 4 dernières années. Nous pouvons facilement illustrer cela sur le graphique suivant.

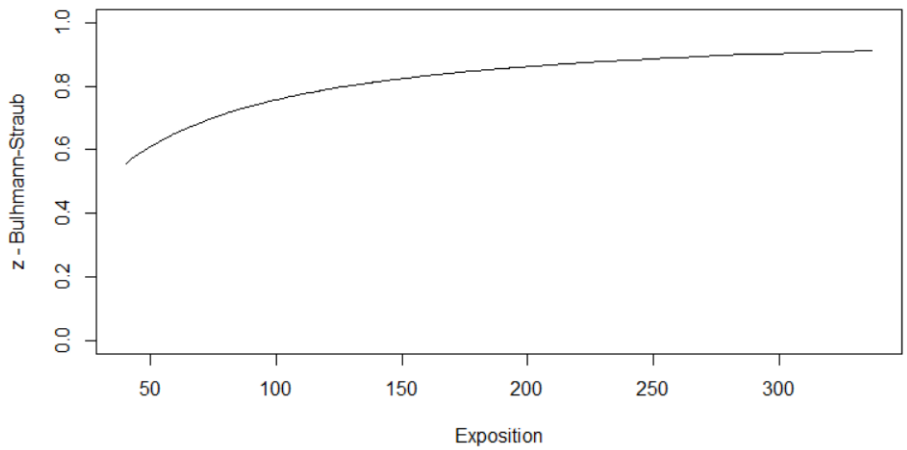

FIGURE 4.2 – Facteur de crédibilité estimé par méthode BS en fonction de l'exposition sur les 4 dernières années

La première remarque que nous pouvons nous faire depuis ce graphique est que le coefficient z commence avec des valeurs très élevées, le minimum est à 0.5581. Le modèle considère donc que dès 40 moteurs (soit 40 moteurs sur 1 année d'expérience), l'expérience est suffisamment importante pour individualiser le tarif de près de 56%. D'après la formule du coefficient de crédibilité, un z élevé peut se traduire par une forte hétérogénéité du portefeuille et/ou une faible variabilité interne du risque. Nous savons en effet que nos fréquences sont surdispersées, la variance de la distribution est donc forte par rapport à l'espérance et nos profils de risques sont assez hétérogènes, au regard notamment des critères tarifaires.

Aussi le coefficient de crédibilité n'atteint pas tout à fait 1. Sa valeur maximale est de 0.9131 pour une exposition de 337. Nous retrouvons toutefois cette croissance plus rapide sur les petites expositions et plus lente sur les plus élevées avec cette allure logarithmique ou racine.

Pour comparer directement avec le coefficient de crédibilité actuellement mis en place, nous représentons les coefficients estimées par méthode de Bühlmann-Straub en fonction du dernier taux de moteurs année.

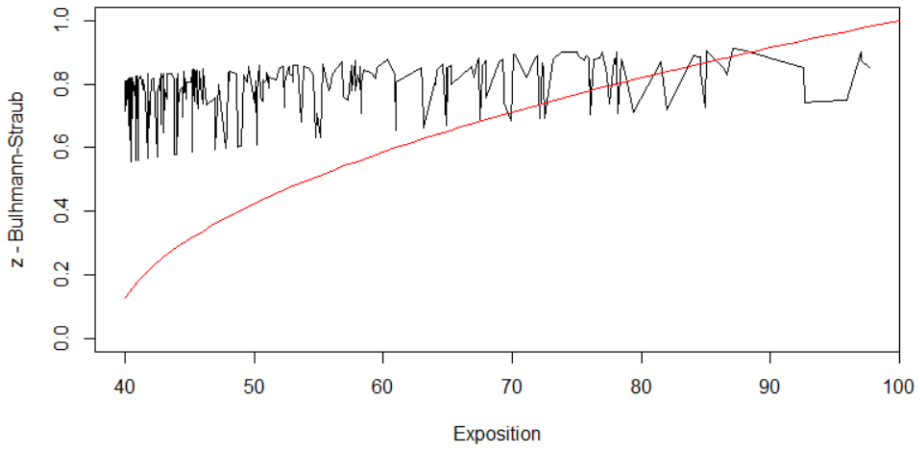

FIGURE 4.3 – Facteur de crédibilité estimé par méthode BS en fonction de la dernière exposition

Sur ce graphique, la courbe noire représente les z estimés et la courbe rouge le modèle actuel de crédibilité. Les points les plus bas des coefficients estimés représentent les contrats avec le moins d'historique, les plus haut ceux qui en ont le plus. La courbe est plus concentrée sur les expositions les plus faibles du fait de la distribution de nos expositions, se référer aux figures 2.15 et 2.16.

Avec ce premier estimateur, nous sommes assez loin de la crédibilité actuellement mise en place. En comparaison, les coefficients sont bien plus élevés sur les petites expositions et légèrement plus faibles sur les expositions plus importantes.

4.4 Le modèle hiérarchique de Jewell

Le modèle proposé par Bühlmann et Straub suppose une certaine homogénéité du risque dans le portefeuille, ce qui en pratique n'est pas forcément le cas, nous avons en effet un portefeuille plutôt hétérogène au regard des critères de zones et d'activité par exemple. Jewell revient donc sur ce modèle en 1975 [26] pour introduire une crédibilité hiérarchique qui permet de subdiviser un portefeuille en différentes classes de risques distinctes.

Dans notre cas, nous avons un modèle à 3 niveaux : les observations, les individus et un regroupement en classes. Mais ce modèle peut s'appliquer sur un nombre plus importants de classes encore. Nous introduisons donc pour la suite, les notations :

- $k = 1, ..., K$: l'indice de classe
- Φ_k : risque de la classe k, une variable aléatoire
- Θ_{ki} : risque de l'individu i qui se trouve dans la classe k $(i = 1, ..., I_k)$, une variable aléatoire
- X_{ijk} : fréquence sinistre sur l'année j, du contrat i qui se trouve dans la classe k

Notre hiérarchie s'organise de la manière suivante :

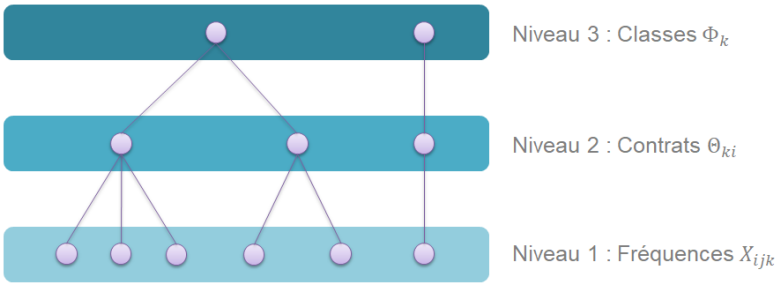

FIGURE 4.4 – Schéma de la hiérarchie du classement pour la mise en place de la crédibilité

Pour l'application de cette crédibilité hiérarchique, les variables utilisées pour construire les niveaux ne doivent pas être liées entre elles au risque de ne pas prendre en compte les interactions entre ces variables.

4.4.1 Rappels de la théorie

Le modèle de Jewell se base sur les hypothèses suivantes :

(H1) Les variables aléatoires Φ_k sont indépendantes et identiquement distribuées

(H2) Les variables aléatoires Θ_{ki} sont, conditionnellement à Φ_k, indépendantes et identiquement distribuées

(H3) Les variables aléatoires X_{ijk} sont, conditionnellement à Θ_{ki} indépendantes et identiquement distribuées. Ses moments conditionnels d'ordre 1 et 2 sont notés :

$$\mu(\Theta_{ki}, \Phi_k) = \mathbf{E}\left[X_{kij}|\Theta_{ki}, \Phi_k\right]$$
$$\sigma^2(\Theta_{ki}, \Phi_k) = \omega_{kij}\mathbf{V}\left(X_{kij}|\Theta_{ki}, \Phi_k\right)$$

Le principe de cette crédibilité hiérarchique, est d'appliquer la crédibilité de Bühlmann-Straub niveau après niveau, à commencer par le plus élevé (le niveau 3 des classes pour nous).

Théorème 3 *Modèle de Jewell. Sous (H1), (H2), (H3), l'estimateur non-homogène de crédibilité est donné par :*

$$\widehat{\widehat{\mu(\Theta_{ki}, \Phi_k)}} = z_{ki}\bar{X}_{ki} + (1 - z_{ki})\widehat{\widehat{\mu(\Phi_k)}}$$

$$\widehat{\widehat{\mu(\Phi_k)}} = z_k\bar{X}_k + (1 - z_k)\widehat{\widehat{\mu_0}}$$

Avec :

- $\bar{X}_{ki} = \sum_{j=1}^{n_{ki}} \frac{\omega_{kij}}{\omega_{ki\bullet}} X_{kij}$ *: moyenne individuelle pondérée des poids*
- $\bar{X}_k = \sum_{i=1}^{I_k} \frac{z_{ki}}{z_{k\bullet}} \bar{X}_{ki}$ *: moyenne de la classe pondérée des coefficients de crédibilité*
- $\widehat{\widehat{\mu_0}} = \sum_{k=1}^{K} \frac{z_k}{z_\bullet} \bar{X}_k$ *: moyenne collective pondérée*

Les coefficients de crédibilité sont de la forme suivante :

$$z_{ki} = \frac{\omega_{ki\bullet}}{\omega_{ki\bullet} + \frac{\sigma^2}{a}} \qquad \omega_{ki\bullet} = \sum_{j=1}^{n_{ki}} \omega_{kij}$$
$$z_k = \frac{z_{k\bullet}}{z_{k\bullet} + \frac{a}{b}} \qquad z_{k\bullet} = \sum_{i=1}^{I_k} z_{ki}$$
$$z_\bullet = \sum_{k=1}^{K} z_k \qquad z_{\bullet\bullet} = \sum_{k=1}^{K} z_{k\bullet}$$

Nous pouvons commencer par déterminer l'estimateur de la variabilité intra-contrat, pour l'estimation de z_{ki} :

$$\widehat{\sigma^2} = \frac{1}{\sum_{k=1}^{K}\sum_{i=1}^{I_k}(n_{ki} - 1)} \sum_{k=1}^{K}\sum_{i=1}^{I_k}\sum_{j=1}^{n_{ki}} \omega_{kij}\left(X_{kij} - \bar{X}_{ki}\right)^2$$

Cet estimateur est sans biais.

Pour les variances intra (a) et inter-classes (b), plusieurs estimateurs peuvent être utilisés. Nous commençons par présenter les estimateurs proposés par Goovaerts :

$$\widetilde{a} = \frac{1}{\sum_{k=1}^{K}(I_K - 1)} \sum_{k=1}^{K} \sum_{i=1}^{I_k} z_{ki} \left(\bar{X}_{ki} - \bar{X}_k\right)^2$$

$$\widetilde{b} = \frac{1}{K-1} \sum_{k=1}^{K} z_k \left(\bar{X}_k - \bar{\bar{X}}_k\right)^2$$

Avec

$$\bar{\bar{X}}_k = \sum_{k=1}^{K} \frac{z_{k\bullet}}{z_{\bullet\bullet}} \bar{X}_k$$

Ces estimateurs sont semblables au modèle de Bühlmann-Straub et sont directement intuités depuis leur définition. Cependant, pour les estimer nous devons procéder de manière itérative, ce qui est en pratique plus difficile à mettre en place, car les poids utilisés pour estimer les paramètres d'un niveau sont les facteurs de crédibilité du même niveau.

Pour palier à cela, il est possible de remplacer les coefficients de crédibilité z par les poids ω. C'est la proposition qui a été notamment apportée par Ohlsson. Les estimateurs suivants sont sans biais :

$$\widehat{\widehat{a}} = \frac{\sum_{k=1}^{K} A_k}{\sum_{k=1}^{K} c_k}$$

Avec :

$$A_k = \sum_{i=1}^{I_k} \omega_{ki\bullet} \left(\bar{X}_{ki} - \bar{X}_k\right)^2 - (I_k - 1)\,\sigma^2$$

$$c_k = \omega_{k\bullet\bullet} - \sum_{i=1}^{I_k} \frac{\omega_{ki\bullet}^2}{\omega_{k\bullet\bullet}}$$

Et

$$\widehat{\widehat{b}} = \frac{B}{d}$$

Avec :

$$B = \sum_{k=1}^{K} z_{k\bullet} \left(\bar{X}_k - \bar{\bar{X}}_k\right)^2 - a(K-1)$$

$$d = z_{\bullet\bullet} - \sum_{k=1}^{K} \frac{z_{i\bullet}^2}{z_{\bullet\bullet}}$$

Afin de palier aux éventuelles valeurs négatives, Ohlsson propose les estimateurs tronqués suivants :

$$\widehat{a} = max(\widehat{\widehat{a}}, 0)$$

$$\widehat{b} = max(\widehat{\widehat{b}}, 0)$$

D'autres méthodes de troncatures des estimateurs existent, nous pouvons mentionner celle de Bühlmann et Gisler par exemple. Ces méthodes sont assez équivalentes à l'utilisation.

4.4.2 Application

Préambule à la classification : l'Analyse des Correspondances Multiples

Avant de pouvoir mettre en place la crédibilité hiérarchique, nous devons créer des classes de risque. Ces classes vont dépendre du croisement zone x activité. Nous avons 6 zones et 8 activités, soit un total de 48 modalités de croisement. Nous ne pouvons pas conserver autant de classes, d'autant plus que certains croisements sont largement sous-représentés, se référer à 3.7. Nous allons donc procéder à une classification.

Nous commençons par implémenter une ACM (Analyse des Correspondances Multiples) avec la fonction *MCA* du package *FactoMineR* et le package *factoextra* qui permet de visualiser les sorties de l'ACM. Cette méthode de

classification se base sur l'analyse factorielle des correspondances appliquée à des variables catégorielles, comme dans notre cas. L'approche ACM a été introduite par Jean-Paul Benzécri en 1963 dans le cours Peccot. Cette méthode cherche à regrouper les individus ayant des profils similaires mais également à identifier les liens entre les variables. L'ACM a l'avantage de considérer des liens linéaires mais aussi non linéaires. Chaque modalité des variables est considérée, et non pas la variable seule. L'ACM permet de créer des classes de risques décorrélées. Ces nouvelles variables sont ainsi appelées axes principaux.

En sortie de l'ACM, ce qui va d'abord nous intéresser c'est le nombre d'axes principaux à prendre en compte. Pour cela, nous pouvons étudier les valeurs propres et le pourcentage de variance expliquée par chacun des axes principaux :

	Valeurs propres	Variance expliquée (%)	Cumul de variance expliquée (%)
dim 1	0.63	10.54	10.54
dim 2	0.59	9.91	20.45
dim 3	0.57	9.51	29.95
dim 4	0.56	9.26	39.22
dim 5	0.54	9.03	48.25
dim 6	0.50	8.33	56.59
dim 7	0.50	8.33	64.92
dim 8	0.46	7.63	72.55
dim 9	0.44	7.40	79.95
dim 10	0.43	7.16	87.11
dim 11	0.41	6.76	93.87
dim 12	0.37	6.13	100.00

TABLE 4.2 – Valeurs propres et variance expliquée

Ce tableau peut également être représenté à l'aide d'un *scree plot*, un graphique des valeurs principales des axes principaux en sortie d'ACM. Ce graphique a été introduit par Raymond B. Cattel (1966) [21].

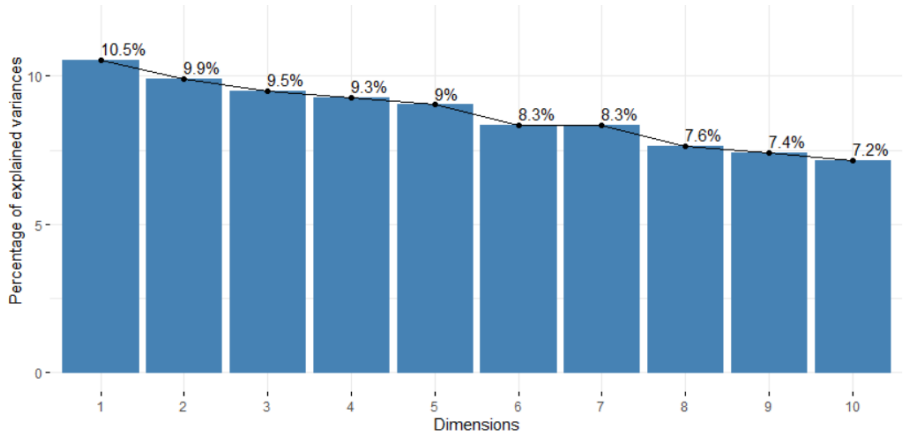

FIGURE 4.5 – Variance expliquée par les axes principaux

Il existe plusieurs méthodes pour déterminer le nombre optimal d'axes principaux à conserver. La méthode du coude par exemple, qui consiste à identifier un décrochage dans le pourcentage de variance expliquée trié en ordre décroissant. En se référant à notre graphique précédent, nous n'observons pas particulièrement de coude, la décroissance est assez stable. Nous pouvons aussi retenir le critère de Kaiser qui consiste à ne sélectionner que les axes dont les valeurs propres sont supérieures à l'inertie moyenne. Dans notre cas la moyenne des valeurs propres est de 0.5. En retenant ce critère, nous considérons 5 axes principaux avec un cumul de variance expliquée de 48.25%.

L'étude de la contribution de chaque modalité sur les axes principaux peut notamment se faire à l'aide d'une matrice de corrélation.

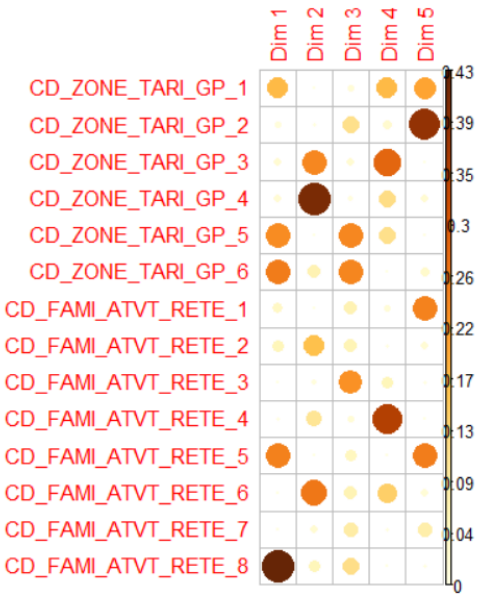

FIGURE 4.6 – Matrice de corrélation des modalités avec les axes principaux

Plus le point au croisement de la modalité et de l'axe principal est foncé et large, plus la contribution de la modalité est forte dans l'axe. De cette matrice, nous déduisons par exemple que la zone 1 est surtout représentée par les axes 1, 4 et 5, la zone 2 par l'axe 5, etc. Seule l'activité 7 n'est pas bien représentée par les axes principaux 1 à 5.

Nous pouvons aussi étudier la contribution des modalités sur chacun des axes.

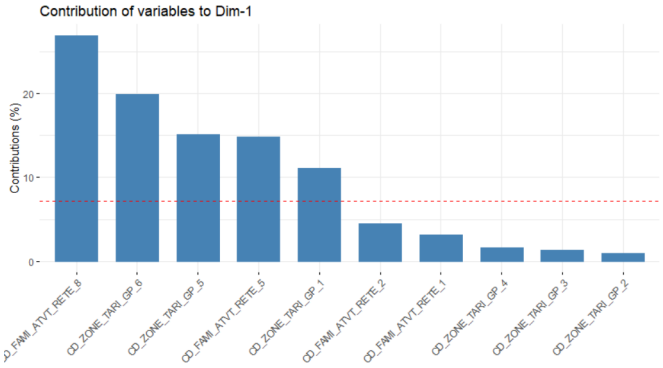

FIGURE 4.7 – Contributions à l'axe principal 1

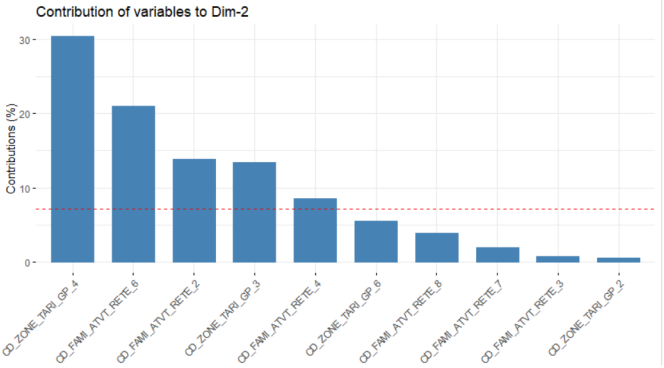

FIGURE 4.8 – Contributions à l'axe principal 2

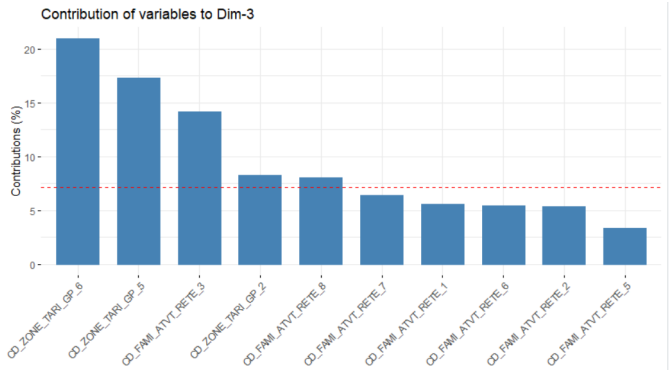

FIGURE 4.9 – Contributions à l'axe principal 3

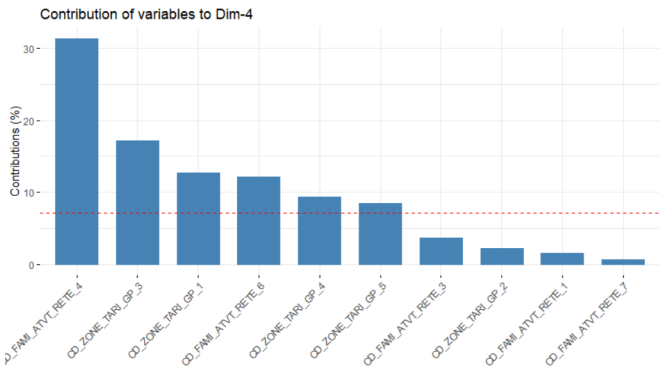

FIGURE 4.10 – Contributions à l'axe principal 4

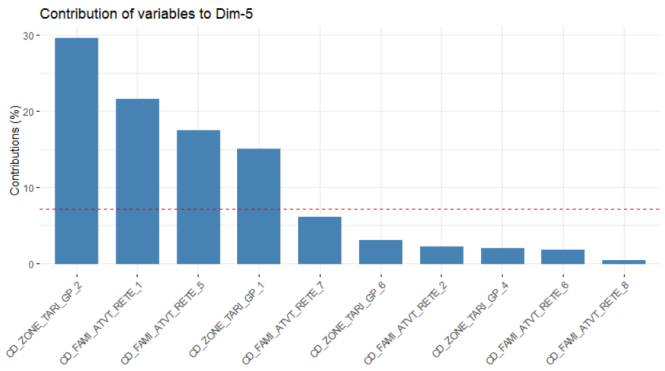

FIGURE 4.11 – Contributions à l'axe principal 5

En se penchant par exemple sur les axes principaux 1 et 2 qui forment le plan, les modalités qui contribuent le plus à l'axe 1 sont les activités 5 et 8 ainsi que les zones 1, 5 et 6. Celles qui contribuent le plus à l'axe 2 sont les activités 2, 4 et 6 et les zones 3 et 4.

Nous pouvons également nous pencher sur les barycentres des modalités sur le plan formé par les axes 1 et 2 et leur ellipse de confiance.

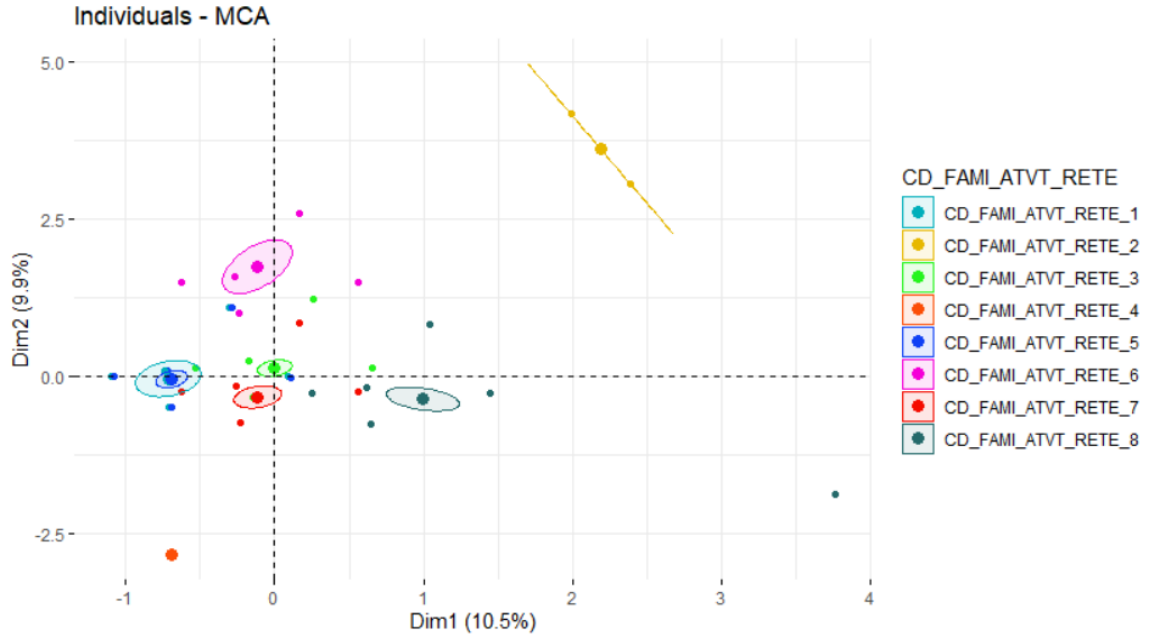

FIGURE 4.12 – Représentation des modalités de la variable activité

Les activités 2 et 4 sont particulièrement éloignées des autres activités. À l'inverse, les activités 1 et 5 se chevauchent. La contribution négative de l'activité 5 et celle positive de l'activité 8 sur l'axe 1 s'observent ici. De même pour l'axe 2 avec l'activité 4 dans le pôle négatif et les activités 2 et 6 dans le pôle positif.

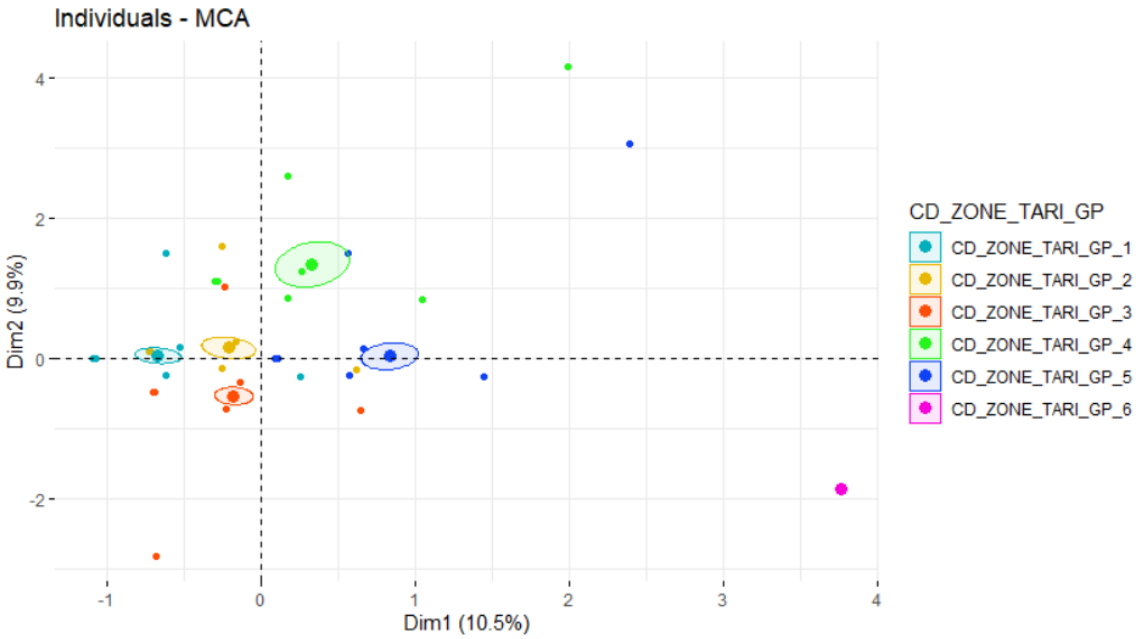

FIGURE 4.13 – Représentation des modalités de la variable zone

Les zones sont ordonnées de gauche à droite par sinistralité croissante, la zone 6 étant particulièrement éloignée des zones 1 à 5. Nous retrouvons bien nos constats précédents, l'axe 1 est bien expliqué par la zone 1 dans son pôle négatif et les zones 5 et 6 dans son pôle positif, ce qui nous semble intuitif. Pour l'axe 2 la zone 3 définit son pôle négatif et la zone 4 son pôle positif, avec des barycentres assez proches mais des modalités qui s'étendent aux opposés.

Suite à cette ACM, les croisements sont maintenant représentées sur 5 axes principaux.

Classification Hiérarchique sur Composantes Principales

Une fois l'analyse des correspondances multiples mise en place, nous pouvons procéder à une classification hiérarchique sur composantes principales (*HCPC* pour *Hierarchical Clustering on Principle Components*) pour créer des classes effectives, en fonction de la zone et de l'activité du contrat. Effectuer une ACM avant la mise en place d'une classification hiérarchique permet de faciliter la construction des classes du fait que l'ACM réduit la taille du jeu de données en créant des axes principaux qui contiennent les informations des variables.

Cette méthode de classification hiérarchique se base sur des arbres de Ward. De tels arbres utilisent le théorème de Huygens qui décompose l'inertie totale en inertie inter et intra. La distance considérée ici est la distance euclidienne.

$$\underbrace{\sum_{k=1}^{K}\sum_{i=1}^{I_k}(x_{ik}-\bar{x})^2}_{\text{Inertie totale}} = \underbrace{\sum_{k=1}^{K}(\bar{x}_k-\bar{x})^2}_{\text{Inertie inter-classe}} + \underbrace{\sum_{k=1}^{K}\sum_{i=1}^{I_k}(x_{ik}-\bar{x}_k)^2}_{\text{Inertie intra-classe}}$$

Avec :

- x_{ik} : l'individu i de la classe k
- $\bar{x}$: la moyenne des individus
- $\bar{x}_k$: la moyenne des individus de la classe k

Cette méthode permet de considérer l'inertie intra et inter-groupe. Ainsi, en partant des individus, deux clusters sont regroupés de manière à minimiser le gain d'inertie intra-groupe ce qui revient à minimiser la réduction de l'inertie inter-groupe [10]. En effet, le but est d'avoir des clusters tels que les individus au sein de ce cluster soient proches et que les individus de clusters différents soient suffisamment éloignés. L'algorithme continue jusqu'à n'avoir plus qu'un groupe, c'est la racine de l'arbre, les feuilles terminales de l'arbre étant les individus. La prochaine étape est de déterminer à quel niveau de l'arbre effectuer le découpage, autrement dit combien de classes nous souhaitons avoir.

Sous R, nous utilisons la fonction *HCPC* du package *FactoMineR*. En sortie de cette fonction, nous avons notamment le graphique qui évalue la perte d'inertie inter-classes quand au passage de la classe $K+1$ à K .

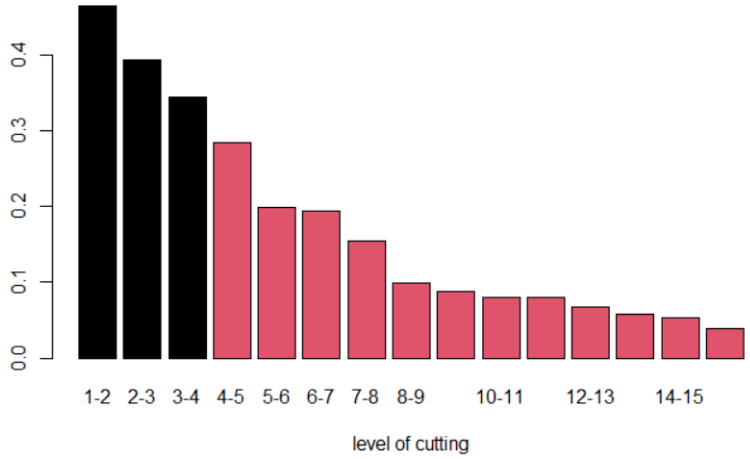

FIGURE 4.14 – Perte d'inertie inter-classes

En regroupant nos classes, nous perdons de l'inertie inter-classe jusqu'au dernier regroupement où l'inertie inter-classe est nulle, puisque nous n'avons plus qu'une classe. Comme mentionné plus tôt, nous voulons minimiser cette perte d'inertie inter-classe tout en restant cohérent avec nos données et notamment leur volumétrie. Nous ne souhaitons pas avoir un nombre de classes trop élevé pour garder une classification significative. Par défaut, R choisit la partition qui a la perte d'inertie intra-classe relative la plus élevée, soit la perte d'inertie inter-classe la plus faible. Dans notre cas de figure, le meilleur partitionnement se fait en 4 classes. Ces 4 classes peuvent être représentées sur le dendogramme par un jeu de couleurs.

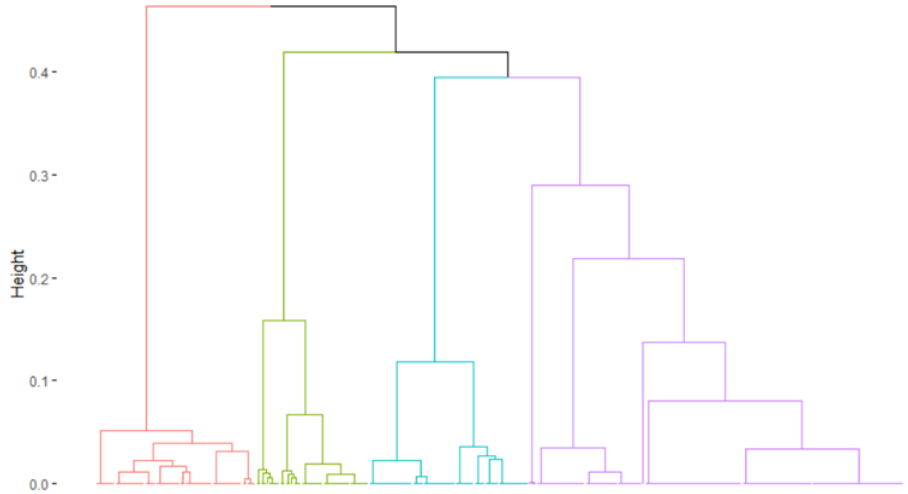

FIGURE 4.15 – Dendogramme de l'*HCPC*

La taille de ce dendogramme correspond à l'inertie intra-classe. Au pied du dendogramme, l'inertie est nulle car il n'y a pas de regroupement en classe.

Avec cette classification, la répartition de nos données dans chacune des classes est la suivante.

1	2	3	4
46.49%	19.06%	15.39%	19.06%

TABLE 4.3 – Répartition du nombre de contrats dans les différents clusters

La répartition dans nos clusters n'est pas tout à fait homogène. La classe 1 contient près de la moitié de nos contrats, la deuxième moitié étant répartie sur les 3 classes restantes. Nous nous satisfaisons tout de même de cette classification, l'objectif étant de créer des classes homogènes dans leur risque et non en volumétrie.

Le graphique suivant permet de visualiser nos individus sur le plan des axes principaux 1 et 2 en fonction de leurs groupes.

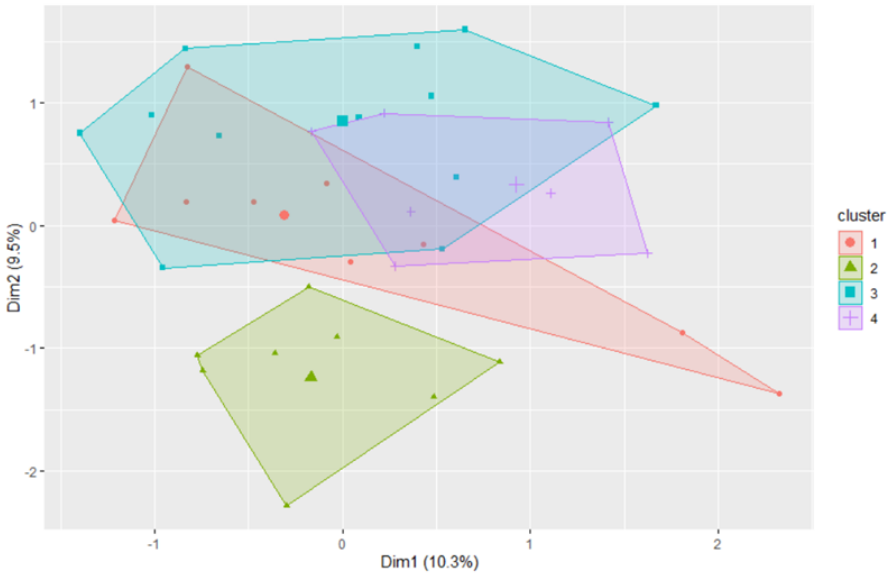

FIGURE 4.16 – Visualisation des individus par groupe

La classe 2 s'oppose à la classe 3 sur le deuxième axe alors que les classes 1, 3 et 4 ont plutôt tendance à se chevaucher sur ce plan.

Afin de comprendre un peu mieux notre classification, il est possible de définir les clusters par les modalités qui les caractérisent. Par exemple la classe 1 est caractérisée par :

- zones : 1, 2, 3, 4, 5
- activités : 1, 3

La classe 2 :

- zone : 6

La classe 3 :

- zones : 1, 2, 3, 4, 5
- activités : 1, 3, 6, 7

La classe 4 :

- zones : 1, 2, 3, 4, 5
- activités : 1, 2, 6, 7

Il est également possible d'expliquer notre classification à l'aide d'un arbre de décision CART. Nous utilisons la fonction *rpart* pour implémenter un tel arbre et *rpart.plot* pour le représenter.

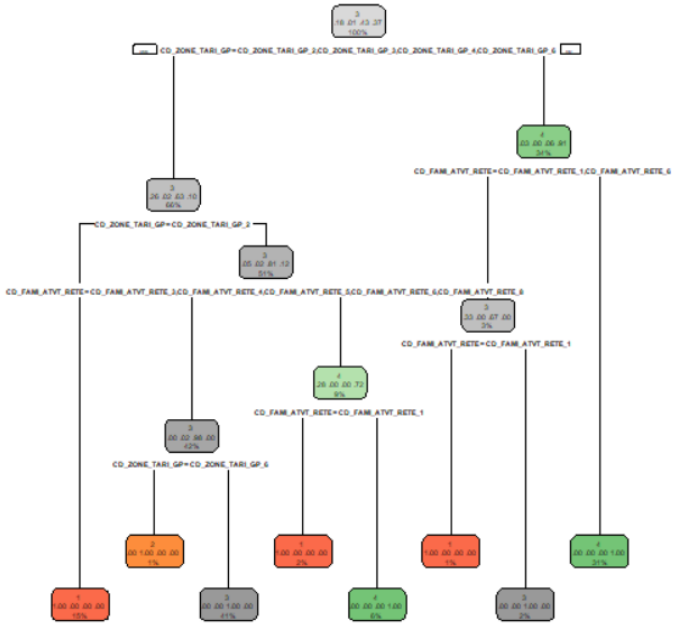

FIGURE 4.17 – Arbre de décision pour la caractérisation de la classification

À la racine, nous testons la zone géographique, si elle est égale à 2, 3, 4 ou 6. Si oui, nous descendons la branche de gauche pour tester si la zone est la 2, si c'est la cas, le contrat est dans la classe 1. Ainsi par exemple, un contrat est dans la classe 1 si il appartient à :

- zone 2
- zone 3, 4 ou 6 et activité 1
- zone 1 ou 5 et activité 1

Nous avons ainsi mis en place une classification avec 4 groupes, sur lesquels nous pouvons maintenant appliquer la méthode de crédibilité de Jewell.

Mise en place de la crédibilité

Pour l'application de la crédibilité hiérarchique dans R, nous utilisons la même fonction *cm* que pour la crédibilité de Bühlmann-Straub. Cette méthode utilise les estimateurs itératifs $(\tilde{a}, \tilde{b})$.

Le modèle nous fournit les estimateurs suivants :

$\widehat{\widehat{\mu_0}}$	$\widehat{\sigma^2}$	$\tilde{a}$	$\widehat{K}$	$\tilde{b}$	$\tilde{K}$
0.0709	0.0550	0.0017	32.0526	0.0001	14

TABLE 4.4 – Estimateurs pour le calcul de z, modèle de Jewell

Nous avons bien les mêmes estimateurs liés à la crédibilité au niveau du contrat.

Comparons maintenant l'évolution de ce nouvel estimateur de crédibilité avec l'estimateur de Bühlmann-Straub.

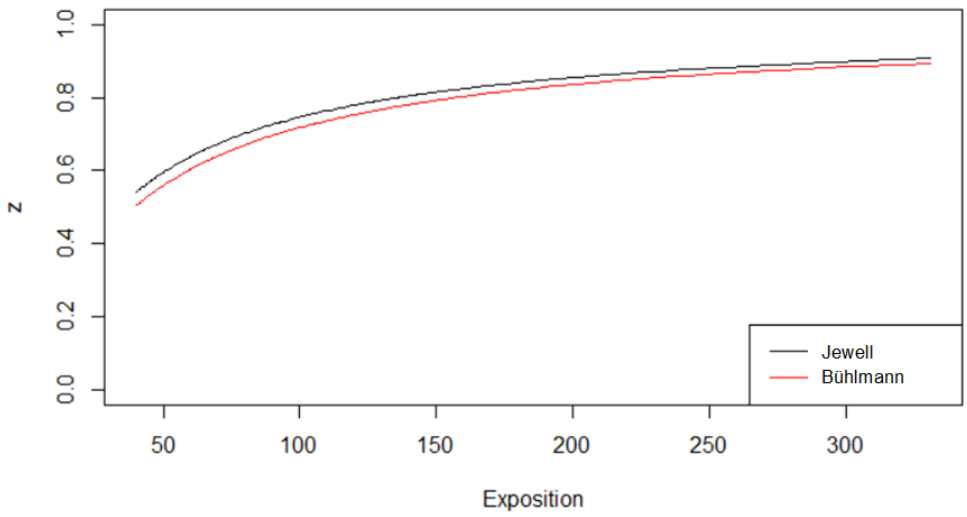

FIGURE 4.18 – Facteurs de crédibilité estimés par méthode de Bühlmann-Straub et de Jewell en fonction de l'exposition sur les 4 dernières années

Les deux courbes sont assez proches. Les estimations de z par la méthode de Jewell sont même un peu plus élevées que les estimations de Bühlmann-Straub. Cette proximité peut être expliquée par la faible hiérarchie dont nous disposons. En effet, nous n'avons qu'un palier de classes qui regroupe les contrats, nos individus. Cette classification n'est peut-être pas pertinente non plus. En effet la variance inter-classes estimées ($\tilde{b}$) est très faible, même plus faible que la variance intra-classes ($\tilde{a}$). Notre risque intra classe est moins homogène que le risque entre les différentes classes.

À la vue de ces résultats, nous préférons conserver le modèle de crédibilité de Bühlmann-Straub plutôt que le modèle de Jewell. En effet, les estimateurs sont assez proches entre les différents modèles, mais nous introduisons un risque de conception supplémentaire en créant des classes avec le modèle hiérarchique et un aléa en mettant en place une classification. Classification pour laquelle nous émettons des doutes quant à sa pertinence au regard des données. Pour assurer la solidité du modèle, nous préférons donc le modèle de Bühlmann-Straub.

4.5 Approximation du coefficient de crédibilité

Avec un modèle de crédibilité sélectionné, le modèle de Bühlmann-Straub, nous cherchons maintenant une formule explicite pour le calcul du coefficient de crédibilité. En effet, nous souhaitons implémenter notre modèle de fréquence crédibilisée sous VBA et appliquer de manière simple, rapide et fiable cette crédibilité à tout nouveau contrat, en fonction de l'exposition au risque. C'est pourquoi nous utilisons maintenant un MLG pour approcher notre coefficient z en fonction du taux de moteurs présents sur les 4 dernières années.

Intuitivement, en se basant sur l'allure de la courbe de z en fonction de l'exposition (graphique 4.2), nous estimons que la courbe a une allure logarithmique. Nous testons donc un premier MLG de formule :

$$z = \beta_0 + \beta_1 \times log(\text{Exposition})$$

Le MLG prédit les paramètres suivants.

β_0	β_1
-0.1736	0.1912

TABLE 4.5 – Estimations des coefficients pour l'approche logarithmique de z

Ces coefficients sont jugés tous deux significatifs avec une p-value inférieure à $2e^{-16}$. Le R^2 de ce modèle est de 0.9837. En considérant ces sorties, nous sommes assez satisfaits de ce modèle qui semble bien estimer notre crédibilité en fonction de l'exposition. Graphiquement, l'approximation donne le résultat suivant :

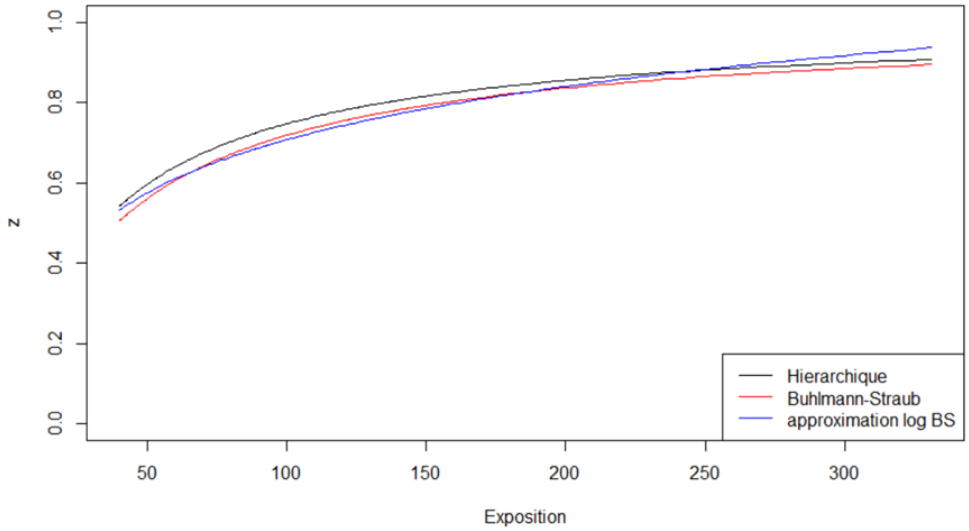

FIGURE 4.19 – Approximation logarithmique par MLG du facteur de crédibilité de Bühlmann-Straub

Cette approximation accorde un peu plus de poids au petit moteur d'entrée et est moins concave que la courbe des coefficients de Bühlmann-Straub. Cependant, l'adéquation est très satisfaisante de 60 à 250 taux moteurs environ. Sur les expositions supérieures, le coefficient approché accorde plus d'individualisation aux contrats. Sur les flottes proches de 400 moteurs, le coefficient tend vers 1 plus rapidement. Cette différence nous est plutôt favorable. En effet, nous nous rapprochons des contrats appartenant au segment 4 avec leur tarif complètement individualisé, l'équivalent d'un coefficient $z = 1$.

Cette approximation nous convient bien, mais nous souhaitons également tester une approximation avec la fonction racine, notamment pour faire écho à la formule actuellement utilisée.

$$z = \beta_0 + \beta_1 \times \sqrt{\text{Exposition}}$$

Les coefficients prédits sont décrits dans la table ci-dessous.

β_0	β_1
0.5001	0.0279

TABLE 4.6 – Estimations des coefficients pour l'approche racine de z

Les coefficients estimés sont ici aussi considérés comme significatifs du fait de leur p-value inférieure à $2e-16$. Le R^2 cependant est un peu plus faible, de l'ordre de 0.9121, ce qui est tout de même assez élevé.

Mais graphiquement, le résultat est moins satisfaisant.

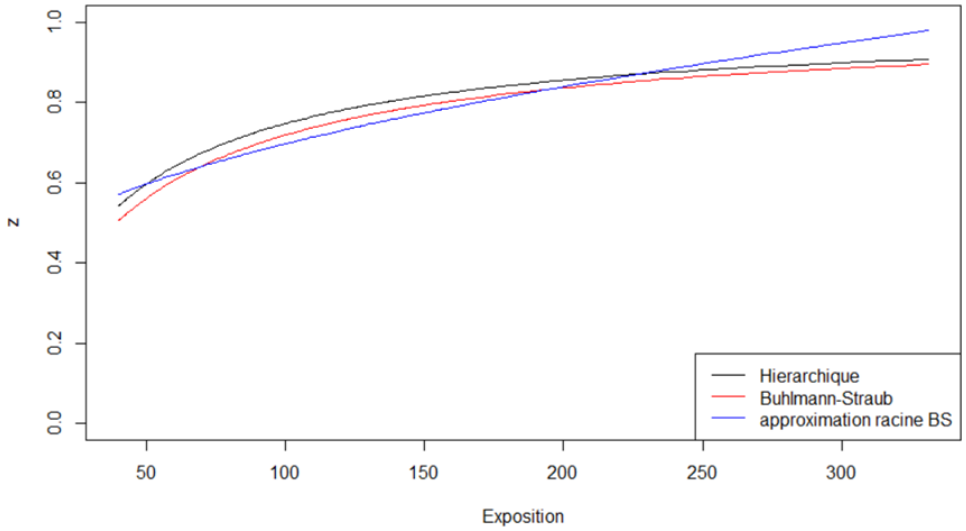

FIGURE 4.20 – Approximation racine par MLG du facteur de crédibilité de Bühlmann-Straub

En effet, sur le segment d'exposition entre 40 et 400 moteurs, la courbe semble presque linéaire, nous ne retrouvons pas la concavité caractéristique de l'estimateur de Bühlmann-Straub. Cette approximation est aussi bien plus éloignée, surtout sur les expositions avant 60 moteurs et après 230 moteurs.

Nous conservons donc l'approximation logarithmique pour prédire le coefficient de crédibilité. Notre modèle de fréquence crédibilisé s'implémente donc simplement avec :

- pour la fréquence collective : des coefficients explicites fonction des critères de zone et d'activité de l'entreprise assurée
- pour le coefficient de crédibilité : une formule explicite fonction de l'exposition au risque, soit le taux de moteurs année sur les 4 dernières années

4.6 Étude d'impact

Ce nouveau modèle de fréquence peut être confronté avec le modèle actuellement mis en place afin d'étudier l'impact de la révision tarifaire. Cette étude peut être menée à plusieurs niveaux : sur la modélisation de la fréquence collective, la fréquence crédibilisée et les coefficients de crédibilité.

Écarts tarifaires sur les fréquences collectives

Au global, l'écart tarifaire entre la fréquence collective révisée et la fréquence collective actuelle est de -56%. Ce qui signifie qu'en moyenne le tarif révisé est 56% moins cher que le tarif actuel à coût moyen égal. Cet écart n'est pas homogène sur les différents croisements des critères zone et activité, comme illustré ci-dessous.

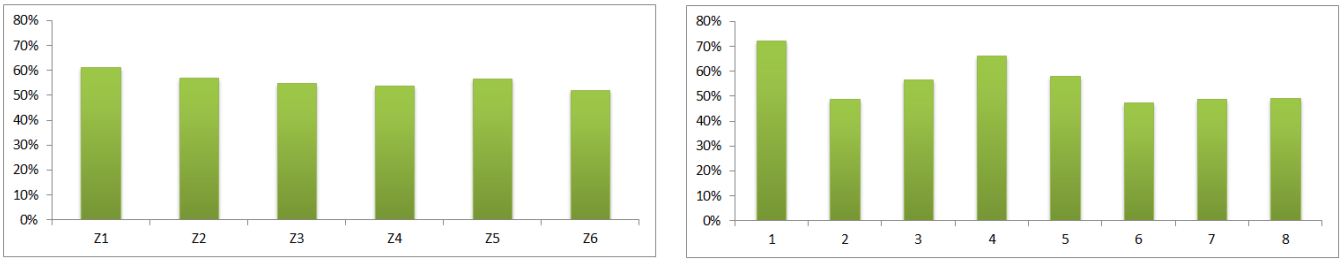

FIGURE 4.21 – Écarts tarifaires en valeur absolue entre la fréquence collective révisée et actuelle sur les zones (gauche) et les activités (droite)

L'écart moyen sur la zone 1 est un peu plus important que sur les zones suivantes. Néanmoins les écarts ne sont pas flagrants avec un écart type à la moyenne de 0.03. À l'inverse, les écarts de fréquence collective sur les activités sont bien plus prononcés avec un écart type 3 fois supérieur. L'écart entre la fréquence actuelle et révisée oscille entre -48% sur l'activité 6 et -72% sur l'activité 1. Ainsi, l'écart le plus faible est tout de même de -43%

et est porté par le croisement zone 6 - activité 6. Le plus important est sur le croisement zone 1 - activité 1 et la fréquence collective révisée est 75% moins élevée sur ce croisement.

La fréquence collective révisée est toujours inférieure à la fréquence actuelle mise en place en 2009, ce qui se représente graphiquement de la manière suivante :

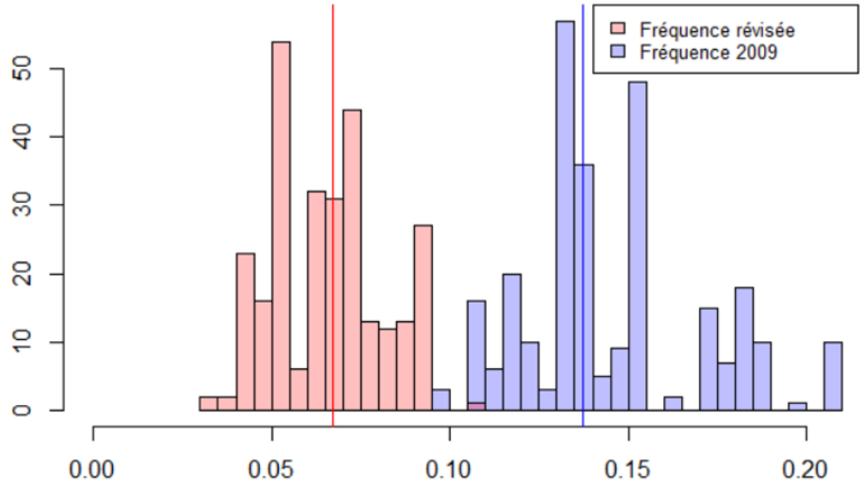

FIGURE 4.22 – Histogramme des fréquences collectives révisées et actuelles

Les droites verticales correspondent aux médianes des fréquences collectives. Cette représentation graphique illustre bien le fort écart entre les deux fréquences collectives.

Écarts tarifaires sur les fréquences crédibilisées

Avec l'individualisation de la fréquence, ces écarts sont voués à s'amoindrir. C'est certes ce qui est observé sur le fréquences crédibilisées, mais en moyenne l'écart tarifaire est tout de même de -36%. L'individualisation creuse les disparités sur les croisements de zone et d'activité, ce qui s'illustre par les graphiques suivants :

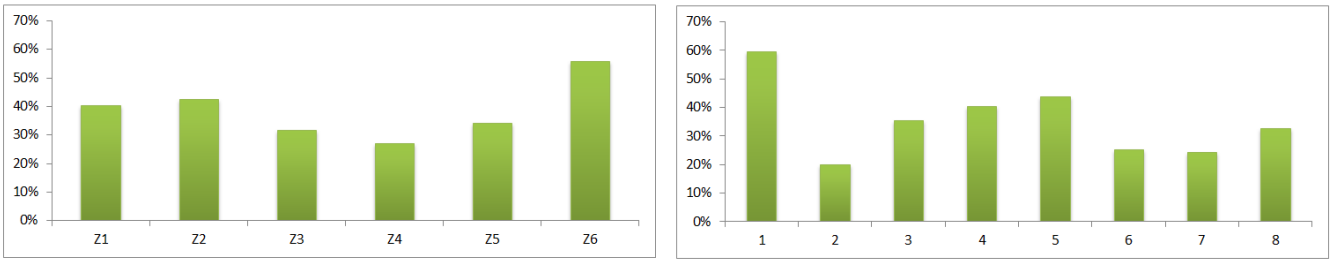

FIGURE 4.23 – Écarts tarifaires en valeur absolue entre la fréquence crédibilisée révisée et actuelle sur les zones (gauche) et les activités (droite)

En effet les écarts sur les différentes zones ne sont plus aussi homogènes que pour la fréquence collective avec un écart type plus de 3 fois supérieur qui atteint donc 0.10. L'écart tarifaire le plus important est sur la zone 6 avec une fréquence révisée 56% moins importante que la fréquence actuelle, contre 52% sur les fréquences collectives. La zone 4 en revanche est la zone avec l'écart tarifaire le moins important à -27% alors qu'il était de l'ordre de -54% sur les fréquences collectives. Sur le critère de l'activité, l'allure du diagramme en barres est relativement similaire que pour les fréquences collectives, mais avec un écart type plus important de l'ordre de 0.13. L'écart tarifaire sur l'activité est compris entre -20% et -60% sur les activités 1 et 2 respectivement. Le croisement avec l'écart le plus faible est le croisement de la zone 4 et de l'activité 6 avec un écart positif de 8%. À l'inverse, le croisement de la zone 4 et de l'activité 1 est celui qui porte l'écart le plus important de -76%.

Graphiquement, les distributions des fréquences crédibilisées révisée et actuelle se superposent de la manière suivante :

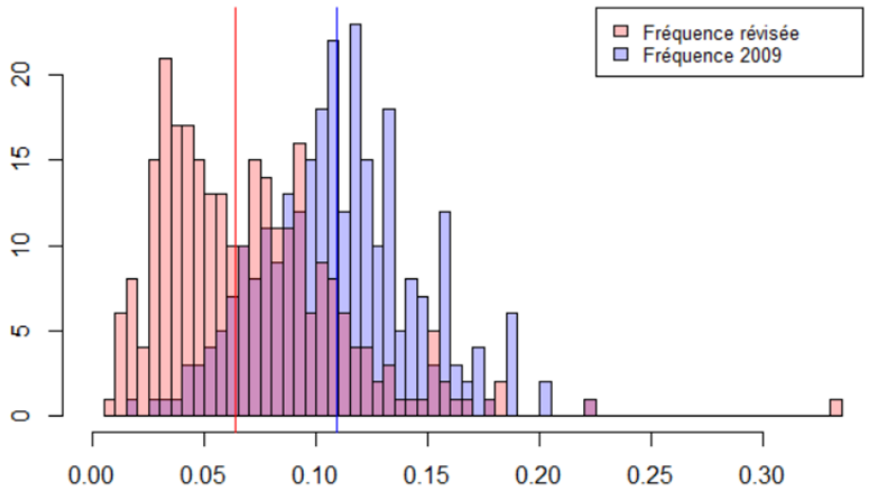

FIGURE 4.24 – Histogramme des fréquences crédibilisées révisées et actuelles

L'écart est bien moins flagrant que pour les fréquences révisées, mais tout de même important avec un écart sur les médianes, représentées par les droites verticales, de -41%.

La fréquence révisée est plus étalée autour de sa moyenne par rapport à la fréquence actuelle comme l'illustre la figure ci-dessous.

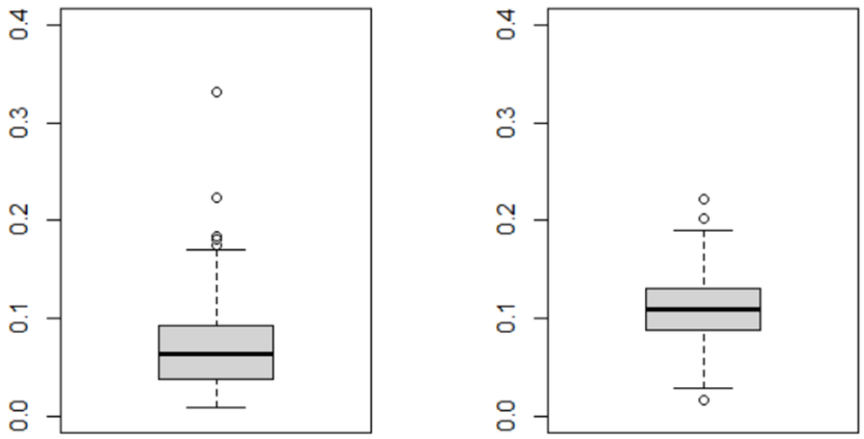

FIGURE 4.25 – Boîtes à moustache des fréquences crédibilisée révisée (gauche) et actuelle (droite)

Sur les fréquences collectives, la tendance est plutôt inverse, la fréquence collective révisée est plus resserrée autour de sa moyenne que la fréquence collective actuelle ne l'est, avec des écarts-types à 0.016 et 0.025 respectivement. Cette inversion de tendance est due à la plus forte individualisation sur le modèle révisé. Plus de poids est accordé à la fréquence individuelle, y compris sur les extrêmes.

Écarts tarifaires par tranche de risque

Pour évaluer ce nouveau modèle, il va être intéressant de se pencher sur les écarts tarifaires par tranche de risque. Cette étape permet de vérifier que les "bons contrats", les contrats avec peu de sinistralité, ne sont pas pénalisés par cette refonte. En d'autres termes, les fréquences modélisées sur les tranches de risque basses doivent rester relativement basses. Même raisonnement sur les tranches de risque hautes. La fréquence collective pénalise les bons contrats avec une fréquence modélisée en moyenne 4 fois supérieure à la fréquence observée sur les tranches de fréquence les plus basses. La crédibilité permet d'être en meilleure adéquation avec les observations et plus proche de la réalité tout en gardant une part forfaitaire. Les écarts tarifaires par tranche de risque sont présentés dans le tableau suivant :

Tranches de fréquence observée	[0, 0.025]	]0.025, 0.05]	]0.05, 0.075]	]0.075, 0.1]	]0.1,0.15]	]0.15, 0.35]
Fréquence révisée VS actuelle	-69.98%	-52.60%	-35.86%	-23.52%	-16.79%	-1.84%

TABLE 4.7 – Écarts tarifaires par tranche de fréquence

Les bornes des tranches ont été choisies de sorte à ce que les tranches ainsi formées soient homogènes en nombre de contrats. L'exposition au risque est relativement constante sur les différentes tranches, de moyenne 140 moteurs. Les bornes inférieures et supérieures (0 et 0.35) sont les bornes observées. Par rapport au tarif actuel, le modèle révisé pénalise beaucoup moins les bons contrats et reste assez équivalent sur les tranches de fréquence plus hautes. Les écarts sur les tranches les plus basses sont particulièrement importants, de près de 70% et 53% sur les deux premières tranches. Même si les écarts vont dans le sens souhaité, plus forts sur les tranches basses que sur les hautes, ils semblent très élevés sur les tranches basses et sont à prendre avec précaution. En effet, il est sage de rester prudent dans le cas de contrats qui n'auraient déclarés aucun sinistre dans le passé pour que leur fréquence modélisée ne soit pas trop basse. Cette prudence est particulièrement nécessaire sur les contrats dont l'absence de sinistre est due à leur faible exposition au risque, en nombre de moteurs ou en profondeur d'historique.

Continuité avec le tarif du segment 2

Sur les flottes de taille importante, le coefficient de crédibilité se rapproche de 1 et la tarification se rapproche donc du tarif individualisé du segment 4. De l'autre côté, vers les flottes de taille moins importante, la liaison semble moins évidente du fait que le coefficient de crédibilité est déjà relativement élevé dès les plus faibles expositions. L'idée ici est donc de vérifier la bonne jointure avec le segment 2 et de vérifier si le coefficient de crédibilité n'est pas trop important sur les flottes de petite taille. L'étude de l'écart tarifaire par tranche de fréquence sur la base des contrats à la limite entre les segments 2 et 3 peut permettre d'y voir plus clair sur cette problématique.

Pour évaluer l'écart entre la tarification forfaitaire du segment 2 et le tarif révisé, la fréquence considérée pour le S2 est la fréquence collective modélisée. En effet, sur le tarif forfaitaire du segment 2, la distinction fréquence et coût moyen n'est pas faite, le tarificateur est automatisé et la prime pure est calculée directement, il n'est donc pas possible d'extraire la composante de la fréquence. Néanmoins il est possible de considérer que la fréquence forfaitaire modélisée est suffisamment proche de la fréquence observée sur le segment 2, du fait notamment des faibles écarts moyens entre les deux fréquences.

Les résultats suivants sont à prendre avec précaution du fait du faible volume de données sur lesquels ils se basent. En effet, réduire le périmètre uniquement aux contrats à la frontière du segment 2 et du segment 3 réduit considérablement notre base de données et ne permet donc pas de faire des résultats exploités une généralité. Néanmoins la répartition sur les classes de risques reste homogène.

Tranches de fréquence observée	[0, 0.025]	]0.025, 0.05]	]0.05, 0.075]	]0.075, 0.1]	]0.1,0.15]	]0.15, 0.35]
Fréquence révisée VS forfaitaire	-55.06%	-28.02%	-13.93%	34.75%	53.46%	167.57%

TABLE 4.8 – Écarts tarifaires par tranche de fréquence sur la liaison avec le S2

Les écarts tarifaires sont plutôt élevés, d'en moyenne 60% en valeur absolue et l'impact de la crédibilité s'observe bien par un écart à la baisse sur les fréquences basses et à la hausse sur les plus élevées. A priori, au vu de ces forts écarts, la continuité avec le segment 2 n'est pas évidente. Les coefficients de crédibilité sur ce périmètre oscille entre 0.6 et 0.84 (selon la profondeur de l'historique, 1 à 4 ans). L'individualisation reste donc très forte, même sur des flottes de petite taille. L'équivalence entre le coefficient de crédibilité et l'exposition correspondante sur les flottes de petites taille est la suivante :

Coefficient de crédibilité	0.6	0.7	0.8	0.9	1
Nombre d'année d'exposition	1	2	3	6	11

TABLE 4.9 – Nombre d'année d'exposition correspondant au coefficient de crédibilité révisé sur les flottes de petite taille

Sur les petites flottes, une exposition de 4 ans suffit pour individualiser le tarif à 84%.

En comparant avec la fréquence actuelle, les écarts tarifaires avec la fréquence collective sont très élevés, de moyenne 104%. Mais l'évolution sur les différentes tranches est relativement constante, en moyenne de 0.1, soit près de 10 fois moins élevés qu'avec la fréquence révisée. Les coefficients de crédibilité sont aussi plus faibles, entre 0.1 et 0.3, l'individualisation est donc bien moins importante.

Application test : utilisation du coefficient de crédibilité actuel pour la fréquence crédibilisée révisée

Pour évaluer l'impact de la révision du coefficient de crédibilité, il est possible de considérer la fréquence suivante :

$$Freq_{test}^{cred} = z_{actuel} \times Freq^{obs} + (1 - z_{actuel}) \times Freq_{\text{révisée}}^{th}$$

Avec z_{actuel} le coefficient de crédibilité issu du modèle actuel.

À partir de cette fréquence, les écarts de tarif avec la fréquence actuelle deviennent :

Tranches de fréquence observée	[0, 0.025]	]0.025, 0.05]	]0.05, 0.075]	]0.075, 0.1]	]0.1,0.15]	]0.15, 0.35]
Fréquence test VS actuelle	-52.85%	-44.39%	-33.15%	-29.40%	-25.21%	-22.19%

TABLE 4.10 – Écarts tarifaires par tranche de fréquence entre la fréquence test et l'actuelle

Les écarts tarifaires sur les tranches basses sont légèrement réduits par rapport aux écarts avec la fréquence révisée, mais sont plus forts sur les tranches plus élevées. En moyenne l'écart tarifaire est légèrement plus important, il passe de -56% à -61%. Les écarts entre la fréquence révisée et la fréquence actuelles sont principalement portés par les écarts entre les fréquences collectives. L'individualisation impacte principalement l'écart autour de la moyenne. En effet en utilisant le coefficient de crédibilité actuel, l'individualisation est moins forte et les fréquences importantes ne sont pas captées. De manière générale la sinistralité est plus centrée autour de la moyenne, par rapport à la fréquence révisée, ce qui se constate sur le graphique suivant :

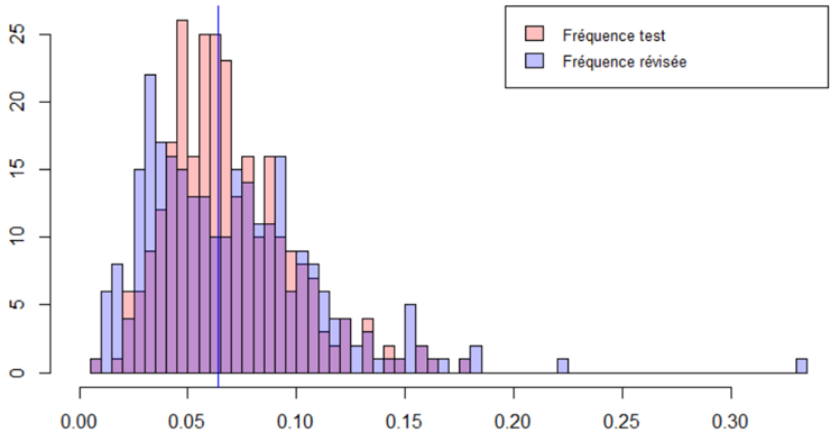

FIGURE 4.26 – Histogramme de comparaison des fréquences test et crédibilisée

D'un autre côté, la jointure entre le segment 2 et le segment 3 se fait naturellement, puisque les coefficients de crédibilité sont compris entre 0 et 0.1. L'écart entre le tarif forfaitaire et crédibilisé sur la jointure est donc de 1.8% en moyenne.

Ce test permet d'illustrer les deux pendants de la problématique de l'individualisation du tarif S3. D'une part en forçant une faible personnalisation du tarif, la jointure entre les segments 2 et 3 se fait naturellement. Aussi cela permet une prudence vis à vis des contrats qui ne déclarent aucun sinistre dans leur historique et sur lesquels la visibilité est faible. D'autre part, une individualisation plus forte permet une meilleure adéquation du tarif à la réalité de la sinistralité observée. Ce qui permet également de favoriser les "bons contrats" et de pénaliser les "mauvais contrats".

Conclusion générale

Dans le cadre de la refonte tarifaire des flottes automobiles de taille intermédiaire, les travaux présentés dans ce mémoire ont permis d'obtenir un premier modèle de fréquence crédibilisée.

Les modèles de fréquence ont été calibrés sur une base de données qui prend en compte les spécificités du marché des flottes de l'entreprise, ainsi que l'actualisation des coûts des sinistres et des sinistres survenus mais non encore déclarés. Cette même base de données pourra ultérieurement être utilisée pour la modélisation du coût moyen.

Pour la modélisation de la fréquence collective plusieurs MLG ont été confrontés. Ces MLG diffèrent par leurs lois sous-jacentes utilisées : Négative Binomiale, Poisson ainsi que des lois de comptage qui prennent en compte la masse de sinistres en 0 grâce aux modèles Zéro-Inflaté et *Hurdle*. Un modèle de *Machine Learning* a également pu être testé : la forêt aléatoire. Pour comparer ces différents modèles, des indicateurs de performance ont été utilisés : le maximum de vraisemblance, l'*AIC*, le *BIC*, la déviance, le *MSE* et l'écart au total. Finalement, au regard de ces indicateurs et des données, le modèle le plus adéquat est le MLG avec loi Négative Binomiale. Ce modèle, comme tout modèle MLG, présente l'avantage d'avoir des coefficients explicites qui expliquent la variable cible en fonction des critères tarifaires. Ces coefficients sont donc facilement modulables et peuvent par exemple être revus pour être en cohérence avec la politique de souscription.

Le coefficient de crédibilité est modélisé par les méthodes de Bühlmann-Straub et de Jewell. Ce dernier est implémenté après une classification des risques par ACM et *HCPC*. Les coefficients de crédibilité obtenus par ces méthodes dépendent de l'exposition au risque des contrats ainsi que de leur poids au portefeuille. De ces deux modèles, celui finalement retenu est le modèle de Bühlmann-Straub. En effet, l'évolution du coefficient de crédibilité en fonction de l'exposition, en sortie des deux modèles, est assez similaire. C'est pourquoi la méthode de Bühlmann-Straub est préférée ici. En effet, pour un résultat équivalent, cette méthode est plus simple d'implémentation et ne contient pas l'aléa dû à la classification. Une fois la méthode choisie et les coefficients estimés, une fonction explicite peut en être déduite. Par une estimation MLG, une fonction de tendance logarithmique prenant en paramètre l'exposition au risque permet d'obtenir un coefficient de crédibilité. Le choix d'avoir une formule explicite provient du tarif existant où le coefficient de crédibilité est une fonction qui a pour paramètre l'exposition au risque. En sortie du modèle implémenté, le coefficient révisé, théoriquement compris entre 0 et 1, commence à 0.5. Cela signifie que les plus faibles expositions sont déjà individualisées à 50%, contre 10% actuellement. Cette individualisation semble forte et pose la question de l'intérêt d'intégrer un plancher sur la fréquence crédibilisée. Ce plancher permettrait d'équilibrer le tarif, notamment dans le cas d'un contrat qui n'aurait aucun sinistre dans son historique, pour ne pas avoir une fréquence trop basse due à l'individualisation.

À ce stade de l'étude, un modèle de fréquence crédibilisée explicite a été obtenu, afin d'estimer la fréquence sinistre du portefeuille des flottes de taille intermédiaire. L'obtention de ce modèle est une première étape dans la révision du tarif du portefeuille cible. Mais avant de pouvoir implémenter cette fréquence révisée, il est nécessaire de s'assurer de sa cohérence, et notamment qu'il fasse bien la jointure avec le tarif forfaitaire des plus petites flottes et avec le tarif individuel des flottes plus importantes. Également, l'impact de cette fréquence révisée par rapport au tarif actuel est a étudier. Cette étude se traduit par l'évaluation de l'écart tarifaire sur les croisements des critères tarifaires mais aussi par tranches de fréquence observée. Cela permet notamment de vérifier que les "bons contrats" avec une fréquence observée basse ne sont pas pénalisés par rapport au tarif actuel. Au global la révision de la partie collective de la fréquence entraîne un fort écart tarifaire à la baisse : le tarif collectif révisé est moitié moins cher que le tarif collectif actuel. Cet écart s'amoindrit avec l'individualisation. Toutefois, l'individualisation du tarif entraîne un fort écart sur les basses fréquences (les "bons contrats") et un écart faible, voire nul sur les tranches de fréquences les plus élevées (les "mauvais contrats"). Ce fort écart sur les tranches de fréquences les plus faibles est à considérer avec prudence et la question de l'intégration d'un plancher est à prendre en compte ici et pourra faire l'objet d'études futures.

Dans la continuité des travaux, la mutualisation de la sinistralité grave devra également être intégrée. En ef-

fet, toute l'étude est effectuée hors-grave et ne tient pas compte des sinistres exceptionnels. Ces sinistres graves peuvent être approchés avec la théorie des valeurs extrêmes notamment.

Actuellement le coût moyen est modélisé de manière forfaitaire et ne prend pas en compte l'expérience de la flotte. Il pourrait être intéressant de crédibiliser le coût moyen, ce qui permettrait d'être plus proche encore de la réalité du risque, mais également de mieux faire la jointure avec le tarif individualisé du S4. Des méthodes de crédibilité pourraient être utilisées pour modéliser la fréquence et la sévérité conjointement.

De manière générale la sinistralité du zonier reflète toujours bien l'idée d'après laquelle il a été construit (une sinistralité croissante sur chaque zone) et lors des modélisations, il n'a donc pas été remis en cause. Cependant, il pourrait être intéressant de le revoir dans un second temps. Étudier ce zonier permettrait notamment de s'assurer que les départements qui constituent les différentes zones correspondent toujours à leur classe de risque.

Ce sujet d'individualisation progressive du tarif sur les flottes de taille intermédiaire est également repris par le laboratoire de recherche de l'Institut du Risque et de l'Assurance (IRA) du Mans dans le cadre de l'initiative de recherche sur les risques atypiques en assurance.

Bibliographie

[1] GUILLOT A. Apprentissage statistique en tarification non-vie : quel avantage opérationnel ?, 2015.

[2] MORNET A. Cours de théorie de valeurs extrêmes. *ISFA*.

[3] PIERRE A. Modélisation de la sévérité des traités en excédent de sinistre, approche par la théorie des valeurs extrêmes., 2021.

[4] GISLER A. BÜHLMANN H. A course in credibility theory and its applications. *Springer*, 2005.

[5] LECOEUR E. Institut des Actuaires. La provision IBNR. Juin 2011.

[6] ROYER M. DUFOUR A.B. Td de croisement de deux variables qualitatives. *LBBE*, 2014.

[7] GOUNO E. Cours 3 de statistiques bayésiennes. 2013.

[8] VALERI F. Surveillance d'un portefeuille de contrats flottes automobiles par l'application de la théorie de la crédibilité, 2021.

[9] BÜHLMANN H. Experience rating and credibility. *ASTIN Bulletin*, 4(3) :199–207, 1967.

[10] PAGES J. HUSSON F., JOSSE J. Principal component methods - hierarchical clustering - partitional clustering : why would we need to choose for visualizing data ? *Technical Report – Agrocampus*, 2010.

[11] MULLAHY J. Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3) :341–365, 1986.

[12] GARDES L. https://irma-web1.math.unistra.fr/ gardes/Poly_extreme.pdf (cours de théorie des valeurs extrêmes).

[13] LALOUM L. Approche hybride au calcul de la provision ibnr : méthode de fréquence / coût pour l'estimation des sinistres tardifs, 2021.

[14] MARFOQ L. Evaluation des provisions techniques en réassurance iard, dans le cadre de solvabilité 2, 2011.

[15] BARBASTE M. Une méthode de provisionnement individuel par apprentissage automatique. 2017.

[16] GUINET N. Construction d'un outil de rentabilité pour les marchandises transportées refonte du tarif sur les flottes de taille intermédiaire, 2021.

[17] VERRAL R.J. NELDER J.A. Credibility theory and generalized linear models. *ASTIN Bulletin*, 27(1) :71–82, 1997.

[18] JOHANSSON B. OHLSSON E. Credibility theory and glm revisited. 2003.

[19] THEROND P. Cours de théorie de la crédibilité. *ISFA*, 2021-2022.

[20] Pietro PARODI P. and BONCHE S. Uncertainty-based credibility and its applications. *Variance*, 4(1) :18–29, 2010.

[21] CATTEL R.B. The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2) :245–276, 1966. PMID : 26828106.

[22] RDocumentation. fitdist function.

[23] GARNIER S. Application of credibility theory to health pricing, 2013.

[24] SRA. https://www.sra.asso.fr/statistiques/rep%C3%A8res%20trimestriels/historique (historique sra de l'évolution des principaux éléments constituant le coût de la réparation des vp / vul).

[25] BENLAGHA N. VASECHKO O.A., GRUN-RÉHOMME M. Modélisation de la fréquence des sinistres en assurance automobile. *Bulletin français d'Actuariat*, 9(18) :41–63, 2009.

[26] JEWELL W.S. The use of collateral data in credibility theory : a hierarchical model. 1975.

[27] ACHIM J. ZEILIS C. and KLEIBER S. Regression models for count data in r. *Journal of Statistical Software*, 27(8) :1–25, 2008.

Annexes

Annexe A - V de Cramer et χ^2 de contingence

Le V de Cramer est une statistique qui permet d'évaluer le degré de dépendance entre 2 variables. Lorsque cette statistique est proche de 0, les variables ne sont pas liées entre elles, à l'inverse elles sont corrélées lorsque le V est proche de 1 [6].

Prenons l'exemple de 2 variables A et B qui prennent chacune 2 modalités ($p = q = 2$) et 20 observations ($n = 20$). La table de contingence observée de ce scénario est la suivante :

	a1	a2
b1	3	7
b2	4	6

La table de contingence théorique pour A et B décorrelés est une répartition uniforme des effectifs, comme suit :

	a1	a2
b1	5	5
b2	5	5

Le χ^2 se calcule alors de la manière suivante :

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

Avec n_{ij} les effectifs observés du tableau de contingence et e_{ij} les effectifs théoriques.

$$\chi^2 = \frac{(3-5)^2 + (7-5)^2 + (4-5)^2 + (6-5)^2}{5} = 2$$

Le V de Cramer se calcule ainsi :

$$V = \sqrt{\frac{\chi^2}{n \times min(p-1, q-1)}} = \sqrt{\frac{2}{20 \times min(1,1)}} = \sqrt{\frac{1}{10}} = 0.32$$

Les variables A et B ne sont, au vu de ce résultat, pas indépendantes, leur corrélation est plutôt faible.

Annexe B - Prime de Bayes

Théorème de Bayes

Soit une hypothèse H et un événement aléatoire B et $\mathbb{P}[B] > 0$, le théorème de Bayes est le suivant :

$$\mathbb{P}[H|B] = \frac{\mathbb{P}[B|H]\mathbb{P}[H]}{\mathbb{P}[B]}$$

Prime de Bayes

Quelques notations :
- $\theta \in \Theta$: le profil de risque de l'assuré, une variable aléatoire de fonction de répartition $U(\theta)$
- X_j : le risque de l'assuré sur la période j, le montant des sinistres

Conditionnellement à $\Theta = \theta$, les variables $X_1, X_2, ...$ sont i.i.d de loi F_θ. Dans le cas contraire, ils sont corrélés positivement, de covariance :

$$Cov(X_1, X_2) = \mathbb{E}[Cov(X_1, X_2|\Theta)] + Cov(\mathbb{E}[X_1|\Theta], \mathbb{E}[X_2|\Theta]) = 0 + Var[\mu(\Theta)]$$

La prime individuelle correcte, qui est une variable aléatoire, se note :

$$P^{ind} = \mathbb{E}[X_{n+1}|\Theta] = \mu(\Theta)$$

La prime collective, qui au contraire est un montant déterministe, se note :

$$P^{coll} = \mathbb{E}[X_{n+1}] = \int_\Theta \mu(\theta) dU(\theta) = \mu_0$$

Ainsi la meilleure prime d'expérience, au regard du critère de l'erreur quadratique moyenne, est la prime de Bayes qui s'écrit :

$$P^{Bayes} = \widetilde{\mu(\Theta)} = \mathbb{E}[\mu(\Theta)|X]$$

Elle est appelée meilleure prime d'expérience car, soit $\widehat{\mu(\Theta)}$ un estimateur de $\mu(\Theta)$:

$$\mathbb{E}[(\widehat{\mu(\Theta)} - \mu(\Theta))^2] = \mathbb{E}[(\widehat{\mu(\Theta)} - \widetilde{\mu(\Theta)} + \widetilde{\mu(\Theta)} - \mu(\Theta))^2] = \mathbb{E}[\mathbb{E}[(\widehat{\mu(\Theta)} - \widetilde{\mu(\Theta)} + \widetilde{\mu(\Theta)} - \mu(\Theta))^2|X]]$$

En développant et avec l'égalité : $\mathbb{E}[(\widehat{\mu} - \widetilde{\mu})(\widetilde{\mu} - \mu)|X] = \widetilde{\mu}\mathbb{E}[\widehat{\mu}|X] - \mathbb{E}[\widehat{\mu}\mu|X] - \widetilde{\mu}^2 + \widetilde{\mu}^2$ (qui se déduit du fait que $\widetilde{\mu}$ soit X-mesurable), l'équation devient :

$$\mathbb{E}[(\widehat{\mu(\Theta)} - \mu(\Theta))^2] = \mathbb{E}[(\widehat{\mu(\Theta)} - \widetilde{\mu(\Theta)})^2] + \mathbb{E}[(\widetilde{\mu(\Theta)} - \mu(\Theta))^2]$$

Ainsi, le critère de l'erreur quadratique est vérifié :

$$\mathbb{E}[(\widehat{\mu(\Theta)} - \mu(\Theta))^2] > \mathbb{E}[(\widetilde{\mu(\Theta)} - \mu(\Theta))^2]$$

Cette prime de Bayes qui est maintenant un montant certain s'interprète comme l'estimation de $\mu(\Theta)$ sachant l'historique X.

Pour déterminer la distribution *a posteriori* de $(\Theta|X)$ et ainsi la prime de Bayes, les éléments suivants sont nécessaires :
- F_θ : la loi conditionnelle de $(X_{n+1}|\Theta)$
- $U(\theta)$: la fonction de structure de portefeuille, aussi appelée *a priori*

Le choix de ces deux éléments peut se faire grâce aux classes de lois conjuguées.

Ainsi, la famille $\mathcal{U}$ est conjuguée à la famille $\mathcal{F}$ si : $\forall\ \gamma \in \Gamma$ et pour toute réalisation x de X, $\exists\ \gamma' \in \Gamma$ tel que $\forall\ \theta \in \Theta$:

$$U_\gamma(\theta|X = x) = U_{\gamma'}(\theta)$$

Quelques exemples de lois conjuguées classiques [7] :

Loi conditionnelle	Loi *a priori*	Loi *a posteriori*	Prime de Bayes
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha,\beta)$	Gamma $\mathcal{G}(\alpha+x,\beta+1)$	$\mathbb{E}[\theta\|x]$ $\frac{\alpha+x}{\beta+1}$
Binomiale $\mathcal{B}(n,\theta)$	Bêta $Beta(\alpha,\beta)$	Bêta $Beta(\alpha+n,\beta+x)$	$\mathbb{E}[\theta\|x]$ $\frac{\alpha+n}{\alpha+n+\beta+x}$
Gamma $\mathcal{G}(n,\theta)$	Gamma $\mathcal{G}(\alpha,\beta)$	Gamma $\mathcal{G}(\alpha+n,\beta+x)$	$\mathbb{E}[\theta\|x]$ $\frac{\alpha+n}{\beta+x}$
Normale $\mathcal{N}(\theta,\sigma^2)$	Normale $\mathcal{N}(\mu,\tau^2)$	Normale $\mathcal{N}\left(\frac{x}{\sigma^2}+\frac{\mu}{\tau^2},[\frac{1}{\sigma^2}+\frac{1}{\tau^2}]^{-1}\right)$	$\mathbb{E}[\theta\|x]$ $\frac{x}{\sigma^2}+\frac{\mu}{\tau^2}$

TABLE 4.11 – Lois conjuguées classiques

Cette méthode nécessite toutefois d'estimer les paramètres des lois utilisées.