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

Par : Océane LIN

Titre du mémoire :

Élaboration d'un tarif rapide pour la garantie dégât des eaux du produit Risques Industriels

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

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

Membres présents du jury de la filière :	Signature :	Entreprise : AXA France Nom : LUCAS Gérald Signature : AXA France I.A.R.D. Société Anonyme au Capital de 214 799 030 € Entreprise régie par le Code des Assurances Siège social : 313, Terrasses de l'Arche 92727 NANTERRE CEDEX 722 057 460 RCS Nanterre
		Directeur de mémoire en entreprise Nom : Sophie KRANZLIN Signature :
Membres présents du jury de l'Institut des Actuaires :	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 :

Année 2020-2021

Mémoire de fin d'études

M2 Actuariat

Océane LIN

Elaboration d'un tarif rapide pour la garantie dégât des eaux du produit Risques Industriels

Tutrice entreprise : Sophie KRANZLIN

Tutrice académique : Anne BONTOUX

Résumé

Le marché de l'assurance est un marché concurrentiel et en innovation constante. Afin de se démarquer de ses concurrents et de faire la différence sur un marché compétitif, les assureurs doivent améliorer leur segmentation régulièrement. Cependant, un bon tarif et un bon produit ne suffisent pas pour séduire la clientèle. L'attractivité est un challenge pour les compagnies d'assurance. Ainsi, les enjeux de ce mémoire seraient d'améliorer la compétitivité et la compréhension du tarif, de faciliter la souscription et de répondre rapidement aux besoins des clients. Ces problématiques sont au cœur de l'offre Risques Industriels (RI). Développer un tarif rapide permettrait d'atteindre ces objectifs. L'étude se concentrera uniquement sur la modélisation de la garantie Dégâts Des Eaux (DDE) car elle présente la sinistralité la plus élevée en termes de fréquence. Les autres garanties suivront le même raisonnement que l'étude de la garantie DDE.

Actuellement, un tarif technique et un tarif rapide existent. Le tarif rapide repose sur une liste de six questions posées au client, définies par connaissance métier et se réfère à une modalité de base pour toutes les variables du tarif technique non posées aux clients. Cette modalité de base permet de s'affranchir du problème des questions non posées au client, mais provoque un fort risque de sur ou de sous-estimation de la sinistralité. Cette observation est constatée en comparant le tarif rapide actuel et le tarif technique actuel. L'objectif de ce mémoire est alors de challenger le tarif rapide mis en œuvre, en élaborant un nouveau tarif rapide reprenant les questions posées au client, mais ne se contraignant pas à la structure tarifaire technique. Dans le cadre de la création de ce nouveau tarif rapide, des étapes de collecte et de traitement des données sont réalisées pour renforcer la qualité des données ainsi que d'assurer une modélisation fiable et performante. Ensuite, des modèles de prime pure par des modèles linéaires généralisés (GLM) en fonction de deux méthodes de zonier seront appliqués. Ces derniers vont permettre d'expliquer et de comprendre le risque étudié. Les approches considérées sont la prime pure (modèle Tweedie) et « fréquence x coût moyen ». Les méthodes de zonier sont la méthode « ajout simple du zonier » qui consiste à un croisement de deux zoniers disponibles en un unique zonier et la méthode « off set » en mettant en off set les variables tarifaires hors zoniers puis en calibrant chaque zonier séparément. Plusieurs analyses ont été effectuées pour sélectionner le meilleur modèle, en fonction des deux approches et des deux méthodes d'implémentation de zonier. Parmi les différents modèles testés, le modèle Tweedie selon la méthode « ajout simple de zonier » a été retenu. Ce dernier modèle présentait les meilleures performances.

Des analyses sur le tarif technique actuel ont également été effectuées. Les comparaisons montrent les mêmes conclusions que le tarif rapide actuel. Le tarif technique actuel sous-estime fortement la sinistralité des gros risques et surestime légèrement la sinistralité des plus petits risques. Nous allons voir tout au long de ce mémoire que cette sous-estimation est directement liée aux données. Le retraitement des variables et le choix de la méthode d'implémentation du zonier ont un impact direct dans les résultats de la modélisation.

Mots clés : Risques industriels, dommages aux biens, tarif rapide / tarif quick quote, tarif technique, dégât des eaux, modèles linéaires généralisés GLM, prime pure, fréquence, coût moyen, zonier, théorie des valeurs extrêmes.

Abstract

The insurance business is in a very competitive market and in constant innovation. In order to stand out from the competitors and make a difference in a competitive market, insurers must regularly improve their pricing segmentations. However, a good tariff and a good product are not enough to attract customers. Attractiveness is a challenge for insurance companies. Thus, the challenges of this thesis would be to improve competitiveness and understanding of the tariff, to facilitate underwriting and to respond quickly to customer needs. These issues are at the heart of the Industrial Risks offer. Developing a quick quote tariff would help achieve these goals. The study will only focus on modeling the Water Damage coverage because it has the highest frequency of claims. The other coverages will follow the same reasoning as the study of the Water Damage coverage.

Currently, a technical tariff and a quick tariff exist. The quick tariff is based on a list of six questions asked to the client, defined by business knowledge, and refers to a basic modality for all the variables of the technical tariff not asked to the clients. This basic modality makes it possible to overcome the problem of questions not asked to the client, but it causes a strong risk of over or underestimation of the claims experience. This observation is noted by comparing the current quick tariff and the current technical tariff. The objective of this thesis is to challenge the current quick tariff by developing a new quick tariff that includes the questions asked to the customer and is not constrained to the technical tariff structure. As part of the creation of this new quick tariff, data collection and preprocessing steps are performed to strengthen the quality of the data as well as to ensure a reliable and efficient modeling. Then, pure premium models by Generalized Linear Models (GLM) based on two zoning methods will be applied. These models will allow to explain and understand the studied risk. The approaches considered are pure premium (Tweedie model) and "frequency x average cost". The zoning methods are the "simple addition of zoning" method which consists in crossing two available zoning systems into a single zoning system, and the "off set" which sets all pricing variables off set except the zoning variable and then calibrate each zoning system separately. Several analyses were performed to select the best model, based on the two approaches and the two zoning implementation methods. Among the different models tested, the Tweedie model using the "simple addition of zoning" method was selected. This model showed the best performance.

Analyses on the current technical tariff were also performed. The comparisons show the same conclusions as the current quick tariff. The current technical tariff strongly underestimates the loss experience of large claims and slightly overestimates the loss experience of smaller risks. We will see throughout this thesis that this underestimation is directly related to the data. The preprocessing of the variables and the choice of the zoning implementation method have a direct impact on the modeling results.

Keywords: Industrial risks, property damage, quick quote tariff, technical tariff, Water Damage, Generalized Linear Models GLM, pure premium, frequency, average cost, zoning, extreme value theory.

Note de synthèse

Contexte

Les compagnies d'assurance sont confrontées aujourd'hui à la concurrence. Un bon tarif et un bon produit ne suffisent pas pour séduire la clientèle. L'attractivité est un challenge permanent, et la digitalisation a accentué la compétition. De nos jours, le choix de l'assureur dépend de l'attractivité de celui-ci. Par conséquent, les compagnies doivent y être représentées sur le marché et y figurer en bonne position afin de se démarquer de la concurrence. De plus, le marché de l'assurance doit s'adapter en fonction de l'évolution du risque. Le sujet de ce mémoire prend alors tout son intérêt. En effet, développer un tarif rapide aide à améliorer la compétitivité et la compréhension du tarif, à faciliter la souscription et à avoir une idée brève du tarif pour répondre rapidement aux besoins des clients. C'est un outil qui permet également de revoir la structure tarifaire, savoir s'il est nécessaire de faire un ajustement.

Problématiques et enjeux

Par définition, un tarif rapide permet d'avoir une idée brève du tarif avec une unique contrainte sur le nombre de questions à poser au client. L'objectif est de cerner le profil de risque du client avec un nombre restreint de questions. Par conséquent, certaines questions ne sont pas demandées et sont négligées. Dans le cadre du nouveau produit Risques Industriels d'AXA France, un tarif rapide a été mis en place. Cependant, outre le nombre limité de questions à poser aux clients, une nouvelle contrainte a été rajoutée : il s'agit de la contrainte de structure tarifaire. En effet, la structure tarifaire du tarif rapide doit être basée sur la structure du tarif technique[1]. De ce fait, le tarif rapide prend en compte a priori l'intégralité des informations relatives au tarif technique. Ainsi, la modélisation actuelle de ce tarif repose sur une liste de six questions posées au client[2], définies par connaissance métier et se référant à une modalité de base pour toutes les variables du tarif technique non posées aux clients. Cette modalité de base permet de pallier le problème des questions non posées au client, mais provoque un fort risque de sur ou sous-estimation de la sinistralité. Par exemple, les antécédents de sinistres ne sont pas posés aux clients pour calculer ce tarif, ainsi, il est convenu d'utiliser une modalité de référence correspondant à la modalité la plus représentée : « sans antécédents » pour tout le monde. La sinistralité prédite sera donc sous-estimée. La Direction Technique souhaite challenger le tarif rapide mis en œuvre, en élaborant un nouveau tarif rapide sans se contraindre à la structure tarifaire technique et en vérifiant la pertinence des questions posées actuellement. Il s'agit de savoir si d'une part les questions posées aux clients sont indispensables (suffisantes ou insuffisantes) pour calculer rapidement le tarif, et d'autre part d'évaluer, et de quantifier la perte engendrée par les questions non posées au client. Est-ce une plus-value de rajouter des questions intervenant dans le tarif technique ? Son abstraction permet-elle d'obtenir un tarif raisonnable ?

[1] Le tarif technique correspond à une modélisation de prime pure de chaque garantie selon une approche actuarielle ou à dire d'expert.

[2] Liste de six questions posées aux clients : l'activité de l'entreprise, l'adresse pour géocoder les zoniers, la surface, la qualité (locataire ou propriétaire), le contenu incendie et le chiffre d'affaires.

Périmètre

Ces problématiques sont au cœur de l'offre Risques Industriels (RI), avec des produits du bas et du haut de segment. Les produits RI étudiés par taille croissante de risque sont la MRP PP (MultiRisques Professionnelles Particuliers et Professionnels), MRP EN (MultiRisques Professionnels Entreprises), MPME (Multirisques Petites et Moyennes Entreprises) et la MRES (MultiRisques Entreprises Simplifiées). La frontière entre les différents produits est définie tout d'abord par le type d'activité de l'entreprise puis par des critères de surface, de chiffre d'affaires et de contenu incendie. L'assurance des risques industriels fait partie des assurances de biens et de responsabilité. Les principaux événements assurés sont tous les dommages aux biens et les dommages découlant des responsabilités incendie et autres dommages aux biens, dont : actes de vandalisme, attentats, inondations, vols, bris de machines. Les pertes d'exploitation et autres pertes financières sont également couvertes. La branche RI est un risque d'intensité, elle est caractérisée par des garanties telles que l'incendie et la perte d'exploitation dont la majeure partie de la charge est causée par très peu de sinistres. La période d'observation est de neuf ans, sur une étude fixée entre 2010 et 2018.

A noter que pour réaliser cette étude, le mémoire se concentrera uniquement sur la modélisation de la garantie Dégâts Des Eaux (DDE) car elle présente la sinistralité la plus élevée en termes de fréquence parmi toutes les garanties faisant appel à des modalités de base. Les autres garanties suivront le même raisonnement que l'étude de la garantie DDE. Le DDE est un risque réel pour l'entreprise. Toute entreprise peut être confrontée à ce risque, dont les origines peuvent être multiples. La garantie DDE en RI, couvre les dommages matériels causés directement par l'eau et consécutifs à une fuite, engorgement de conduits ou de canalisations, infiltrations accidentelles, gel des tuyaux, etc.

Analyses et retraitements des données

La collecte et la préparation des données sont des étapes importantes à ne pas négliger. Renforcer la qualité des données permet d'assurer une modélisation fiable et performante. Ainsi, un travail important de qualité de données a été mené.

Les variables explicatives à analyser et à retraiter sont les questions posées au client à savoir : l'activité, l'adresse permettant d'établir des zoniers, la surface, la qualité (locataire ou propriétaire), le contenu incendie et le chiffre d'affaires. Par ailleurs, la variable « antécédent de sinistres » sera également comptée parmi les variables explicatives afin de construire le nouveau tarif technique. Ces données sont issues de plusieurs bases :

- des bases contrats permettant de sélectionner les contrats qui ont été en cours ou résiliés sur le périmètre d'étude, à savoir les produits RI avec un historique entre 2010 et 2018
- des bases risques en fonction du produit, regroupant les données sur le risque à assurer telles que l'adresse de l'entreprise, le montant de chiffre d'affaires déclaré lors de la souscription, le contenu incendie assuré, la surface, la qualité, le code activité de l'entreprise, …
- des bases révisables apportant des informations sur la révision du chiffre d'affaires lorsque le contrat est révisable (25% des cas)
- des bases sinistres recensant la sinistralité par contrat et par année de vision. Ces bases permettront d'obtenir les antécédents de sinistres.

Les bases sont ensuite fusionnées et une réconciliation des données est réalisée avec les différents suivis RI disponibles dans la Direction Technique. Cette réconciliation a pour but de vérifier et de valider les différentes extractions de données opérées. Ensuite, des étapes de nettoyage, retraitement et analyse sont primordiales : complétion des données manquantes spécifique aux variables, catégorisation des variables quantitatives, regroupement des activités…

Les variables à expliquer sont le nombre de sinistres et la charge de sinistre dans le cadre d'un modèle de fréquence x CM. Dans le cadre de la modélisation « Prime Pure » avec l'approche Tweedie, la variable charge de sinistre est la variable à expliquer. Des spécificités sur la charge sont à considérer et à retraiter :

- les charges y compris franchises, permettant de mieux refléter le risque de l'assuré et de rendre les sinistres comparables ;
- les charges négatives, pouvant être expliquées par les recours. Ces charges ne sont pas exclues, mais forcées à zéro. Ce choix permet d'une part de garder la sinistralité en fréquence et d'autre part de modéliser la charge globale ou le coût moyen non négatif selon l'approche de modélisation utilisée ;
- la charge ultime et inflatée par produit, représentant le coût final des sinistres avec une même unité monétaire. La charge ultime sera estimée par des méthodes classiques de provisionnement. Le retraitement de la réévaluation permettra de travailler sur des distributions de charges ultimes non biaisées par l'inflation. Ainsi, il faudra réévaluer l'ensemble des charges et capitaux assurés pour les mettre en vision 2018 suivant deux indices d'indexation : l'indice RI et l'indice FFB (Fédération Française du Bâtiment) ;
- la distinction de la charge attritionnelle et grave. Le seuil préconisé par AXA qui distingue la charge attritionnelle de la charge grave est de 150 000€ pour l'ensemble des garanties de la branche RI. Or, ce seuil n'est pas forcément optimal, adapté et se relève trop important pour la garantie DDE. Ainsi, un nouveau seuil grave de 19 000€; sera déterminé à l'aide de la théorie des valeurs extrêmes (TVE). Cette dernière permettra de définir un seuil qui séparera les sinistres fréquents et peu coûteux (sinistres attritionnels) et les sinistres rares et coûteux (sinistres graves). La charge grave ou la sur-crête grave désignera les sinistres dont la charge est supérieure au seuil grave et ne dépassant pas le seuil atypique ;
- la charge atypique, regroupant les sinistres dont la charge est supérieure au seuil atypique d'AXA. Ce seuil peut être différent selon le produit, il est spécifié à 500 000€ pour les produits MRP EN et MPME, 600 000€ pour le produit MRP PP et 1 200 000€ pour le produit MRES. Ce sont des sinistres atypiques, très rares qui ne reflètent pas la sinistralité réelle du contrat, ils ne seront donc pas pris en compte dans la modélisation ;
- la charge grave mutualisée, il s'agit de mutualiser la sur-crête grave au prorata des primes acquises DDE des sinistrés. La mutualisation est appliquée sur les contrats sinistrés pour ne pas pénaliser les assurés qui n'ont pas eu de sinistres.

Modélisations

Afin de répondre à l'ensemble de ces problématiques, deux tarifs ont été créés. Un nouveau tarif rapide, basé sur les six questions posées au client et sans structure tarifaire imposée. La création de ce tarif rapide a aussi entraîné la conception d'un nouveau tarif technique dont la structure est basée sur le nouveau tarif rapide avec l'ajout de la variable « antécédent ». La conception de ces deux tarifs va permettre de réaliser différentes comparaisons entre l'actuel et le nouveau. Pour modéliser la prime pure, des modèles linéaires généralisés (GLM) ont été appliqués selon deux approches. Les approches prime pure (Tweedie) et « fréquence x coût moyen » ont été considérées avec deux méthodes d'implémentation de zonier différentes. L'approche de prime pure directe consiste à modéliser la valeur de la charge réellement observée sur tout le portefeuille. L'approche fréquence x coût moyen nécessite deux modélisations. Une modélisation de la fréquence qui correspond au nombre de sinistres sur une période d'exposition et une modélisation du coût moyen observé sur les contrats sinistrés uniquement. Ensuite, pour obtenir la valeur de la prime pure pour chaque contrat, les deux modèles sont agrégés, c'est-à-dire que les coefficients des deux modèles sont multipliés.

Plusieurs analyses ont été effectuées pour sélectionner le meilleur modèle, en fonction des deux approches et des deux méthodes d'implémentation de zonier. Parmi les différents modèles testés, le modèle Tweedie selon la méthode « ajout simple de zonier » a été retenu. Ce dernier modèle présentait les meilleures performances.

Résultats et conclusions

La méthode retenue pour le nouveau tarif rapide est l'approche prime pure (modèle Tweedie hors charges atypiques, mais y compris franchises et mutualisation des sur-crêtes graves au prorata des primes pour les contrats sinistrés). De plus, cette modélisation est caractérisée par l'implémentation du zonier selon la méthode d'ajout simple du zonier (croisement des deux zoniers disponibles en un zonier). Les analyses graphiques et les indicateurs de performance ont permis d'une part de valider les modèles et d'autre part de juger de la pertinence des six questions posées au client. Les performances des modèles ont été examinées à partir des courbes de Lorenz et Lift, du spread, des résidus, de l'indice de Gini, et de l'erreur quadratique moyenne. Les deux nouveaux tarifs sont considérés comme plus performants que les deux tarifs actuels. En effet, en challengeant le tarif rapide actuel et le nouveau tarif rapide, l'une des premières problématiques était d'étudier l'impact sur le tarif rapide de s'imposer la structure tarifaire du tarif technique. L'existence des modalités de base dans le tarif rapide est-elle problématique ? En comparant le tarif rapide actuel et le tarif technique actuel, l'utilisation des modalités de base entraîne une forte sous-estimation de la sinistralité. En effet, par simplicité, le tarif rapide actuel utilise les mêmes coefficients générés par le tarif technique actuel. Or, en appliquant les mêmes coefficients, le tarif rapide actuel est largement sous-estimé car généralement la modalité de base est la moins sinistrée et la plus représentée. De plus, cette sous-estimation est d'autant plus visible lorsque dès le départ le tarif technique sous-estime la sinistralité. Pour intégrer la modalité de base de façon plus appropriée, il faudrait recalibrer le modèle, qui réhausserait les coefficients des autres modalités. Ainsi, les conclusions obtenues sont en faveur du nouveau tarif rapide sans contrainte de structure tarifaire. En effet, la modélisation actuelle ne retraite pas les valeurs manquantes ou aberrantes (valeurs saisies à zéro et correspond à une part non négligeable des données) pour les variables comme le chiffre d'affaires, le contenu incendie et la surface. Par exemple, un chiffre d'affaires nul est considéré comme une valeur aberrante puisque dès lors qu'une entreprise souscrit une assurance, elle exerce une activité normale et courante pour payer sa cotisation. Ne pas avoir traité les chiffres d'affaires nuls a un impact sur la modélisation actuelle, car cette variable est considérée comme non tarifaire, mais est réellement tarifaire si des retraitements sont réalisés. Ainsi, des propositions de retraitements seront présentées pour ces variables lors de cette étude. Cela nous permettra de voir que la qualité des données a une contribution importante dans la modélisation. Finalement, les variables discriminantes retenues dans le nouveau tarif rapide sont l'activité, le zonier, la surface, le contenu incendie et le chiffre d'affaires.

Le deuxième enjeu lié à ce sujet est de quantifier et d'évaluer la perte liée aux questions non posées, mais intervenant dans le tarif technique. Est-ce une plus-value de les ajouter parmi les questions posées au client ? Dans le cadre de la modélisation de la garantie DDE, la variable étudiée est les antécédents de sinistres. Bien que les antécédents expliquent un signal important, sa suppression permet tout de même d'obtenir un tarif raisonnable, proche de la prime pure observée. La comparaison entre la prime pure du nouveau tarif rapide et du nouveau tarif technique a permis de conclure que l'écart est faible, donc la perte engendrée par cette question non posée est faible. Cela permet de déduire également que les six questions sont suffisantes pour établir un tarif rapide de la garantie DDE.

Un tableau récapitulatif des indicateurs numériques des quatre modèles sur la base de validation :

	Base de validation					
	Tarif rapide actuel	**Nouveau tarif rapide**	Ecart	Tarif technique actuel	**Nouveau tarif technique**	Ecart
RMSE	752,8	**751,3**	-1,5	751,4	**751,2**	-0,2
Gini	49,24%	**52,12%**	2,88 points	50,64%	**53,24%**	2,6 points

Tableau - Comparaison des indicateurs numériques de performance des modèles actuels et nouveaux

Le tarif rapide actuel possède un Gini de 49,24% et le **nouveau tarif rapide** possède un Gini de **52,12%**. Avec les différents retraitements réalisés sur la base de données, le tarif est mieux segmenté et plus performant avec la nouvelle modélisation. Il y a également une amélioration de la segmentation dans le tarif technique : le tarif technique actuel a un Gini de 50,64% contre **53,24%** pour **le nouveau tarif technique**. Les nouvelles modélisations apportent donc une meilleure segmentation du tarif avec un gain de presque trois points en Gini et des RMSE légèrement plus faibles. Cet apport n'est pas négligeable et confirme l'importance du retraitement et de la qualité des données. De plus, d'autres analyses sur la base de validation permettront de valider les modèles construits et d'apprécier leur robustesse. Cependant, le nouveau tarif technique est basé sur la structure du nouveau tarif rapide, c'est-à-dire les six questions posées au client. Il serait plus judicieux de ne pas conditionner les variables en entrée du modèle aux questions uniquement.

Le schéma ci-dessous synthétise les résultats finaux des différentes comparaisons de prime pure :

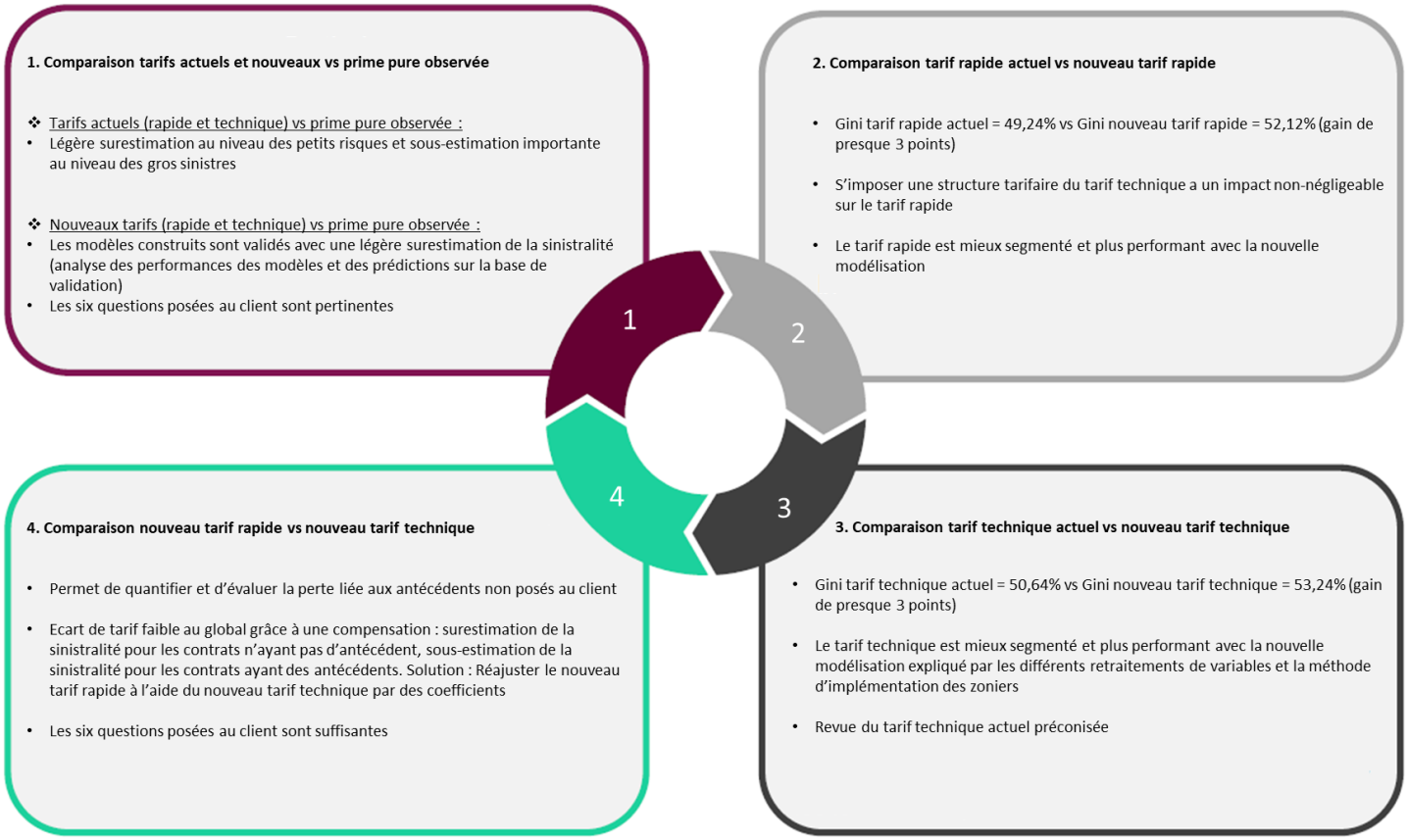


Figure - Schéma synthétique des différentes comparaisons de prime pure

Executive summary

Context

Nowadays, insurance companies are faced with competition. A good tariff and a good product are not enough to attract customers. Attractiveness is a permanent challenge, and digitalization has increased the competition. The choice of insurer depends on their attractiveness. Therefore, companies need to be represented in the market and to be in a good position in order to stand out from the competition. Moreover, the insurance market must adapt to the evolution of the risk. This is where the thesis topic comes into its own. Indeed, developing a quick tariff helps to improve competitiveness and understanding of the tariff, to facilitate underwriting and to have a brief idea of the tariff to quickly meet the customer needs. It is also a tool that allows to review the tariff structure and to know if it is necessary to make an adjustment.

Problems and issues

By definition, a quick tariff allows to have a brief idea of the rate with a single constraint on the number of questions asked to the customer. The objective is to identify the client's risk profile with a limited number of questions. As a result, some questions are not asked and are overlooked. In the context of AXA France's new Industrial Risks product, a quick tariff was implemented. However, in addition to the limited number of questions to be asked to customers, a new constraint has been added: it is the tariff structure constraint. Indeed, the tariff structure of the quick tariff must be based on the structure of the technical tariff[3]. As a result, the quick tariff takes into account all the information related to the technical tariff. Thus, the current modeling of this tariff is based on a list of six questions asked to the clients[4], defined by business knowledge, and refers to a basic modality for all the variables of the technical tariff not asked to the clients. This basic modality makes it possible to overcome the problem of questions not asked to the customer, but it causes a strong risk of underestimation of the claims experience. For example, clients are not asked about their claims history in order to calculate this tariff, so it is agreed to use a reference modality corresponding to the most represented modality: "no claims history" for everyone. The predicted claims/loss experience will therefore be underestimated. The Technical Department wishes to challenge the implemented quick tariff, by developing a new quick tariff without being constrained by the technical tariff structure and by checking the relevance of the questions currently asked. It is a question of knowing whether the questions asked to customers are essential (sufficient or insufficient) to calculate the tariff quickly, and also of evaluating and quantifying the loss caused by the questions not asked to the customer. Is it an added value to add questions to the technical tariff? Does its abstraction allow us to obtain a reasonable tariff?

Perimeter

[3] The technical tariff corresponds to a pure premium model of each guarantee according to an actuarial approach or expert opinion.

[4] List of six questions asked to the clients: business activity, address to geocode the zoning variable, surface area, quality (tenant or owner), fire content and turnover.

These issues are at the heart of the Industrial Risks (IR) offer, with products at the bottom and top of the segment. The industrial risk (IR) products studied by increasing size of risk are MRP PP (MultiRisques Professionnelles Particuliers et Professionnels), MRP EN MultiRisques Professionnels Entreprises), MPME (Multirisques Petites et Moyennes Entreprises) et la MRES (MultiRisques Entreprises Simplifiées). The boundary between the different products is defined firstly by the type of activity of the company and then by criteria of surface area, turnover and fire content. Industrial risk insurance is part of property and liability insurance. The main insured events are all damages to property and damages resulting from fire and other property liabilities, including: vandalism, attacks, floods, theft, machinery breakdown. Business interruption and other financial losses are also covered. The industrial risk branch is an intensity risk, characterized by coverages such as fire and business interruption where the majority of the expenses is caused by very few claims. The observation period is nine years, over a study set between 2010 and 2018.

It should be noted that in order to carry out this study, the study will only focus on the modeling of the Water Damage (WD) coverage because it has the highest frequency of claims among all coverages using basic modalities. The other coverages will follow the same reasoning as the study of the WD coverage. WD is a real risk for the company. Any company can be confronted with this risk, whose origins can be multiple. The WD guarantee in IR, covers material damages caused directly by water and consecutive to a leak, clogging of conduits or pipes, accidental infiltrations, freezing of pipes, etc.

Data analysis and preprocessing

Data collection and preprocessing are important steps that should not be overlooked. Strengthening the quality of the data allows to ensure a reliable and efficient modeling. Thus, an important work of data quality has been carried out.

The explanatory variables to be analyzed and preprocessed are the questions asked to the client, namely: business activity, address allowing the establishment of the zoning variable, surface area, quality (tenant or owner), fire content and turnover. In addition, the "claims history" variable will also be counted among the explanatory variables in order to construct a new technical tariff. These data come from several databases:

- contract databases allowing the selection of contracts that have been in progress or terminated within the scope of the study, i.e. IR products with a history between 2010 and 2018
- risk databases according to the product, gathering data on the risk to be insured such as the company's address, the amount of turnover declared at the time of subscription, the insured fire contents, the surface area, the quality, the company's activity code, etc.
- revisable databases providing information on the revision of the turnover when the contract is revisable (25% of cases)
- claims databases listing the number of claims per contract and per year of vision. These databases will make it possible to obtain the claims history.

The databases are then merged and the data is reconciled with the various IR monitoring systems available in the Technical Department. The purpose of this reconciliation is to verify and validate the various data extractions performed. Then, cleaning, preprocessing and analysis steps are essential: completion of missing data specific to the variables, categorization of quantitative variables, grouping of activities, etc.

The variables to be explained are the number of claims and the claims expenses in a “frequency x average cost” model. In the framework of the "Pure Premium" modeling with the Tweedie approach, the claims expenses variable is the variable to be explained. Specificities on the claim expenses have to be considered and restated:

- claims expenses including deductibles, which can better reflect the risk of the insured and make the claims comparable;
- negative claims expenses, which can be explained by recourse. These claims expenses are not excluded, but forced to zero. This choice makes it possible to keep the frequency of claims and to model the overall expense or the non-negative average cost, depending on the modeling approach used;
- the ultimate and inflated cost per product, representing the final cost of the claims with the same monetary unit. The final claims expenses will be estimated using traditional reserving methods. The revaluation adjustment will allow to work on final cost distributions, which are not biased by inflation. Thus, it will be necessary to revalue all the expenses and insured capital to put them in the 2018 vision according to two indexation indices: the IR index and the FFB index (*Fédération Française du Bâtiment*);
- the distinction between attritional and large claims cost. The threshold recommended by AXA which distinguishes the attritional load from the large claims cost is € 150,000 for all the guarantees of the IR branch. However, this threshold is not necessarily optimal and is too high for the WD coverage. Thus, a new severe threshold of € 19,000 will be determined, using the extreme value theory (EVT). The latter will allow to define a threshold that will separate frequent and low-cost claims (attritional claims) and rare and high-cost claims (large claims). Large claims expenses or large over-peak will refer to claims with expenses above the severe threshold and not exceeding the atypical threshold;
- Atypical claims expenses, which includes claims with a load above AXA's atypical threshold. This threshold can be different depending on the product, it is specified at € 500,000 for the MRP EN and MPME products, €600,000 for the MRP PP product and € 1,200,000 for the MRES product. These are atypical, very rare claims that do not reflect the real loss experience of the contract and will therefore not be taken into account in the modeling;
- the mutualized severe cost, which consists in mutualizing the severe excess of loss on a pro-rata basis of the WD earned premiums of the claimants. The mutualization is applied to the contracts with claims in order not to penalize the policyholders who have not had any claims.

Modelling

In order to respond to all these issues, two tariffs have been created. A new quick tariff, based on the six questions asked to the client and without an imposed tariff structure. The creation of this quick tariff also led to the design of a new technical tariff whose structure is based on the new quick tariff with the addition of the "claims history" variable. The design of these two tariffs will allow for different comparisons between the current and the new tariff. To model the pure premium, generalized linear models (GLM) were applied using two approaches. The pure premium (Tweedie) and "frequency x average cost" approaches were considered with two different zonal implementation methods. The direct pure premium approach involves modeling the value of the load actually observed over the entire portfolio. The frequency x average cost approach requires two modeling runs. A frequency model which corresponds to the number of claims over a period of exposure and a model of the average cost observed on the contracts that have been claimed. Then, to obtain the value of the pure

premium for each contract, the two models are aggregated, i.e. the coefficients of the two models are multiplied.

Several analyses were performed to select the best model, based on the two approaches and the two zoning implementation methods. Among the different models tested, the Tweedie model using the "simple addition of the zoning variable" method was selected. This model showed the best performance.

Results and conclusions

The method chosen for the new quick tariff is the pure premium approach (Tweedie model excluding atypical charges, but including deductibles and pooling of large over-peak on a pro-rata basis to premiums for loss contracts). In addition, this modeling is characterized by the implementation of the zonier according to the “simple addition of the zoning” method (crossing the two available zoning variables into one zoning variable). Graphical analyses and performance indicators were used to validate the models on the one hand and to judge the relevance of the six questions asked to the client on the other. The performance of the models was examined on the basis of the Lorenz and Lift curves, the spread, the residuals, the Gini index, and the mean squared error. The two new tariffs are considered to be more efficient than the two current tariffs. Indeed, in challenging the current fast tariff and the new fast tariff, one of the first issues was to study the impact on the fast tariff of imposing the tariff structure of the technical tariff. Is the existence of the basic terms in the fast tariff problematic? When comparing the current fast tariff and the current technical tariff, the use of basic terms and conditions results in a significant underestimation of claims. Indeed, for simplicity's sake, the current quick tariff uses the same coefficients generated by the current technical tariff. However, by applying the same coefficients, the current quick tariff is largely underestimated because the basic modality is generally the least claimed and the most represented. This underestimation is even more apparent when the technical tariff underestimates claims from the start. In order to integrate the basic modality more appropriately, the model would have to be recalibrated, which would increase the coefficients of the other modalities. Thus, the conclusions obtained are in favor of the new fast tariff without tariff structure constraints. We will see later that this underestimation is directly linked to the data. Indeed, the current modeling does not preprocess missing or outlier values (values entered at zero and corresponding to a non-negligible part of the data) for variables such as turnover, fire content and area. For example, zero turnover is considered an outlier because once a company buys insurance, it is conducting normal, routine business to pay its premium. Not having treated zero turnovers has an impact on the current modeling, as this variable is considered non-rate, but is actually rate if adjustments have been made. Thus, proposed adjustments will be presented for these variables in this study. This will allow us to see that data quality has an important contribution in the modeling. Finally, the discriminant variables retained in the new quick tariff are business activity, zoning, surface area, fire content and turnover.

The second issue related to this topic is to quantify and evaluate the loss related to the questions not asked, but involved in the technical tariff. Is there any added value in adding them to the questions asked to the customer? In the context of modeling WD coverage, the variable studied is “claims history”. Although history explains an important signal, its removal still allows us to obtain a reasonable tariff, close to the observed pure premium. Comparing the pure premium of the new quick tariff with the new technical tariff, we conclude that the difference is small, so the loss generated by this unasked question is small. This also leads to the conclusion that the six questions are sufficient to establish a quick tariff for WD coverage.

A summary table of the numerical indicators of the four models on the validation basis:

	Validation basis					
	Current quick quote tariff	**New quick quote tariff**	Difference	Current technical tariff	**New technical tariff**	Difference
RMSE	752,8	**751,3**	-1,5	751,4	**751,2**	-0,2
Gini	49,24%	**52,12%**	2,88 points	50,64%	**53,24%**	2,6 points

Table - Comparison of numerical performance indicators for current and new models

The current quick quote tariff has a Gini of 49.24% and the new quick quote tariff has a Gini of 52.12%. With the various adjustments made to the database, the tariff is better segmented and performs better with the new modeling. There is also an improvement in segmentation in the technical tariff: the current technical tariff has a Gini of 50.64% compared to 53.24% for the new technical tariff. The new models therefore provide a better segmentation of the tariff with a gain of almost three points in Gini and slightly lower RMSEs. This contribution is not negligible and confirms the importance of preprocessing and data quality. Moreover, other analyses on the validation basis will allow to validate the models built and to assess their robustness. Nevertheless, the new technical tariff is based on the structure of the new quick rate, i.e., the six questions asked to the customer. It would be better to not condition the input variables of the model on the questions only.

The diagram below summarizes the final results of the different pure premium comparisons:

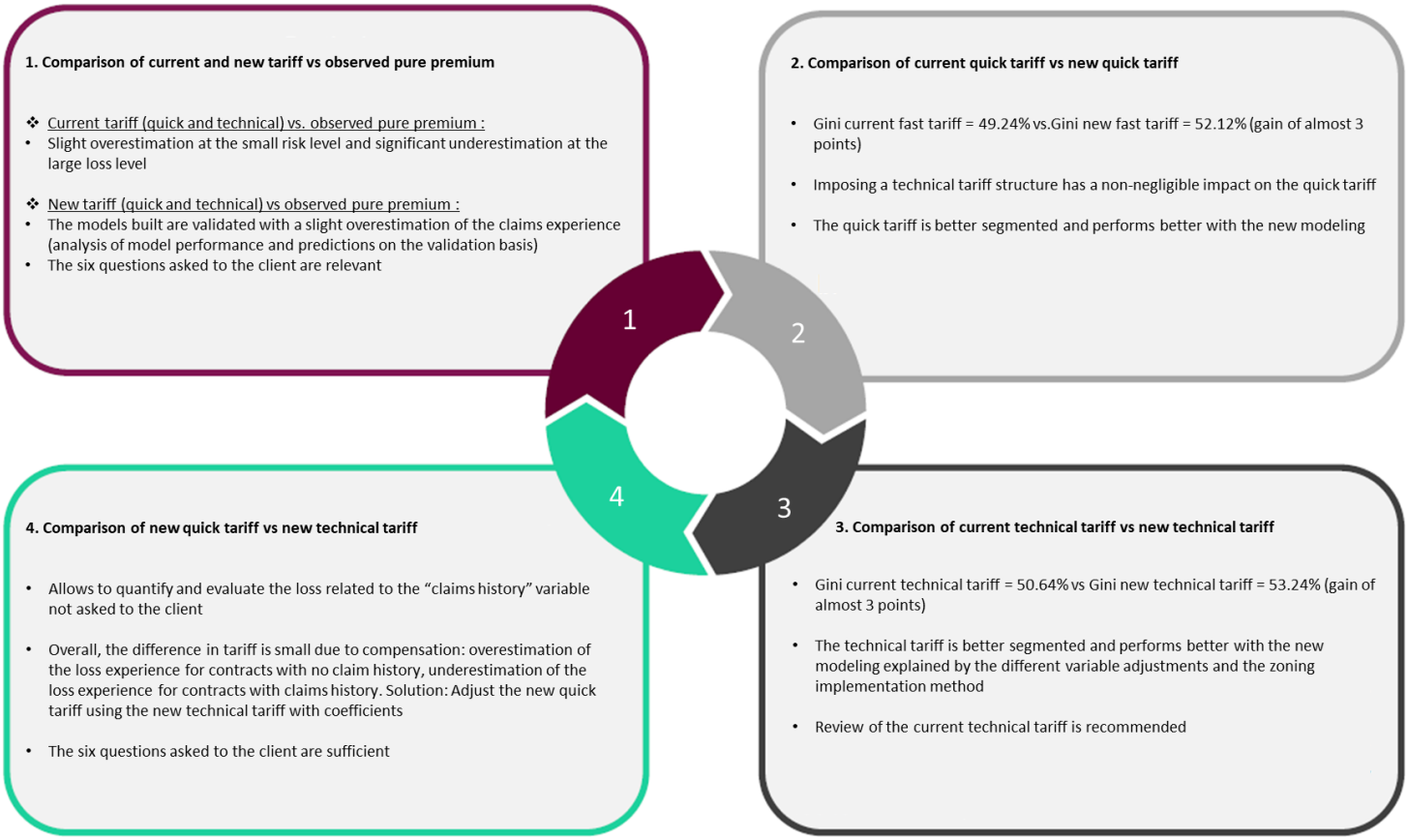


Figure - Summary diagram of the different pure premium comparisons

Remerciements

Je tiens avant tout à remercier Sophie KRANZLIN, ma tutrice en entreprise et conseiller d'études actuarielles en Risques Industriels à la Direction Actuariat Pilotage Entreprises (DAPE). Son encadrement, ses conseils, sa disponibilité, ses encouragements et ses relectures m'ont permis d'accomplir pleinement mes missions en tant qu'apprentie chargée d'études actuarielles et de réaliser ce mémoire de fin d'études.

Je remercie également toute l'équipe Risques Industriels du marché des entreprises et professionnels, Eliane KABORE KONATE, Franck PEKA KADJE, Khalil OUHABI, Oksana ALLAIRE de m'avoir éclairée, aidée sur mon sujet et encouragée tout au long de la réalisation de ce mémoire.

J'adresse mes remerciements à Gérald LUCAS, responsable du marché des entreprises en Risques Industriels, Construction et Responsabilité Civile à la DAPE et Mohamed HALIMI, responsable du marché des professionnels pour leur expertise, leur disponibilité et les nombreux conseils octroyés pour mes missions et mon mémoire. Je remercie aussi Loïc CHENU, ancien responsable du marché entreprises de m'avoir accompagnée et suivie les premiers mois de mon alternance.

Des remerciements à Véronique MARPILLAT de m'avoir accordé sa confiance et accueillie dans le service.

Merci à tous les membres de l'équipe DAPE pour leur accueil et leur bienveillance qui m'a permis de m'intégrer très rapidement dans ce nouvel environnement. Une très bonne ambiance dans l'équipe est maintenue malgré la situation sanitaire grâce à de nombreuse activités soutenues par notre manager Gérald.

Je souhaiterais aussi remercier toute l'équipe pédagogique de l'ISUP pour avoir assuré les cours théoriques et pour nous avoir donné l'occasion de réaliser une alternance de début d'expérience pour le métier d'actuaire. Je remercie plus particulièrement ma tutrice académique Anne BONTOUX pour sa disponibilité, ses remarques et ses conseils.

Je remercie également Nathalie OBERT-BEN-TAIEB et Alexia GONZALEZ, de nous avoir aidés à trouver une alternance, en nous envoyant plusieurs offres.

Enfin, je tiens à remercier toutes les personnes qui m'ont conseillée et relue lors de la rédaction de ce mémoire.

Table des matières

Introduction

Ce mémoire a été réalisé chez AXA France dans la Direction Actuariat et Pilotage Entreprises (DAPE), au sein de l'équipe Risques Industriels (RI) pour les entreprises, dont l'objet est l'élaboration d'un tarif quick quote (tarif rapide) afin de répondre rapidement aux besoins de nos clients.

L'assurance des Risques Industriels fait partie des assurances de biens et de responsabilité. Elle permet de couvrir en cas de sinistre, les biens assurés de l'entreprise, tant sur le plan matériel (bâtiments, matériels, mobiliers, marchandises, …) qu'immatériel/financier (frais consécutifs, perte de marge brute, …).

L'assurance protège contre les risques de la vie quotidienne. Cependant, le contexte de risque évolue rapidement. Par conséquent, le marché de l'assurance s'adapte en fonction de l'évolution du risque : en ajustant son tarif, ses garanties ou même en ajoutant des exclusions. Face à cette évolution, le sujet du mémoire prend tout son intérêt. En effet, un tarif quick quote est un outil aidant d'une part à répondre rapidement aux besoins des clients et d'autre part permettant de revoir la structure tarifaire, savoir s'il est nécessaire de faire un ajustement. Son développement permet également d'améliorer la compétitivité, la compréhension, de faciliter la souscription et d'avoir une idée rapide du tarif.

Actuellement, un **tarif technique** et un **tarif quick quote** par garantie du produit d'assurance Risques Industriels sont calibrés. Le **tarif technique** correspond à une modélisation de prime pure de chaque garantie selon deux approches :

- Actuarielle : modèle de crédibilité pour l'incendie et modèle de fréquence * coût moyen pour les autres garanties majeures,
- A dire d'expert : pour les plus « petites garanties » où l'on dispose de peu de données ou portant plus de marge.

Par définition, un **tarif quick quote** est un tarif rapide, donc il est recommandé de poser le moins de questions possible. Par conséquent, certaines questions ne sont pas posées au client. La modélisation actuelle du tarif rapide doit respecter deux contraintes, une contrainte de questions posées au client et une contrainte de structure tarifaire technique. Cette dernière repose sur une liste de six questions posées aux clients, définies par connaissance métier, et se référant à une modalité de base pour toutes les variables du tarif technique non posées aux clients. Cette modalité de référence permet de pallier le problème des questions non posées aux clients. Par exemple, les antécédents de sinistre ne sont pas posés aux clients pour calculer le tarif rapide, ainsi, il est convenu d'utiliser une modalité de référence, correspondant à la modalité la plus représentée : « sans antécédents » pour tout le monde. La Direction Technique souhaite challenger le tarif rapide mis en place, en élaborant un nouveau tarif rapide avec les six questions, mais sans la contrainte de structure tarifaire imposée, et en vérifiant la pertinence des questions posées actuellement. Il s'agit de savoir si d'une part les questions posées aux clients sont indispensables (suffisantes ou insuffisantes) pour calculer rapidement le tarif, et d'autre part d'évaluer, et de quantifier le risque engendré par les modalités de base. Son existence est-elle problématique ? Est-ce une plus-value de les ajouter parmi les questions posées aux clients ? Sa suppression permet-elle d'obtenir un tarif raisonnable ? Quel est l'impact sur le tarif rapide de s'imposer une structure tarifaire du tarif technique ? Quelle est la perte liée aux questions non posées mais intervenant dans le tarif technique ? Une série de questions se posent à ce sujet, elles seront répondues tout au long du mémoire.

Pour le produit d'assurance RI, dans le cadre du tarif rapide, sept garanties font intervenir des variables tarifaires dont certaines ne sont pas posées sous forme de question au client. Or, il est nécessaire d'avoir toutes les variables tarifaires renseignées pour calculer le tarif, ainsi, des modalités de base sont utilisées pour compléter les informations non posées au client. Sur ces sept garanties, six d'entre-elles utilisent une modalité de base pour les antécédents. Pour étudier l'impact de cette modalité de base, nous nous sommes intéressés à la garantie Dégâts Des Eaux (DDE) car elle présente la sinistralité la plus élevée, en termes de fréquence, parmi toutes les garanties faisant appel à des modalités de base. Les autres garanties suivront le même raisonnement que l'étude de la garantie DDE.

Ce mémoire amène donc à modéliser la garantie DDE. Pour ce faire, une étape de sélection des variables est revue et deux approches de modélisation sont appliquées, des modèles linéaires généralisés (fréquence x coût moyen) et Tweedie, avec une contrainte des variables choisies parmi les questions posées au client. A partir de ce nouveau tarif rapide, une étude approfondie est réalisée pour déterminer son efficacité et sa raisonnabilité sur les deux tarifs actuels.

La première partie de ce mémoire est consacrée à la présentation du contexte, le périmètre de l'étude et les enjeux qui lui sont liés. Ensuite, un chapitre détaillant le processus de la création de la base de modélisation, puis une section rappelant quelques aspects théoriques utiles pour la modélisation, enfin une dernière partie sera accordée à la modélisation de la prime pure et la comparaison des différentes primes pures.

CHAPITRE 1

1 Contexte, objectifs et périmètre de l'étude

1.1 Présentation de l'assurance de dommages

Au sein du secteur de l'assurance, les directives européennes distinguent deux branches :
- la branche vie (assurances vie, décès, bons de capitalisation, fonds de retraite)
- la branche non-vie

Selon une classification habituelle de la profession, les produits commercialisés relèvent soit de l'assurance de personnes, soit de l'assurance de dommages (ou IARD "Incendie Accident et Risques Divers").

L'assurance IARD comprend
- les assurances de biens, qui couvrent un risque relatif à un élément d'actif patrimonial
- les assurances de responsabilité, qui couvrent les dettes liées à l'obligation de réparer les dommages causés à autrui, y compris éventuellement les dommages corporels

Les principales assurances de dommages sont :
- l'assurance des biens particuliers (contrats MultiRisques Habitation MRH)
- l'assurance des biens professionnels (Risques industriels Entreprise, agriculteurs, commerçants, artisans et prestataires de services, collectivités locales…)
- l'assurance de construction
- l'assurance automobile
- l'assurance de transports (assurances ferroviaire, maritime, fluviale, aérienne, spatiale, marchandises transportées)
- l'assurance de responsabilité civile
- l'assurance crédit
- l'assurance de protection juridique

1.2 Présentation de l'assurance des Risques Industriels (RI)

Dans ce mémoire, nous nous concentrons uniquement sur l'assurance des Risques Industriels qui fait partie de l'assurance de dommages aux biens professionnels.

L'assurance des Risques Industriels fait partie des assurances de biens et de responsabilité.
Le concept de l'assurance des Risques Industriels est apparu à la naissance de l'assurance des dommages aux biens à la fin du XVIIème siècle, après le grand incendie de Londres de 1666 qui a causé la destruction de plus de 13 000 bâtiments.

Les principaux évènements assurés sont tous les dommages aux biens et les dommages découlant des responsabilités incendie et autres dommages aux biens, dont : actes de vandalisme, attentats, inondations, vols, bris de machines. Les pertes d'exploitation et autres pertes financières sont également couvertes.

La couverture peut inclure :

- Garanties de base (Socle) : Incendie (INC), Catastrophes naturelles (CN), Événements climatiques (EVTCL), Dommages électriques (DEL) ;
- Garanties optionnelles : Dégâts des eaux et gel (DDE), Vol (y compris les détériorations), Bris des glaces (BDG), Bris de machines (BDM, y compris tous risques informatiques), Perte de marchandises en installation frigorifique (MIF), Dommages aux marchandises et matériels transportés (MMT), Perte d'exploitation (PE), perte de revenus, Responsabilité Civile Professionnelle, Perte d'exploitation limitée, Garantie automatique des investissements, Frais supplémentaires additionnels, Responsabilité Civile propriétaire des locaux, Pénalités de retard, Perte de valeur du fonds de commerce, Carence de fournisseurs, Pénalités de retard, Indemnités de licenciement, Protection Juridique, Assistance.

La garantie incendie comprend elle-même plusieurs sous garanties : Incendie, explosion, vandalisme, risques divers, responsabilité liée à l'occupation des locaux, dommages lors des salons, foires et manifestations, frais de reconstitution d'archives, frais consécutifs, effondrement à la suite d'une cause externe, dommages lors des salons, foires et manifestations, attentats et actes de terrorisme.

Dans une branche à déroulement lent, les indemnisations surviennent plusieurs années après la survenance du sinistre. Il est possible que la sinistralité finale des polices existantes ne se soit pas encore complètement réalisée. Au contraire, pour une branche à développement court, les remboursements se font rapidement. Par exemple, le risque à responsabilité civile et le risque de construction sont considérés comme étant à développement long alors que le remboursement de soins courants en branche santé est considéré comme à développement court. Le risque industriel de la garantie DDE peut être considéré comme un risque dommages de biens et serait donc à classer comme un risque à développement court. Cependant, la garantie incendie peut être considérée comme un risque à développement court comme long car l'indemnisation des plus gros incendies peut prendre plusieurs années.

La branche RI est un risque d'intensité, elle est caractérisée par des garanties telles que l'incendie et la perte d'exploitation dont la majeure partie de la charge est causée par très peu de sinistres.
L'indemnisation en cas de sinistres est plus complexe, elle peut se calculer de 2 manières :

- En fonction des biens touchés et des valeurs déclarées qui sont les "Capitaux assurés"
- En fonction du sinistre Maximum Possible (SMP), qui va dépendre en plus du capital assuré, de l'éloignement des différents sites ou bâtiments ainsi que des critères de prévention et protection mis en place par l'entreprise.

Cette distinction est faite car l'ensemble des capitaux assurés ne sont pas toujours endommagés en totalité à la suite d'un unique sinistre.

1.3 Positionnement d'AXA France sur la branche RI par rapport au marché

La Fédération Française de l'Assurance (FFA) publie chaque année une synthèse d'un questionnaire, rempli par une majorité de sociétés d'assurance, sur la production de contrats et sinistres de la branche Dommages Aux Biens (DAB).

Ci-dessous la répartition des cotisations en 2019 sur le marché de l'assurance des entreprises, accompagnée d'un zoom sur le marché des DAB.

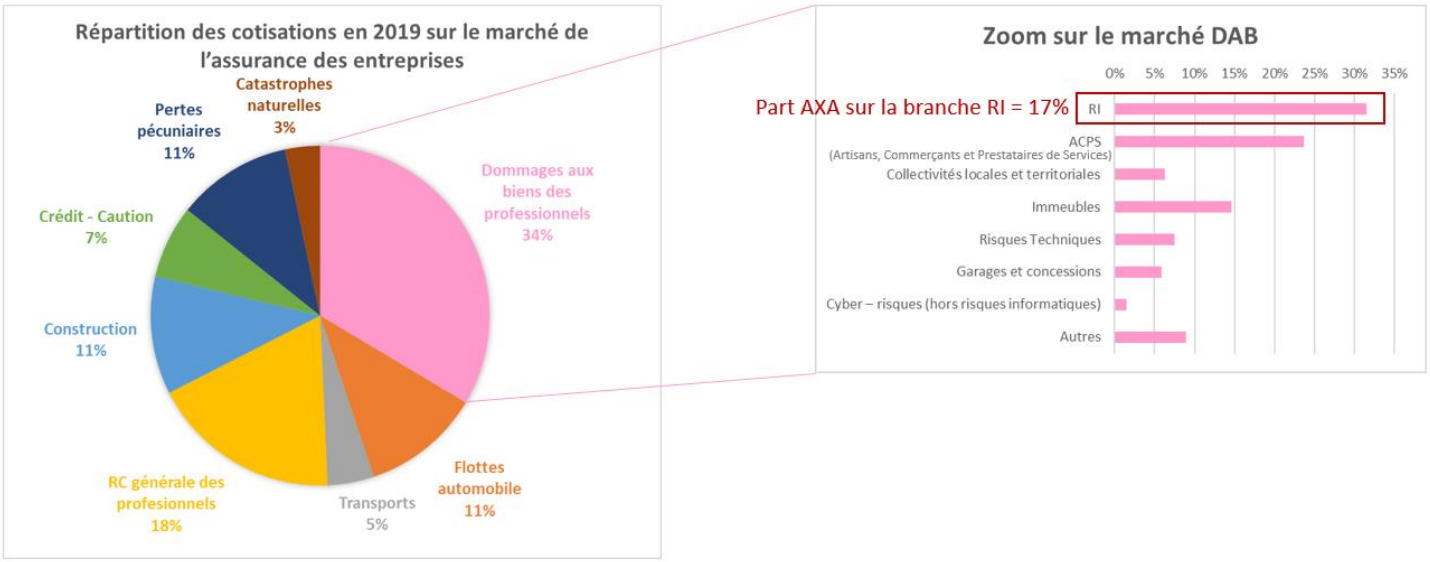


Figure 1.3.1 – Répartition des cotisations en 2019 sur le marché de l'assurance des entreprises, accompagnée d'un zoom sur le marché des DAB

En 2019, l'assurance dommages aux biens des professionnels représente 34% des primes sur tout le marché de l'assurance des entreprises. La branche RI présente la part de cotisation la plus élevée (32%) parmi tout le marché DAB. Afin d'apprécier le positionnement d'AXA sur la branche RI, une étude de comparaison a été réalisée. Les résultats ont montré qu'en 2019, AXA constitue une part représentative sur le marché RI en termes de cotisation (17%).

1.4 Les offres de la branche RI d'AXA France

La branche RI est composée de trois produits avec deux segmentations différentes, le bas de segment et le haut de segment.
La frontière entre les différents produits est définie tout d'abord par le type d'activité de l'entreprise (3 classements possibles : Activités commerciales et artisanales, Commerces de gros et petites activités industrielles, Activités industrielles Grandes activités commerciales) puis par des critères de surface, chiffre d'affaires et de contenu incendie.

Ci-dessous un schéma et un tableau qui montrent la distinction des trois produits avec les différents critères :

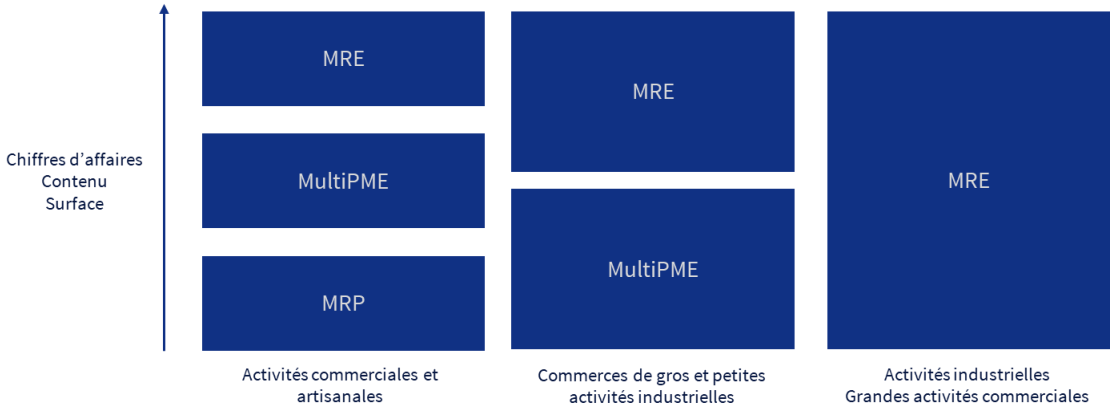


Figure 1.4.1 - Distinction des trois produits

Segmentation	Bas de segment		Haut de segment
Produit	MRP	MPME	MRE
Critères de passage	Critère 1 : - Activité commerciales et artisanales - Une superficie inférieure à 3000 m² - Un chiffre d'affaires annuel inférieur à 5 M€ - Un contenu incendie inférieur à 1 M €	Si l'un des critères 1 n'est pas respecté et que le critère 2 ci-dessous est respecté : - Commerces de gros et petites activités industrielles - Une superficie inférieure à 7000 m² - Un chiffre d'affaires annuel inférieur à 10 M€ - Un contenu incendie inférieur à 3 M€	Si les deux critères ne sont pas respectés et que le critère 3 ci-dessous est respecté : - Activités industrielles, grandes activités commerciales - Une superficie inférieure à 8000 m² - Un chiffre d'affaires annuel inférieur à 20 M€ - Un contenu incendie inférieur à 20 M€

Tableau 1.4.1 - Distinction des trois produits avec les différents critères

Dans les sections suivantes, des précisions sur les produits seront abordées.

1.4.1 Bas de segment

Le bas de segment de la branche RI, comprend les produits Multirisques Professionnelles et Multirisques Petites et Moyennes Entreprises.

Multirisques Professionnelles (MRP), commercialisées depuis 1996, avec une séparation entre Particuliers-Professionnels (PP) et Entreprises (EN) :

- **Multirisques Professionnels Particuliers et Professionnels (MRP PP)**
- **Multirisques Professionnels Entreprises (MRP EN)**

Multirisques Petites et Moyennes Entreprises (MPME), commercialisées depuis 2012

Le portefeuille 2010 à 2018 est composé majoritairement du produit MRP PP, soit 89% des contrats et constitue 10% des primes acquises.

Quelques exemples d'activité pour ces deux produits : salon de coiffure, auto-école, vente de produits de beauté, station de lavage automobile en libre-service, tailleur, …
Pour distinguer les différentes activités pour ces trois produits, il existe un code activité sur sept caractères dont les quatre premiers caractères permettent de déterminer le groupe d'activité. Les trois autres caractères permettent à AXA de faire une différenciation des sous-activités à une maille plus fine.

1.4.2 Haut de segment

Le haut de segment de la branche RI se caractérise par le produit Multirisques Entreprises.

Multirisques Entreprises (MRE), commercialisées depuis 2004, avec deux approches :

- **Multirisques Entreprises simplifiées (MRES)**, où le SMP est inférieur à 20 M€ de capitaux assurés et correspond à un monosite
- **Multirisques Entreprises complexes (MREC)**, où le SMP est supérieur ou égal à 20 M€ de capitaux assurés

La MREC constitue une part de prime acquise très importante (88%) justifiée par son risque conséquent, mais ne représente seulement 5% des contrats.

Quelques exemples d'activité pour ce produit : hôtel 5 étoiles, industrie pharmaceutique, travail des métaux, piscine, gymnase, patinoire, laboratoire de recherches, …

Pour déterminer le type d'activité en MRE, AXA utilise le TRE (Traité des Risques d'Entreprises) qui contient 177 rubriques qui se regroupent elles-mêmes en 10 fascicules. Ainsi, chaque type d'entreprise est associé à un code TRE, dont la structure est composée de deux lettres suivies de trois chiffres. Le premier des trois chiffres correspond au numéro de fascicules. Les deux derniers chiffres permettent de distinguer les différentes activités au sein d'un même fascicule.

La nomenclature des dix fascicules est la suivante :

- Fascicule 0 : Extraction et préparation de minerais et minéraux divers, de combustibles minéraux solides – Métallurgie,
- Fascicule 1 : Production de matériaux de construction – Industries des céramiques – Industries du verre,
- Fascicule 2 : Travail des métaux – Industries électriques et électroniques – Construction automobile, aéronautique et navale – Carrosserie et réparation de véhicule en tout genre – Garages et stations – service,
- Fascicule 3 : Industries chimiques et para-chimiques – Transformation de matières plastiques et de caoutchouc,
- Fascicule 4 : Industries textiles – Bonneterie – Confection de vêtements et autres articles textiles,
- Fascicule 5 : Industries du papier et du carton – Imprimeries – Industries du cuir et du délainage,
- Fascicule 6 : Industries du bois,
- Fascicule 7 : Industries agro-alimentaires,
- Fascicule 8 : Traitement des déchets urbains et industriels – Production et distribution d'énergie,
- Fascicule 9 : Autres risques d'entreprises.

Ci-dessous un schéma descriptif détaillé du périmètre de la branche RI avec des exemples d'activités par segment (portefeuille 2010-2018).

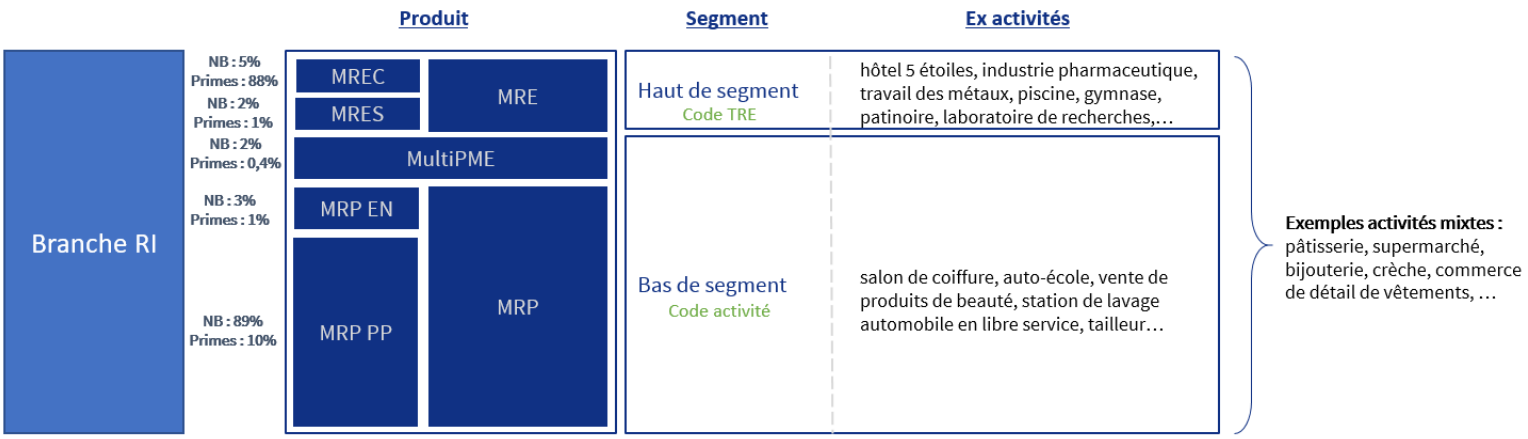


Figure 1.4.2 - Périmètre de la branche RI avec des exemples d'activités par segment

Le schéma ci-dessus met en évidence la différence des tailles de risque, la MRP PP est caractérisée par beaucoup de contrats et peu de primes tandis que le haut de segment avec peu de contrats, mais des primes importantes.

Il existe également des activités mixtes, ouvertes aux marchés professionnels et entreprises.
Exemple : pâtisserie, supermarché, bijouterie, crèche, commerce de détail de vêtements, commerce de gros de vêtements, d'alimentation, …
Ces activités mixtes peuvent relever à la fois de la MRP, de la MPME ou de la MRE.

1.4.3 Tarification des différents produits

Historiquement, nous proposons un tarif DAB par produit, ce qui nous amène à trois tarifs pour l'assurance DAB.

Les trois produits sont tarifés séparément et de manière différente :

- **Méthode analytique**
  Les produits MRP et MPME sont tarifés par des méthodes analytiques avec une structure par garantie et par niveau de couverture, ensuite enrichies par d'autres coefficients selon les clauses du contrat.
- **Théorie de la crédibilité hiérarchique de Jewell**
  Le tarif du produit MRE s'inspire du tarif des risques d'entreprises de la FFA proposé en 2009 qui applique le modèle de crédibilité hiérarchique de Jewell. Dans ce modèle, la seule variable discriminante retenue est l'activité de l'assuré. Il nécessite l'évaluation de paramètres dits "structuraux" à partir de l'observation d'un historique de sinistres.

La FFA propose des taux de prime pure par TRE. AXA s'est inspiré de ces taux pour créer sa prime commerciale. Ci-dessous la structure tarifaire en 5 étapes :

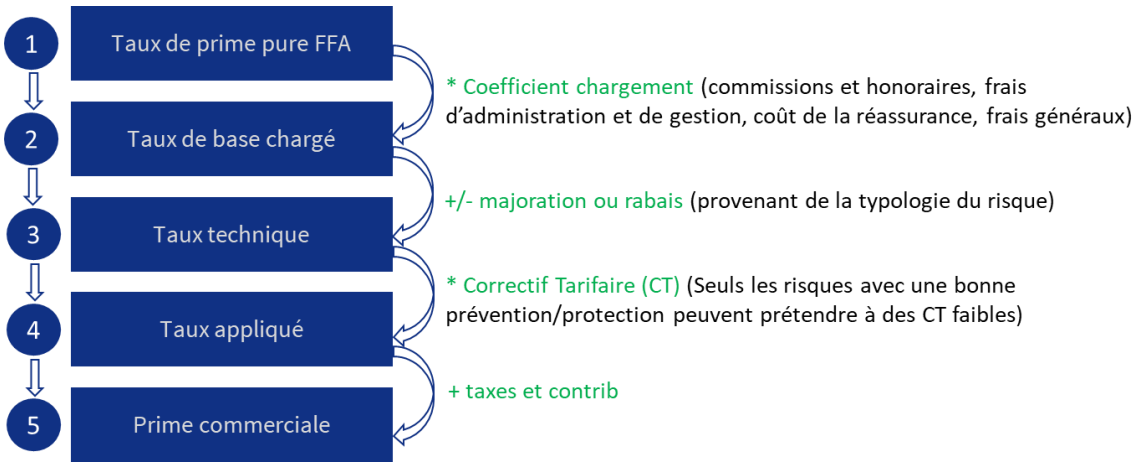


Figure 1.4.3 - Passage de la prime pure à la prime commerciale (ancien tarif du produit MRE)

1.4.4 Gestion et sélection des risques

Avant la tarification, des étapes primordiales de sélection, d'analyse et d'évaluation des risques sont réalisées au niveau de la souscription.

Comme dans tout contrat, les deux parties sont tenues à des obligations. L'assureur doit honorer son contrat en remboursant l'assuré en cas de sinistre. Réciproquement, l'assuré doit respecter les termes de son contrat. A la souscription, il est tenu de répondre de manière exacte aux questions posées par l'assureur ou le distributeur, et doit régler sa cotisation. A chaque échéance du contrat, l'assuré signale tout évènement pouvant aggraver les risques pris en charge (par exemple en déclarant une augmentation du chiffre d'affaires ou du contenu assuré). Par ailleurs, tout sinistre doit être déclaré dans les délais prévus par le contrat d'assurance.

Une bonne appréciation de la santé financière des assurés est un des axes de réduction de la sinistralité. En étant très vigilant sur les entreprises en grande difficulté financière, il est possible d'éviter un certain nombre de sinistres graves :

- Sinistres de malveillance dus à des salariés d'entreprises en réduction d'effectifs ;
- Sinistres volontaires de la part de propriétaires d'entreprises en grande difficulté financière ;
- Sinistres causés ou aggravés par une mauvaise maintenance des outils de production ou par un mauvais entretien des systèmes de protection dans les entreprises dont les moyens financiers ainsi que la motivation sont insuffisants.

La connaissance de la sinistralité est indispensable avant la souscription. Compte tenu des faibles fréquences de sinistres enregistrées en Risques Industriels, cette sinistralité est demandée sur une période d'au moins 3 ans. La connaissance de la sinistralité passée nous renseigne sur la pathologie sinistre (survenance de sinistres graves ou de sinistres répétitifs, type de sinistres) et sur les mesures mises en œuvre par l'exploitant. Elle peut également révéler des conditions d'exploitation défavorables, une absence de maintenance et/ou de révision, un manque de protection contre le vol, une situation économique tendue, etc.
Dans le cas d'une résiliation par le précédent assureur, il est important d'analyser quel en a été le motif.

La connaissance des antécédents pourra conduire à prescrire des mesures de prévention ou de protection préalable à la souscription, à modifier les franchises et les conditions tarifaires, voire à ne pas accepter le risque.

Déterminer la cotisation (prime) et le montant de franchise est un enjeu également important pour les souscripteurs.

La cotisation est forfaitaire ou révisable avec mise à jour annuelle :

- Les contrats avec la cotisation forfaitaire ne sont pas assujettis à une évolution de prime liée à un changement de chiffre d'affaires (CA) ou autre. Dans ce cas, le CA n'est pas révisé, mais est tout de même indexé selon le rapport entre les indices de base à la souscription et de quittancement. Cette gestion forfaitaire présente des avantages pour l'assuré et l'assureur. L'assuré a une prime assez stable car il n'est pas impacté par les changements liés à l'activité de l'entreprise. Pour l'assureur, cela permet de diminuer les frais liés à la gestion de la révision des contrats.
- Les contrats avec la cotisation révisable ont un CA qui est révisé chaque année, c'est-à-dire que l'assuré doit déclarer annuellement son CA.

La franchise correspond au montant qui reste à la charge de l'assuré en cas de sinistre ou en d'autres termes, la partie des coûts des dommages non pris en charge par AXA. Elle est indexée et évolue à chaque échéance.

La suite du premier chapitre est consacrée à la description de l'objectif de l'étude pour mieux appréhender les enjeux de la refonte du tarif rapide.

1.5 Objectifs de l'étude

1.5.1 Evolution des offres risques industriels d'AXA France

Les produits Multirisques couvrent actuellement beaucoup d'activités (~350) et les parcours de souscription sont peu lisibles et flexibles. La souscription entièrement déléguée avec des garanties modulables est limitée à 30 activités. Ainsi, le processus de souscription actuel ne permet pas de répondre efficacement et rapidement aux exigences et besoins d'une offre transparente, modulable et au meilleur prix sur le marché cible des Entreprises dont le chiffre d'affaires (CA) est inférieur à 20 M€.

Les ambitions à travers la nouvelle offre RI sont multiples :

- Créer un parcours de souscription unique, simple dynamique qui faciliterait la vente et répondrait aux attentes des clients à l'aide d'un questionnaire risque optimisé. Une simplification avec une unique structure tarifaire mieux segmentée mais autorisant une flexibilité du tarif par marché/activité.
- Permettre la convergence Professionnels Particuliers et Entreprise (PP/EN, respectivement 93% et 7% en nombre de contrats), avec un passage fluide du bas de segment au haut de segment sans effet de seuil.
- Donner la possibilité de choisir le niveau de couverture : léger, partiel ou sur mesure.
- Pouvoir maintenir la part de marché d'AXA sur le réseau Agents face à l'agressivité de la concurrence notamment des banques-assureurs.
- Développer l'activité des Agents sur les entreprises supérieures à 50 salariés, ainsi que du Courtage sur les entreprises inférieures à 50 salariés.
- Réduire les coûts de développement en mutualisant la refonte de 3 produits (MRP, MPME, MRE).

Cette nouvelle offre est ouverte aux artisans, professions libérales, commerçants (détaillants ou grossistes) et entreprises situées en France, en principautés de Monaco ou d'Andorre et dont le chiffre d'affaires est inférieur à 20 M€.
Cette offre divise le marché PP et l'EN, avec une **offre Pro** (CA inférieur 5 M€) et une **offre Entreprise** (CA compris entre 5 M€ et 20 M€) en proposant 3 niveaux de couvertures : léger, partiel ou sur-mesure.

- Léger, « Essentielle » : Offre qui représente une couverture intégrant les garanties essentielles (garanties de base/du socle : incendie, évènements climatique et dommages électriques) de l'activité et un tarif ajusté
- Partiel, « Optimale » : Offre qui porte des éléments différenciant pouvant être adaptés. Ce second niveau constitue le niveau équilibré de couverture avec un tarif toujours ajusté
- « Sur-mesure / A la carte » : offre avec la plus grande souplesse au niveau des garanties et du choix des montants maximum de couverture.

Les différentes offres répondent à une logique de sélection du risque, des garanties choisies permettant d'afficher un tarif ajusté et un risque maîtrisé.

1.5.2 La garantie Dégâts Des Eaux

Le Dégât Des Eaux (DDE) est un risque réel pour l'entreprise. Toute entreprise peut être confrontée à ce risque, dont les origines peuvent être multiples. La garantie DDE en RI, couvre les dommages matériels causés directement par l'eau et consécutifs à une fuite, engorgement de conduits ou de canalisations, infiltrations accidentelles, gel des tuyaux, etc. A noter que les inondations ne sont pas comprises dans cette garantie, elles font partie de la garantie évènement climatique. Sur la période observée (2010-2018), 93% des contrats ont la garantie DDE, cette dernière ne fait pas partie des garanties de base, obligatoires de couverture. Les statistiques ont montré que cette garantie relève une fréquence de sinistre de 4%, correspondant à la garantie la plus sinistrée en termes de nombre de sinistres (1ère position) et montant des charges (2ème position) parmi toutes les garanties proposées.

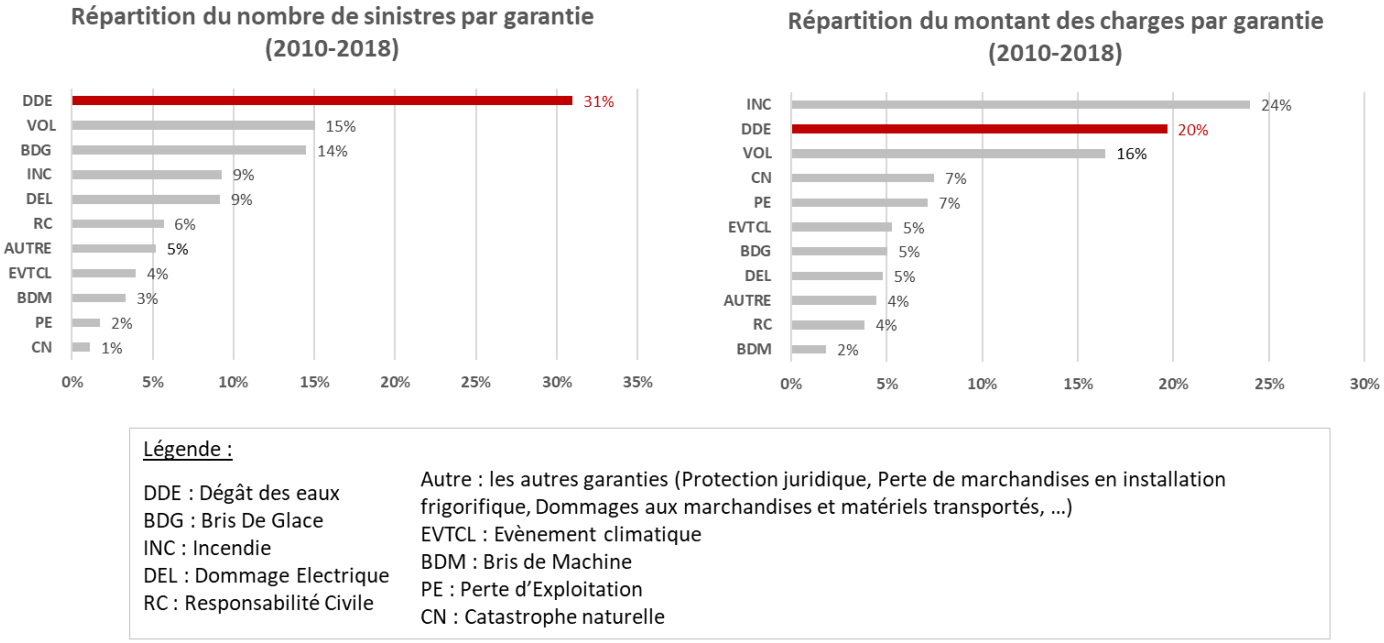


Figure 1.5.1 - Répartition des sinistres en nombre et montant par garantie

Des statistiques sur la fréquence, le coût moyen et la prime pure en base 100 par marché et par produit sur la période observée 2010 à 2018 sont visibles ci-dessous.

Un rappel des formules sur les trois variables d'analyses :

- $\text{Fréquence (Freq) DDE observée} = \frac{\text{Nombre de sinistres DDE}}{\text{Exposition DDE}}$
- $\text{Coût moyen (CM) DDE observé} = \frac{\text{Charges DDE}}{\text{Nombre de sinistres DDE}}$
- $\text{Prime pure (PP) DDE observée} = \frac{\text{Charges DDE}}{\text{Exposition DDE}} = \text{Fréquence} * \text{Coût moyen}$

Les graphiques ci-dessous sont mis en base 100 pour des raisons de confidentialité.

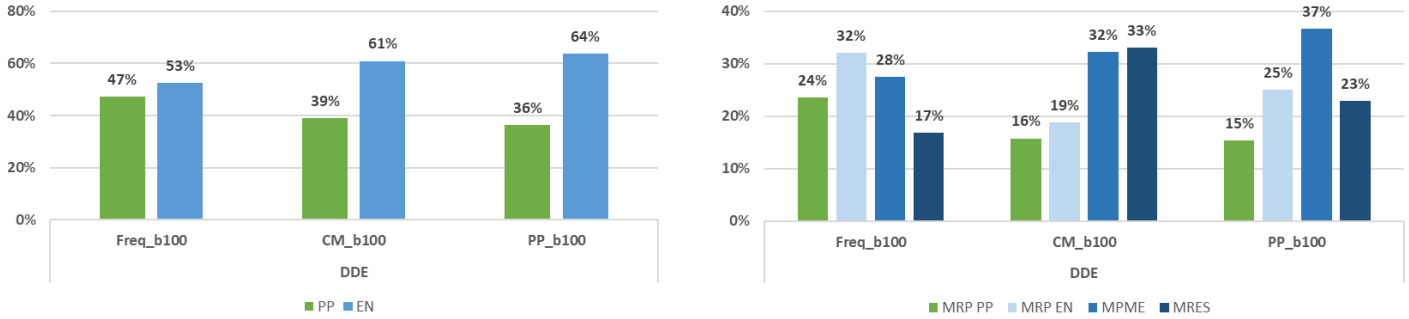


Figure 1.5.2 - Indicateurs de sinistralité par marché et par produit

Par marché : Le marché Particuliers-Professionnels (PP) a une taille de risque plus petite que le marché Entreprise (EN) observation confirmée par le graphique à gauche, il ne représente que 39% en CM et 36% en PP sur toute la base.

Par produit : Le produit MRES - haut de segment RI, présente la plus faible fréquence et la deuxième prime pure la plus faible. Ceci peut être justifié par les différents types de protection et prévention disponibles dans les très grandes entreprises donc ils sont moins impactés par les sinistres DDE. Cependant, lorsqu'un sinistre a lieu, le coût moyen pour ce produit reste tout de même le plus élevé (représentant 33% du CM).

Il est également intéressant d'étudier le ratio de sinistralité (rapport entre la somme des charges et la prime acquise DDE) par année de vision. Ci-dessous, une dégradation de la rentabilité est observée depuis 2015. Ce constat remet en cause l'ancien tarif et justifie la refonte du tarif réalisé en 2019 pour le nouveau produit RI.

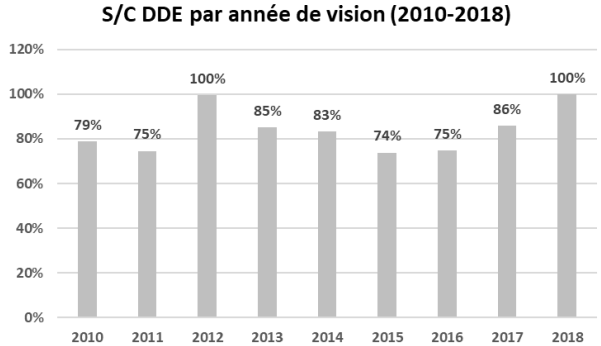


Figure 1.5.3 - Evolution du S/C par année de vision

1.5.3 Simplification et amélioration du tarif rapide

- **Tarif technique actuel**

A la différence de l'ancienne offre, cette nouvelle offre nous permettra d'obtenir deux tarifs :

- **Un tarif technique** unique au lieu de trois tarifs, où la prime pure de chaque garantie est modélisée selon deux approches :
  - Actuarielle : modèle de crédibilité pour l'incendie et modèle de fréquence * coût moyen pour les autres garanties majeures.
  - A dire d'expert : pour les plus « petites garanties » où l'on dispose de peu de données ou portant plus de marge.

  Ensuite, chaque activité dispose de coefficients multiplicateurs pour ajuster le risque sur chaque garantie.
- **Un tarif commercial** adapté à la stratégie du marché (Professionnel ou Entreprise), comprenant les frais généraux : frais de gestion des sinistres, loyers, salaires des employés, impôts et taxes, et un coefficient permettant de respecter les stratégies tarifaires de chaque marché pour chaque activité (coefficient activité * marché).

Ci-dessous la liste des variables tarifaires du tarif technique actuel (prime HT) :

- Activité
- Surface
- Qualité de l'occupant : Locataire Occupant (LO), Locataire Exonéré de RC locative (LE), Locataire qui souscrit pour le Compte de son propriétaire (LC) il est alors chargé d'assurer les bâtiments comme le propriétaire, et Propriétaire Occupant (PO)
- Contenu par garantie : incendie, dommages électriques, vol
- Chiffre d'affaires
- Période d'indemnisation
- Antécédents de sinistres par garantie
- Zoniers par garantie

Tarif technique DDE actuel

Actuellement, la garantie DDE est modélisée par un modèle de fréquence x CM.
La structure tarifaire est la suivante :

$$\text{Constante} * \text{Activité} * \text{Contenu INC} * \text{Surface} * \text{Antécédents DDE} * \text{zonier DDE} * \mathbb{1}_{\text{statut zonier}}$$

Deux types de zoniers interviennent dans le tarif : zonier Voronoi et le zonier INSEE. Ils sont calibrés en mettant en off set les coefficients des autres variables tarifaires.

- Le **zonier Voronoi** est utilisé lorsque l'adresse est renseignée complètement, et que le géocodage d'AXA a réussi à lui associer des coordonnées longitude et latitude. Les cas Voronoi sont de l'ordre de 40% dans la base.
- Le **zonier INSEE** est appliqué le cas échéant. Il s'agit d'un géocodage à la maille commune, grâce au code INSEE. La France métropolitaine est composée d'un peu plus de 36 000 codes INSEE.

Le zonier est plus amplement expliqué dans la section 2.2. Traitement des variables explicatives retenues.

- **Tarif quick quote / tarif rapide actuel**

Par définition, un tarif quick quote est un tarif rapide, donc il possède une contrainte sur le nombre de questions à poser au client. Par conséquent, certaines questions ne sont pas demandées et sont négligées.
Dans le cadre du tarif rapide du nouveau produit RI d'AXA France, en outre de la contrainte de questions, une autre contrainte de structure tarifaire du tarif technique est imposée. En raison de l'existence de cette deuxième contrainte, le tarif rapide actuel doit prendre en compte l'intégralité des variables du tarif technique. Ainsi, la modélisation actuelle du tarif rapide repose sur une liste de six questions posées aux clients, définies par connaissance métier, et se référant à une modalité de base pour toutes les variables du tarif technique non posées aux clients. Cette modalité de référence dans ce cadre permet de pallier le problème des questions non posées aux clients. Par exemple, les antécédents de sinistre ne sont pas posés aux clients pour calculer le tarif rapide, ainsi il est convenu d'utiliser une modalité de référence correspondant à la modalité la plus représentée : « sans antécédents » pour tout le monde.

Ci-dessous la liste des variables posées aux clients :
- Activité
- Adresse pour géocoder les zoniers
- Surface
- Qualité
- Contenu incendie
- Chiffre d'affaires

Pour le produit d'assurance RI, sept garanties du tarif font intervenir des variables non posées au client dans le cadre du tarif rapide, faisant ainsi appel à des modalités de base pour calculer le tarif de ces garanties.

Les sept garanties avec les variables non posées au client sont :

1. Dégâts des eaux DDE (Antécédents DDE)
2. Dommages électriques DEL (Antécédents DEL, Contenu DEL)
3. Bris des glaces, BDG (Antécédents BDG, Longueur de vitrine, murs rideaux)
4. Bris de machine BDM (Antécédents BDM)
5. Responsabilité civile RC (Antécédents RC)
6. Vol (Antécédents vol, Contenu vol)
7. Perte d'exploitation PE (période d'indemnisation)

Sur ces sept garanties, six d'entre-elles utilisent une modalité de base pour les antécédents. Pour étudier l'impact de cette modalité de base, nous nous sommes intéressés à la garantie Dégâts Des Eaux (DDE) car elle présente la sinistralité la plus élevée, en termes de fréquence, parmi toutes les garanties faisant appel à des modalités de base. Les autres garanties suivront le même raisonnement que l'étude de la garantie DDE.

- **Nouveau tarif rapide DDE**

La Direction Technique souhaite challenger le tarif rapide actuel mis en place, en élaborant un nouveau tarif rapide sans se contraindre à la structure tarifaire technique, et en vérifiant la pertinence des questions posées actuellement. Il s'agit de savoir si d'une part les questions posées aux clients sont indispensables (suffisantes ou insuffisantes) pour calculer rapidement le tarif, et d'autre part d'évaluer,

et de quantifier la perte engendrée par les questions non posées au client. Est-ce une plus-value de rajouter des questions intervenant dans le tarif technique ? Son abstraction permet-elle d'obtenir un tarif raisonnable ?

Pour ce faire, deux tarifs ont été créés :

- **Nouveau tarif rapide**, basé sur les six questions posées au client et sans structure tarifaire imposée
- **Nouveau tarif technique avec la structure du nouveau tarif rapide et la variable antécédent,** basé sur les six questions posées au client et ajout de la variable antécédent intervenant dans le tarif technique, mais non demandées au client. Ce tarif permet de quantifier l'impact des antécédents dans le modèle.

Pour valider les deux tarifs créés, une analyse des performances des modèles et des prédictions sur la base d'apprentissage et la base de validation est réalisée. Une fois les modèles validés, pour connaître l'impact des modalités de référence sur la performance des modèles, nous comparons le tarif rapide actuel avec le nouveau tarif rapide. La deuxième étude serait de quantifier la perte liée aux questions non posées, mais intervenant dans le tarif technique. Pour cela, nous analysons le nouveau tarif rapide avec le nouveau tarif technique (avec la structure du nouveau tarif rapide et la variable antécédent). Si l'écart entre les deux tarifs est faible alors nous jugeons que les questions posées au client sont suffisantes pour modéliser la garantie DDE.

Le périmètre du mémoire est le bas de segment (MRP, MPME) et le haut de segment (MRES uniquement). La MREC est exclue de la modélisation car elle ne fait pas partie du périmètre du nouveau produit RI, où AXA préfère limiter la souscription de la MREC à cause de la sensibilité aux sinistres atypiques. La garantie modélisée est le Dégât des Eaux (DDE). La période d'observation est de 2010 à 2018. Les variables d'analyse sont les variables posées au client à savoir : l'activité, les zoniers, la surface, la qualité, le contenu incendie, le chiffre d'affaires.

CHAPITRE 2

2 Création de la base de données

Ce deuxième chapitre est destiné à détailler les étapes permettant la création de la base de données. Il sera divisé en trois sections : l'agrégation des différentes bases, les traitements des variables explicatives retenues et des variables à expliquer.

2.1 Agrégation des différentes bases

Le produit d'assurance RI se décompose en deux applicatifs de souscription. AXAPAC qui est l'outil de souscription pour le bas de segment (MRP, MPME), et NAE (Nouvelle Application Entreprise) pour le haut de segment (MRE). Les données issues des deux outils sont stockées dans deux bases risques différentes. Ainsi cela peut créer des asymétries de complétude des données dans la base agrégée.

Etant donné que la nouvelle offre réunit les trois produits en un tarif technique unique, la base de données sera une combinaison de différentes sources. A noter que cette fusion correspond à un premier biais dans la modélisation, car la typologie des contrats diffère selon les produits.

L'accès, la création, la gestion et le traitement des bases sont simples et efficaces. Toutes les bases décrites ci-dessous sont annuelles, plusieurs fusions sont indispensables pour réunir les informations disponibles d'un contrat de chaque année de vision et de chaque base dans une base de données unique.

2.1.1 Bases contrats

La toute première étape de la construction de la base de données est de sélectionner les contrats dans le périmètre d'étude, à savoir les produits RI étudiés (MRP PP, MRP EN, MPME et MRES) et l'historique de portefeuille (2010-2018). Le périmètre réduit à l'année 2018 est expliqué par le choix d'un vieillissement des charges sur deux ans (plus de détails dans la section 2.3.1.2 Vieillissement). La base des contrats est obtenue par fusion des bases 2010 à 2018 sur les produits RI étudiés. Ensuite, des traitements spécifiques sur la période de vision sont réalisés, à savoir conserver uniquement les contrats qui ont été en cours ou résiliés sur cette période et supprimer les affaires nouvelles sans effets.

Les variables brutes renseignées dans les bases sont le numéro de contrat, la date d'affaire nouvelle (date de création du contrat), date de résiliation, l'année en cours (année pendant laquelle le contrat a été présent), le type de produit (MRP PP, MRP EN, MPME, MRES), le numéro de client (un client peut avoir plusieurs contrats), le marché (PP, EN), la CoPHT (Cotisation Potentielle Hors Taxe annuelle), les primes émises DDE, le code INSEE de la ville de risque (maille plus fine que le code postal), les indices d'indexation etc. Certaines variables utiles pour la modélisation ne sont pas renseignées

directement dans la base, ainsi des manipulations de création de variables à partir des variables brutes ont été effectuées (par exemple pour l'exposition et pour la prime acquise DDE)

Ci-dessous quelques traitements nécessaires pour la modélisation et les analyses.

- Traitement pour obtenir la variable exposition :

$$\text{Exposition} = \begin{cases} \dfrac{\text{date de fin} - \text{date de début} + 1}{366} & \text{si année bissextile} \\ \dfrac{\text{date de fin} - \text{date de début} + 1}{365} & \text{sinon} \end{cases}$$

Avec

- date de début = max(date d'affaire nouvelle ; 01/01/année de vision)
- date de fin = min(date de résiliation – 1 ; 31/12/ année de vision)

Le but étant de modéliser la prime pure de la garantie DDE, cette variable permettra d'obtenir l'exposition de la garantie étudiée :

$$\text{Exposition DDE} = \text{Exposition} * \mathbb{1}_{\text{garantie DDE présente}}$$

- Traitement pour obtenir la variable prime acquise DDE, utilisée dans les graphiques de S/C (Section 1.5.2 La garantie Dégâts Des Eaux) :

$$\text{Prime acquise DDE} = \text{Prime émise DDE} * \text{Exposition DDE}$$

2.1.2 Bases risques

Les données liées au risque assuré sont stockées dans les bases risques. Ces données sont majoritairement déclarées à la souscription. Les bases risques sont enrichies par les applicatifs de souscription. Pour cette étude, quatre bases risques ont été sollicitées, deux pour les produits hors MRES et deux pour les produits MRES :

- **Base risque carte :** contient les données des produits MRP PP, MRP EN et MPME qui ont la formule « carte ». Cette dernière est un niveau de couverture libre, où l'assuré peut choisir les garanties à couvrir.
- **Base risque spéciale :** contient les données des produits MRP PP, MRP EN et MPME qui ont la formule « spéciale ». Cette dernière est un niveau de couverture où les garanties sont déterminées sans possibilité d'adaptation.
- **Base risque NAE :** contient les données du produit MRES. Cette base recense la plupart des informations risques de l'assuré.
- **Base risque données libres (DL) :** contient les données du produit MRES. La dénomination de « données libres » est liée au fait qu'il n'y a ni formatage ni obligation d'entrer les données déclarées à la souscription. Cette dernière complète les données risques de la base risque NAE, par exemple le Sinistre Maximum Possible et les capitaux incendies assurés proviennent de cette base.

Quelques exemples de variables récupérées des bases risques : Numéro de contrat, année en cours (année de vision), adresse de l'entreprise, montant de CA déclaré lors de la souscription, les différents capitaux assurés des garanties, surface de l'entreprise, qualité de l'occupant (locataire ou propriétaire),

code activité de l'entreprise, SMP, etc. Dans la grande majorité des cas, les bases risques contiennent les mêmes informations mais avec des noms de variables différents. Un renommage des variables s'impose pour harmoniser les données recueillies.

2.1.3 Bases révisables

Les bases révisables apportent les informations sur la révision du CA. Pour les contrats où la cotisation est révisable (25% des cas), le CA annuel révisé de cette base est utilisé. Pour les contrats avec la cotisation forfaitaire, nous appliquons le CA déclaré lors de la souscription, disponible dans les bases risques, ce dernier est indexé selon le rapport entre les indices de base à la souscription et de quittancement. Cependant, il faut noter qu'il est possible que le CA ne soit pas ou mal renseigné, tout comme les autres variables (Cf. section 2.2.1 Détection, imputation des valeurs manquantes).

2.1.4 Bases sinistres

Cette base recense la sinistralité par contrat et par année de vision. Afin de prendre en compte au moins trois ans d'antécédents de sinistre pour chaque année de vision, la sinistralité sera étudiée sur un historique de douze ans (2007-2018). La sinistralité de 2007 à 2009 correspond donc aux antécédents de sinistre de l'année 2010 de la modélisation. La branche RI étant un risque d'intensité, elle est caractérisée par peu de sinistres et assimilée à une charge très importante. De ce fait, la sinistralité est demandée sur une période d'au moins 3 ans pour avoir assez de recul. La définition et le principe des antécédents sont détaillés dans la section suivante.

Chaque sinistre est désigné par un numéro de sinistre, associé à un numéro de contrat et une date de survenance. Un contrat peut avoir un sinistre qui impacte une ou plusieurs garanties. Dans le cas où plusieurs garanties sont touchées par un sinistre, le numéro de sinistre, le numéro de contrat et la date de survenance du sinistre sont répétés autant de fois qu'il y a de garanties impactées (survenance d'un sinistre incendie, DDE ou catastrophe naturelle engendrant une baisse d'activité ou un arrêt total de la production de l'entreprise et déclenchant ainsi la garantie Perte d'Exploitation). Les variables garantie, charge par garantie[5] et charge totale des garanties[1] permettent de caractériser le sinistre. Ci-dessous la structure de la base sinistre initiale :

Base sinistre						
Numéro de sinistre	Numéro de contrat	Date de survenance	Date d'observation	Garantie	Charge garantie	Charge totale
1	A	01/01/2010	01/01/2012	INC	1 000	1 000
2	A	31/07/2010	31/07/2012	DDE	200	200
3	A	01/12/2010	01/12/2012	DDE	100	100
4	B	01/01/2018	01/01/2020	DDE	15 000	20 000
4	B	01/01/2018	01/01/2020	PE	5 000	20 000

Tableau 2.1.1 - Structure de la base sinistre initiale

[5] Charge y compris franchise

La date d'observation correspond à la date de vision de la charge. Dans le cadre de la modélisation de la garantie DDE, il est convenu un vieillissement 2 ans (Cf. section 2.3.1.2 Vieillissement). Pour étudier la sinistralité d'un contrat par garantie et par année de survenance, il faut sommer les sinistres par garantie survenus la même année et indiquer le nombre de sinistres et la charge totale correspondants. Dans l'exemple du tableau ci-dessus, cela reviendrait à avoir pour le numéro de contrat A, un sinistre incendie 2010 à 1 000€, deux sinistres DDE 2010 pour une charge DDE totale à 300€ et au global, trois sinistres avec une charge globale de 1 300€. Ainsi, chaque ligne d'observation contient : le numéro de contrat, l'année de survenance, le nombre de sinistres et charge globale, nombre de sinistre et charge par garantie. Ci-dessous la structure de la base sinistre retraitée :

Base sinistre retraitée							
Numéro de contrat	Année de survenance	Nb_sin_tot	Charge_tot	Nb_sin_DDE	Charge_DDE	Nb_sin_INC	Charge_INC
A	2010	3	1 300	2	300	1	1 000

Tableau 2.1.2 - Structure de la base sinistre retraitée

D'autres traitements sur la charge ont été effectués, tels que l'annulation des charges négatives, le vieillissement et l'indexation des charges, l'obtention de la charge ultime par méthode de Chain Ladder, la distinction des charges attritionnelles, graves, atypiques (Cf. section 2.3.1 Charge des sinistres).

2.1.5 Base agrégée

Plusieurs suivis RI sont effectués mensuellement pour analyser la production et la sinistralité de la branche. Avant de fusionner les bases, une réconciliation des bases créées est réalisée avec les différents suivis disponibles. Cette réconciliation permet de vérifier et de valider les différentes extractions de données opérées.

La structure de l'agrégation des différentes bases peut se synthétiser par le schéma suivant :

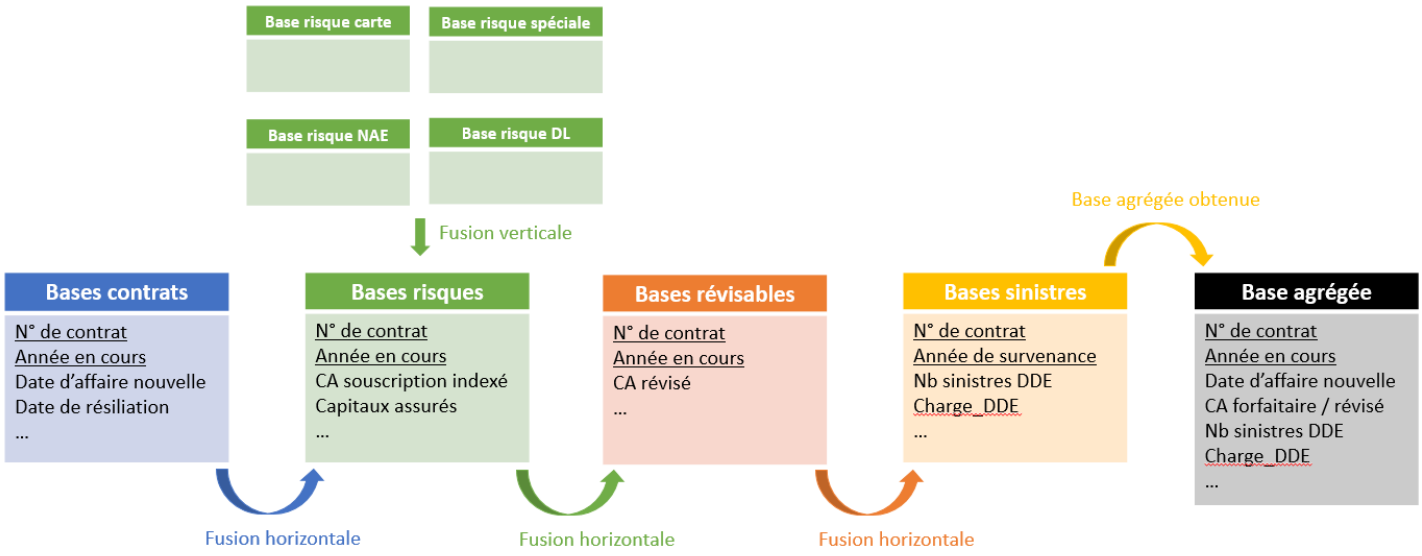


Figure 2.1.1 - Schéma de l'agrégation des différentes bases de données

Une fois la base agrégée obtenue, une étape de nettoyage des données est primordiale. Sur cette base, il est possible de rajouter des filtres pour bien sélectionner le périmètre d'étude. Par exemple : les entreprises présentant un CA supérieurs à 20 M€ et certaines activités d'entreprise (propriétaires non occupants, exposition, foires expositions, fonderie de moulage, …) ne peuvent pas être souscrites dans le nouveau produit RI.

Un nettoyage de données plus détaillé sur les variables explicatives retenues est disponible dans la section suivante.

2.2 Traitement des variables explicatives retenues

Les variables explicatives sont tout d'abord les variables tarifaires du tarif technique DDE actuel, à savoir l'activité, les zoniers, la surface, le contenu incendie et les antécédents.

2.2.1 Détection, imputation des valeurs manquantes

Activité

L'activité est distinguée par le code activité (bas de segment) et le code TRE (haut de segment). Une variable activité a été créée pour combiner en une variable l'activité selon le segment.

$$\text{Activité} = \begin{cases} \text{Code activité} & \text{si bas de segment} \\ \text{Code TRE} & \text{si haut de segment} \end{cases}$$

Zoniers DDE

Les deux zoniers DDE sont le zonier Voronoi et le zonier INSEE. Nous rappelons que le zonier Voronoi est utilisé lorsque le statut du contrat est considéré comme Voronoi c'est-à-dire que son adresse est bien renseignée et détectée par le géocodage d'AXA ; cas échéant, le zonier INSEE est appliqué.

Le zonier Voronoi est divisé en 20 risques croissants : V01, V02, … , V20. Le zonier INSEE est décomposé en 14 risques croissants : I01, I02, … , I14. Où les modalités V01 et I01 correspondent aux zones les moins risqués.

Une variable qui fusionne ces deux zoniers a été créée pour la modélisation de la prime pure afin de challenger la méthode actuelle avec deux zoniers calibrés séparément. Ce croisement de zonier consiste à une méthode de complétion des zoniers Voronoi manquants par des zoniers INSEE. Pour cela, une vérification des zones de risques Voronoi et INSEE a été nécessaire. Ci-dessous une matrice de confusion des deux zoniers, en ligne le zonier Voronoi et en colonne le zonier INSEE.

Somme	INSEE														
Voronoi	I01	I02	I03	I04	I05	I06	I07	I08	I09	I10	I11	i12	I13	I14	Total général
V01	89%	6%	3%	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%
V02	50%	23%	16%	4%	2%	3%	1%	0%	0%	0%	0%	0%	0%	0%	100%
V03	6%	32%	32%	11%	7%	7%	3%	1%	0%	0%	0%	0%	0%	0%	100%
V04	2%	6%	40%	13%	13%	15%	6%	3%	1%	1%	0%	0%	0%	0%	100%
V05	1%	2%	24%	15%	18%	20%	10%	5%	3%	2%	0%	0%	0%	0%	100%
V06	0%	1%	9%	18%	20%	23%	14%	7%	3%	4%	0%	0%	0%	0%	100%
V07	0%	0%	5%	8%	19%	27%	17%	10%	5%	7%	1%	1%	0%	0%	100%
V08	0%	0%	2%	5%	9%	31%	21%	13%	7%	9%	1%	1%	0%	0%	100%
V09	0%	0%	1%	2%	5%	25%	24%	16%	8%	14%	2%	1%	0%	0%	100%
V10	0%	0%	1%	2%	3%	14%	29%	22%	9%	15%	2%	2%	0%	0%	100%
V11	0%	0%	0%	1%	2%	9%	24%	22%	13%	20%	4%	4%	1%	0%	100%
V12	0%	0%	0%	1%	1%	6%	14%	26%	15%	25%	7%	5%	1%	0%	100%
V13	0%	0%	0%	1%	1%	3%	8%	17%	19%	34%	7%	8%	1%	0%	100%
V14	0%	0%	0%	0%	0%	2%	4%	11%	11%	40%	12%	17%	3%	1%	100%
V15	0%	0%	0%	0%	0%	0%	3%	5%	8%	29%	12%	30%	10%	3%	100%
V16	0%	0%	0%	0%	0%	0%	1%	2%	4%	16%	14%	30%	27%	7%	100%
V17	0%	0%	0%	0%	0%	0%	0%	1%	1%	6%	7%	31%	44%	10%	100%
V18	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	3%	23%	51%	19%	100%
V19	0%	0%	0%	0%	0%	0%	0%	0%	0%	3%	2%	16%	49%	30%	100%
V20	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	1%	9%	36%	53%	100%
Total général	**20%**	**6%**	**9%**	**5%**	**5%**	**9%**	**8%**	**6%**	**4%**	**8%**	**2%**	**6%**	**7%**	**4%**	**100%**

Figure 2.2.1 - Matrice de confusion des zoniers

Si le zonier est V01 alors dans 89% des cas, il est considéré comme I01. Par ailleurs, la plupart des contrats sont concentrés sur la diagonale. Donc l'INSEE est une bonne estimation du Voronoi.

Par la suite, cette méthode sera appelée **méthode ajout simple du zonier** : complétion du zonier Voronoi manquant par le zonier INSEE. Le résultat est une unique variable composée de 34 modalités : I01, …, I14, V01, …, V20. Ci-dessous le schéma de cette méthode avec le modèle multiplicatif sur les six variables d'étude.

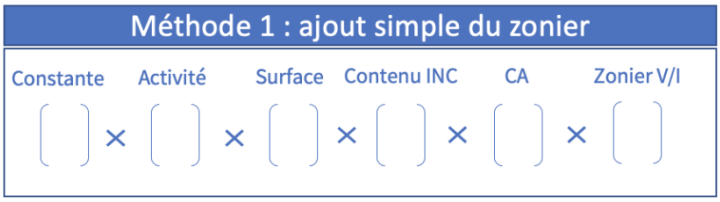


Figure 2.2.2 - Méthode 1 d'implémentation du zonier : "Ajout simple du zonier"

La méthode utilisée actuellement (tarif technique actuel et tarif rapide actuel) pour calibrer le zonier sur la prime pure sera appelée **méthode off set** : en mettant en off set les variables tarifaires hors zoniers puis en calibrant les zoniers Voronoi et INSEE séparément. Le zonier Voronoi concerne uniquement les cas Voronoi et le zonier INSEE porte sur toute la base, car l'information INSEE est disponible pour tous les contrats. Ci-dessous le schéma de cette deuxième méthode :

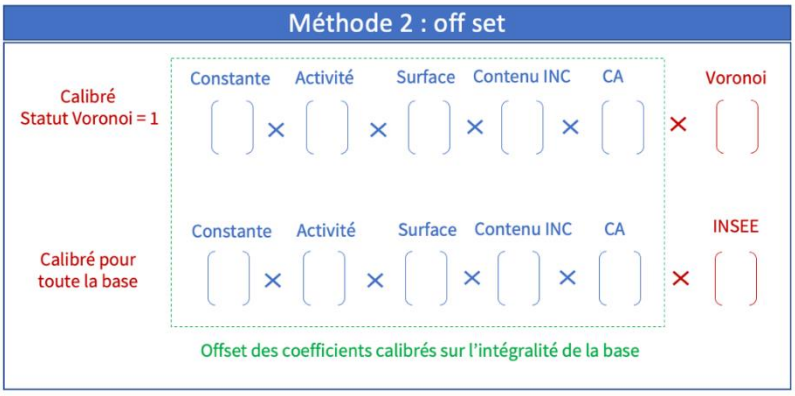


Figure 2.2.3 - Méthode 2 d'implémentation du zonier : "Off set"

Chiffre d'affaires

Dans les bases risques ou les bases révisables, il se peut que le CA ne soit pas renseigné. La valeur est, soit manquante non renseignée, soit renseignée mais saisie à zéro. Le cas des manquants non renseignés représente 2% de la base et le cas des CA saisis à zéro correspond à 31% de la base. Ces deux cas sont considérés comme manquants liés à des anomalies (soit 33% de manquants).

Dans la modélisation du tarif technique actuel, les valeurs manquantes ont été imputées selon la moyenne du CA par activité, cependant les valeurs à zéro ont été laissées à zéro. Dans le cadre du nouveau tarif quick quote, plusieurs étapes d'imputation ont été proposées pour retraiter les CA laissés à zéro.

Après avoir soulevé le problème des valeurs manquantes, plusieurs étapes d'imputation ont été effectuées. Tout d'abord, les CA supérieurs à 20M€ n'étant pas dans le périmètre d'étude, sont directement supprimés de la modélisation.

- **1ère étape d'imputation :** Imputer selon les informations antérieures

Exemple : Lorsque le contrat est présent sur les 9 ans de la modélisation (2010-2018), le CA de l'année 2018 est manquant, la première réflexion serait de prendre son chiffre d'affaires 2017 si ce dernier est renseigné, il est ensuite indexé selon les indices de l'année 2018. Le même raisonnement s'applique sur toutes les années de vision. Après cette première étape d'imputation, il reste 30% de manquants.

- **2èmpe étape d'imputation :** Enrichir avec les données externes

Si le CA est toujours manquant après le premier retraitement, nous l'enrichissons par des données externes. AXA a accès grâce à ses partenariats, à un data Warehouse qui recueille un historique des bilans des entreprises. Il est possible de récupérer le CA avec le numéro SIREN de l'entreprise. Cependant, la loi définit une durée de conservation de maximum dix ans des données personnelles, ainsi pour les années 2010 et 2011, cette méthode d'imputation n'est pas fonctionnelle. À la suite de cet enrichissement, il reste encore 29% de manquants.

- **3èmpe étape d'imputation :** Déterminer un taux de passage ‰ entre la prime perte d'exploitation (PE) et le CA par segment et activité

Cette troisième approche est réalisée sur les contrats qui ont souscrit à la garantie PE, dont le CA est renseigné et retraité par les deux premières approches. Elle consiste à déterminer un taux de passage entre la prime PE et le CA par segment et activité. A noter que cette méthode suppose que la prime PE soit bien calibrée et que le CA et l'activité soient tarifaires dans le modèle actuel pour la garantie PE. Ce qui est bien le cas.

La formule est la suivante :

$$\text{Taux de passage PE CA} = \frac{\sum \text{Prime PE}}{\sum \text{CA}_{\text{renseigné et traité}}} * 1\,000$$

Ainsi, pour obtenir le CA d'un contrat, la formule sera :

$$\text{CA}_{\text{estimé}} = \frac{\text{Prime PE}}{\text{Taux de passage PE CA}} * 1\,000$$

Quelques résultats des taux de passage PE CA ‰ sur le **bas de segment** :

Activité	Tx_passage_PE_CA_bas_segment ‰
Laverie automatique	1,30
Boulangerie - Pâtisserie industrielle	1,14
…	…
Agence de voyage	0,15
Bureau	0,12

Tableau 2.2.1 - Taux de passage entre la prime perte exploitation et le chiffre d'affaire sur le bas de segment

Application des taux de passage ‰ :

Exemple d'un assuré qui assure une laverie automatique et payant une prime PE = 60€ avec un montant de CA manquant. Le taux de passage pour la laverie automatique est de 1,30 donc son CA estimé est :

$$\mathrm{CA}_{\text{estimé}} = \frac{60}{1{,}30} * 1\,000 = \ 46\,154€$$

Quelques résultats des taux de passage PE CA ‰ sur le **haut de segment** :

Activité	Tx_passage_PE_CA_haut_segment ‰
Hotel-restaurant	1,48
Salle de cinéma	1,21
…	…
Ambulancier	0,56
Agence matrimoniale	0,29

Tableau 2.2.2 - Taux de passage entre la prime perte exploitation et le chiffre d'affaires sur le haut de segment

Application des taux de passage ‰ :

Exemple d'un hôtel-restaurant payant une prime PE = 500€ avec un montant de CA manquant. Le taux de passage pour l'hôtel-restaurant est de 1,48 donc son CA estimé est :

$$\mathrm{CA}_{\text{estimé}} = \frac{500}{1{,}48} * 1\,000 = \ 337\,838€$$

La prime PE varie selon le niveau de risque global de la structure à assurer. Il peut prendre en compte la nature de l'activité de l'entreprise (CA, valeur des locaux et du matériel utilisé au quotidien), la période d'indemnisation choisie. Nous pouvons remarquer que la laverie automatique, la boulangerie, l'hôtel-restaurant et salle de cinéma sont des activités qui ont un taux de passage élevé, autrement dit la prime PE payée est surtout liée au CA de l'entreprise. En effet, lorsqu'un sinistre de type incendie ou dégât des eaux survient pour ces deux activités, l'entreprise ne pourra plus du tout exercer d'activité et l'assureur devra indemniser une perte de CA considérable en plus des dommages de matériel…Tandis que l'agence de voyage, le bureau, l'ambulancier, et l'agence matrimoniale, ces activités peuvent toujours fonctionner en télétravail ou en changeant de locaux par exemple, par conséquent la prime PE payée est moindre.

A la fin de cette troisième imputation, il reste toujours 20% de manquants.

- **4ème étape d'imputation :** Moyenne par activité et cinq tranches de prime (Copht : Cotisation Potentielle HT)

Après les trois étapes d'imputations présentées précédemment, les dernières valeurs manquantes sont retraitées selon la moyenne par activité et cinq tranches de prime déterminées par quantile.

Ci-dessous quelques indicateurs statistiques sur la variable CA avant et après retraitement :

	Q1	Médiane	Q3	Moyenne	Valeurs manquantes et nulles
CA avant retraitement	0	76 537	243 749	291 247	33%
CA après retraitement	92 000	186 253	390 282	374 950	0%

Tableau 2.2.3 - Indicateurs statistiques sur la variable CA avant et après retraitements

Avant retraitement, la base de données est constituée de 33% de valeurs manquantes dont 31% de données nulles. Le chiffre d'affaires est largement sous-évalué à cause de cette sur-représentation de valeurs nulles. Après les différentes imputations, les valeurs manquantes sont remplacées et corrigées. Il reste très peu de valeurs manquantes après les retraitements (passage de 33% à quasiment 0%).

Surface / Contenu incendie / Qualité de l'occupant

Pour ces trois variables, le même raisonnement est appliqué avec la première et la quatrième approche d'imputation présentée sur la variable CA.

La même observation de sous-estimation liée à la sur-représentation des valeurs manquantes est constatée pour les variables surface et contenu incendie. Ci-dessous les indicateurs statistiques sur ces variables avant et après retraitement :

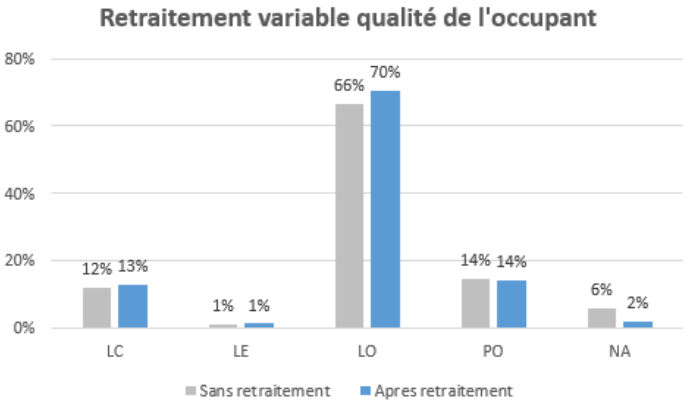
	Q1	Médiane	Q3	Moyenne	Valeurs manquantes et nulles
Surface avant retraitement	50	100	230	212	5%
Surface après retraitement	60	110	248	223	0%

Tableau 2.2.4 - Indicateurs statistiques sur la variable surface avant et après retraitement

	Q1	Médiane	Q3	Moyenne	Valeurs manquantes et nulles
Contenu INC avant retraitement	10 269	27 977	71 958	101 098	6%
Contenu INC après retraitement	13 016	34 591	82 439	110 167	0%

Tableau 2.2.5 - Indicateurs statistiques sur la variable contenu incendie avant et après retraitements

Ci-dessous un diagramme en barres représentant la répartition de la qualité des occupants avant et après retraitement :



Légende :
LC : Locataire qui souscrit pour le compte de son propriétaire,
LE : Locataire Exonéré de RC locative,
LO : Locataire,
PO : Propriétaire Occupant,
NA : Valeurs manquantes

Figure 2.2.4 - Variable qualité avant et après retraitement

Après retraitement, les valeurs manquantes passent de 6% à 2%, créant une augmentation des locataires occupants.

Antécédents de sinistres

Ce terme désigne les informations relatives au passé de l'assuré en termes de sinistre. Cependant, la définition donnée est fictive dans le cadre de cette étude, car naturellement les antécédents demandés à l'affaire nouvelle sont des antécédents des précédents assureurs. Or, la sinistralité passée des affaires nouvelles n'est pas renseignée dans les bases sinistres d'AXA. Ceci s'explique par la troncature à gauche, où il y a un manque d'information sur la sinistralité passée des affaires nouvelles.

Exemple d'une entreprise assurée chez l'assurance X, ayant eu un sinistre en 2017 et résilie son contrat en 2018 pour souscrire chez AXA. L'information de l'antécédent ne sera pas remontée dans les bases sinistres d'AXA et sera considérée comme une perte d'information car elle est non observable. Par conséquent, il n'est pas possible de récupérer les antécédents de sinistres pour les affaires nouvelles mais il est tolérable de créer des antécédents pour les contrats qui sont présents dans le portefeuille d'AXA depuis au moins un an. Exemple de l'entreprise qui arrive chez AXA en 2018, à vision 2018 il s'agit d'une affaire nouvelle donc il n'y aura pas d'antécédent, mais à vision 2019, s'il a eu un sinistre en 2018, il sera donc considéré comme un antécédent sur les douze derniers mois. Dans cette étude, le problème de la troncature n'a pas été pris en compte, mais cela pourrait faire l'objet d'un développement ultérieur, en faisant un modèle de troncature (Cf. annexe troncature).

Pour la modélisation, nous avons calculé les antécédents sur les douze derniers mois, les vingt-quatre derniers mois et les trente-six derniers mois. Le modèle sélectionnera le meilleur fractionnement. Ainsi, dans ce mémoire, les antécédents n'ont de sens que s'ils sont survenus entre le 1er janvier 2007 et le 31 décembre 2017, car la période d'étude est entre 2010 et 2018.

Dans un premier temps, il faut définir :

- L'année de vision
- La date de survenance des antécédents de sinistres compris entre 01/01/2007 et 31/12/2017
- La date de début du contrat de l'année civile $\stackrel{def}{=}$ max (date affaire nouvelle ; 01/01/année de vision)

Ensuite, si un sinistre est constaté entre la date de début du contrat et les trente-six mois qui la précède alors ce sinistre correspond à un antécédent de sinistralité. La variable de topage des antécédents sera égale à 1.

Ci-dessous un exemple, avec :

- Une date d'affaire nouvelle = 05/12/2008,
- Une année de vision = 2018,
- Une date de survenance sinistre = 20/12/2016,
- Une année de début du contrat de l'année civile
  $\stackrel{def}{=}$ max(date d'affaire nouvelle ; 01/01/année de vision) = max (05/12/2008 ; 01/01/2018) = 01/01/2018

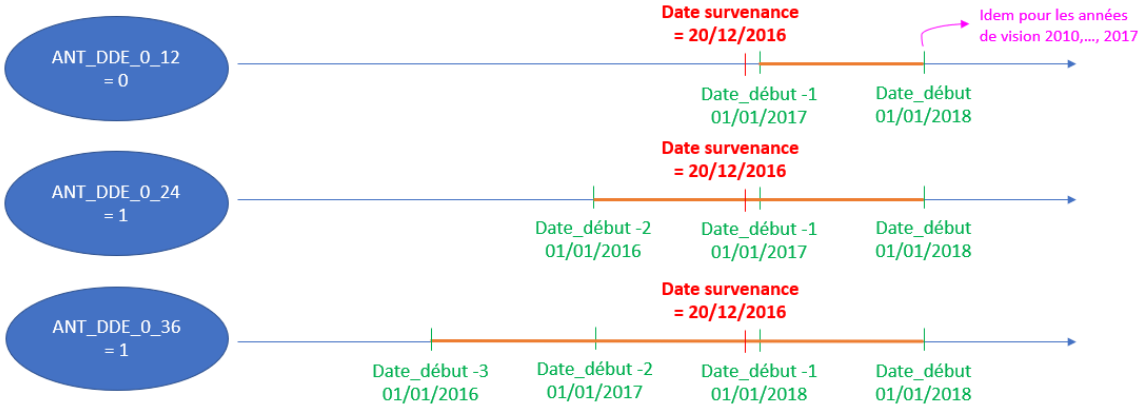


Figure 2.2.5 - Méthodologie pour créer la variable « antécédent de sinistres »

Pour l'année de vision 2018, il n'y a pas d'antécédent sur les douze derniers mois car la date de survenance du sinistre n'est pas comprise dans la période des douze derniers mois. Il y a un antécédent sur les vingt-quatre et trente-six derniers mois car la date de survenance est comprise dans les périodes référentes.

Remarque : Le sinistre vu à l'année 2017 sera considéré comme antécédent douze mois.

Lorsque plusieurs antécédents sont survenus, une somme est effectuée pour compter le nombre d'antécédents par année de vision et mois de recul (douze, vingt-quatre ou trente-six).

2.2.2 Catégorisation des variables quantitatives

Les variables quantitatives ont été segmentées afin de créer des tranches de risques homogènes. Pour cela, des premières tranches fines d'une centaine de classes ont été créées par quantile. Ensuite, ces tranches sont affinées grâce au résultat du modèle linéaire généralisé (GLM), qui attribue un coefficient pour chaque tranche. A partir de ces coefficients, des tests d'égalité de coefficient (test de Wald) sont appliqués permettant de regrouper les différentes tranches des variables ordinales. Ainsi les tranches ayant le même ou un coefficient significativement proche sont regroupées.

Test de Wald :

Le test de Wald peut être utilisé pour tester une seule hypothèse sur plusieurs paramètres, ainsi que pour tester conjointement plusieurs hypothèses sur un ou plusieurs paramètres.

Hypothèses :

$$\begin{cases} \mathrm{H}_0 = \mathrm{R}\beta = 0 \\ \mathrm{H}_1 = \mathrm{R}\beta \neq 0 \end{cases}$$

Avec :

- R une matrice d'ordre p x K ($p \leq K$)
- β coefficients du modèle

Statistique de test :

Sous H_0 : $\mathrm{W} = \left(\mathrm{R}\hat{\beta}\right)^{\mathrm{t}}[\hat{\sigma}_{\mathrm{n}}^2 \mathrm{R}(\mathrm{X}^{\mathrm{t}}\mathrm{X})\mathrm{R}^{\mathrm{t}}]^{-1}\left(\mathrm{R}\hat{\beta}\right) \sim \chi_{\mathrm{p}}^2$

Où

- $\hat{\beta}$ estimateur du maximum de vraisemblance des paramètres ou coefficients β
- $\hat{\sigma}_n^2$ estimateur du maximum de vraisemblance de la variance des erreurs
- χ_p^2 distribution de khi-deux avec p degré de liberté

Règle de décision :

Si $w^{obs} \leq \mathrm{q}_{\alpha}^{\chi_p^2}$ ou si la p-valeur du test $\geq 5\%$ alors on ne rejette pas H_0 au niveau de significativité α.

Si $w^{obs} > \mathrm{q}_{\alpha}^{\chi_p^2}$ ou si la p-valeur du test $< 5\%$ alors on rejette H_0 au niveau de significativité α.

Par exemple, pour tester l'égalité entre deux coefficients, la matrice $\mathrm{R} = (1 \quad -1)$ et $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$.

Il s'agit alors de tester :

$$\begin{cases} \mathrm{H}_0 = \beta_1 - \beta_2 = 0 \text{ ie } \beta_1 = \beta_2 \\ \mathrm{H}_1 = \beta_1 - \beta_2 \neq 0 \text{ ie } \beta_1 \neq \beta_2 \end{cases}$$

Les deux modalités sont regroupées lorsque l'on ne rejette pas H_0, c'est-à-dire qu'une égalité de coefficient est observée.

Le test du ratio de vraisemblance et le test du Score sont deux autres tests statistiques permettant de regrouper les modalités selon l'égalité des coefficients. Les trois tests sont équivalents asymptotiquement et leurs distributions statistiques suivent chacune la loi du khi-deux.

2.2.3 Regroupement des activités

L'activité est une variable discriminante dans la tarification. Le nombre d'activités disponibles dans la base était très important (environ 500 activités). Certaines activités étaient représentées par très peu de contrats, parfois un contrat. Il était indispensable de regrouper les activités avec une très faible part d'exposition pour une modélisation pertinente. La segmentation des activités est réalisée à partir d'un code NAF (Nomenclature d'Activités Française) et du nombre de contrats impliquant ce code. Il s'agit plus précisément d'un code activité spécifique à AXA, composé de sept caractères. Les trois premiers caractères permettent de différencier la famille d'activité et de faire une distinction sur le type d'activité. Les quatre autres caractères apportent un détail supplémentaire sur l'activité. Lorsque des catégories d'activité étaient compatibles, il a été convenu de regrouper les activités à faible exposition avec les activités les plus représentées (supérieures à 350 contrats). Ainsi, les activités sont regroupées selon les trois premiers caractères. Lorsque le regroupement n'est pas faisable un jugement à dire d'expert est adopté pour regrouper les activités ayant les mêmes risques de sinistralité. Exemple de regroupement des activités : le code activité **158**A100 (Boulangerie - pâtisserie industrielle), représenté par moins de 350 contrats, sera regroupé avec le code activité **158**C100 (Boulangerie pâtisserie, fabrication artisanale et commerce) qui est une activité avec plus de 350 contrats. Ce regroupement a du sens puisqu'il permet de réunir les activités qui sont similaires dans un même groupe tout en ayant une exposition suffisante pour la modélisation.

2.3 Traitement des variables à expliquer

Les variables à expliquer sont le nombre de sinistres et la charge de sinistre dans le cadre d'un modèle de fréquence x CM. Dans le cadre de la modélisation « Prime Pure » avec l'approche Tweedie, la variable charge de sinistre est la variable à expliquer.

2.3.1 La charge des sinistres

2.3.1.1 Distinction des charges

Charges négatives

Les sinistres ayant un montant négatif peuvent être expliqués par les recours. En cas de sinistre non responsable, AXA va tout d'abord indemniser son assuré puis faire un recours contre l'assureur du responsable. Lorsque ce montant est négatif, alors AXA a remboursé un montant inférieur au montant du recours touché in fine. Ces charges n'ont pas été exclues mais forcées à 0. Ce choix permet d'une part de garder la sinistralité en fréquence et d'autre part de modéliser la charge globale ou le coût moyen non négatif, respectivement par une loi de Tweedie ou une loi Gamma. La somme des montants négatifs des recours est ensuite ajustée sous forme de mutualisation, c'est-à-dire que la prime pure des assurés sera réduite de manière homogène selon le montant total négatif (correspondant à 0,30 centime par assuré).

La décomposition de la charge en attritionnelle, grave et atypique sera abordée après les sections de pré-traitements, vieillissement et inflation de la charge. Il est important de faire ces pré-traitements avant la distinction de la sinistralité car il se peut que la sinistralité vue dans le passé ne soit pas grave mais vue dans le présent, elle pourrait être considérée comme grave. Ainsi, cela permet de ramener tous les sinistres à leur coût réel et à une même unité monétaire.

2.3.1.2 Inflation

Le périmètre étudié correspond à un historique de 9 ans (2010-2018). Un sinistre ayant lieu en 2010 n'a pas le même coût en 2018. Il convient alors de prendre en compte le phénomène d'inflation. Ce traitement permettra de travailler sur des distributions de charges ultimes non biaisées par l'inflation. Ainsi, il faut réévaluer l'ensemble des charges et capitaux assurés pour les mettre en vision 2018 suivant deux indices d'indexation :

- **Indice FFB (Fédération Française du Bâtiment) :** Indice mis à jour trimestriellement (janvier, avril, juillet, octobre) par la FFB, appliqué pour le bas de segment. Il est calculé à partir du prix de revient d'un immeuble à Paris tout en prenant en compte les matériaux de construction, la main d'œuvre, les frais administratifs, etc. Il ne tient pas compte de la valeur des terrains.
- **Indice des Risques Industriels RI :** Indice mis à jour trimestriellement par la FFA (Fédération Française de l'Assurance), appliqué pour le haut de segment. Cet indice est calculé lui-même sur quatre indices :

  $$I_{RI} = 45 + 2{,}26\,A + 19{,}43\,B + 5{,}64\,C + 8{,}60\,D$$

  Où :

  - A est l'indice FFB du coût de la construction (source FFB - Base 1 au 1er janvier 1941) ;
  - B est l'indice mensuel du coût horaire du travail pour les industries mécaniques et électriques (source : INSEE - Base 100 décembre 2008 - Identifiant : 1565183) ;
  - C est l'indice du prix de vente industriel des métaux (source : INSEE - Base 2015 - Identifiant : 10534652) ;
  - D est l'indice du prix de vente des biens intermédiaires (source : INSEE - Base 2015 - Identifiant : 10534800).

Pour calculer l'inflation entre les différentes années, les indices du dernier trimestre ont été choisis. Ci-dessous l'évolution des deux indices au quatrième trimestre depuis 2010 :

	Indice FFB	Indice RI
2010	851,2	5240
2011	879,8	5573
2012	903,1	5690
2013	920,8	5753
2014	930,8	5772
2015	929,5	5819
2016	942	5783
2017	974,8	5948
2018	988,2	6100

Tableau 2.3.1 - Evolution des indices FFB et RI

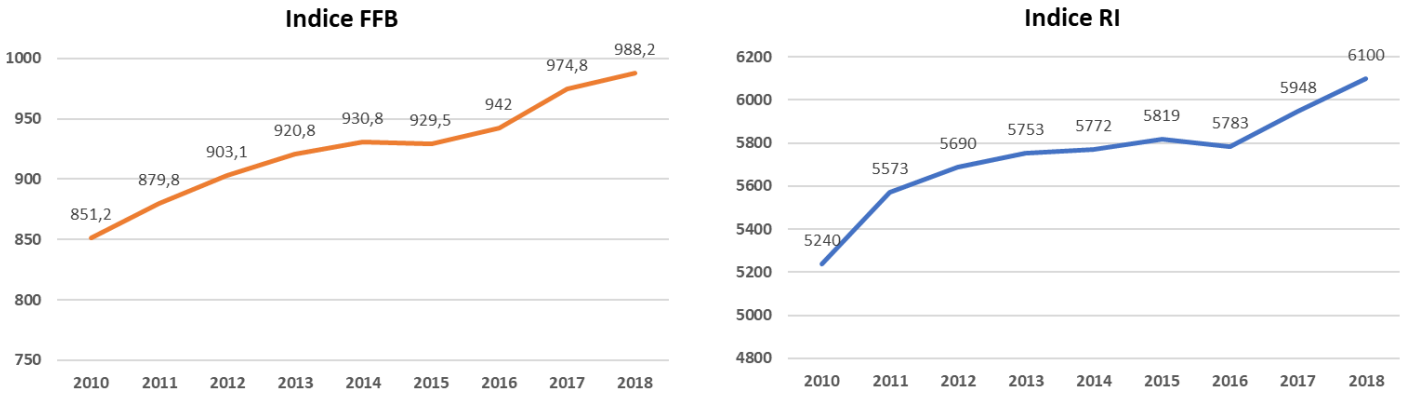


Figure 2.3.1 - Evolution des indices FFB et RI

La tendance est à la hausse pour les deux indices. Une légère diminution en 2015 pour l'indice FFB et en 2016 pour l'indice RI est constatée.

Les charges ultimes et capitaux assurés sont mis en « as if 2018 » en appliquant les coefficients d'inflation appropriés. Pour évaluer un montant d'une année en vision 2018, il suffit d'appliquer la formule suivante :

$$\text{Montant}_{\text{as if 2018}} = \text{Montant}_{\text{année}} * \frac{\text{Indice}_{2018}}{\text{Indice}_{\text{année}}}$$

Par exemple, un contrat MRP PP ayant eu un sinistre de 10 000€ en 2016, correspondra à un montant de $10\,000 * \frac{988,2}{942} = 10\,490€$ en vision 2018.

Ci-dessous les coefficients d'inflation calculés en vision 2018 :

	Indice FFB	Indice RI
2010	1,16	1,16
2011	1,12	1,09
2012	1,09	1,07
2013	1,07	1,06
2014	1,06	1,06
2015	1,06	1,05
2016	1,05	1,05
2017	1,01	1,03
2018	1,00	1,00

Tableau 2.3.2 - Coefficients d'inflation calculés en vision 2018

2.3.1.3 Vieillissement

La base de données créée présente une sinistralité selon la date d'observation de la charge. Pour la modélisation de la garantie DDE, il a été convenu initialement d'un vieillissement de deux ans. Ainsi, pour chaque année de vision, la charge est vue à N+2, soit deux ans plus tard, dans les bases sinistres d'AXA.

Cependant, le développement N+2 paraissait insuffisant. En effet, nous avons créé des triangles des charges par produit et vérifié l'écart entre la charge vue à N+2 et la charge ultime. Au global, en prenant la charge N+2, cette dernière est surestimée de 10% (une surestimation de 4% pour le produit MRP PP, 10% pour MRP EN et MPME et 7% pour la MRES).

Ci-dessous un graphique représentant le développement des charges de 2010 à 2018.

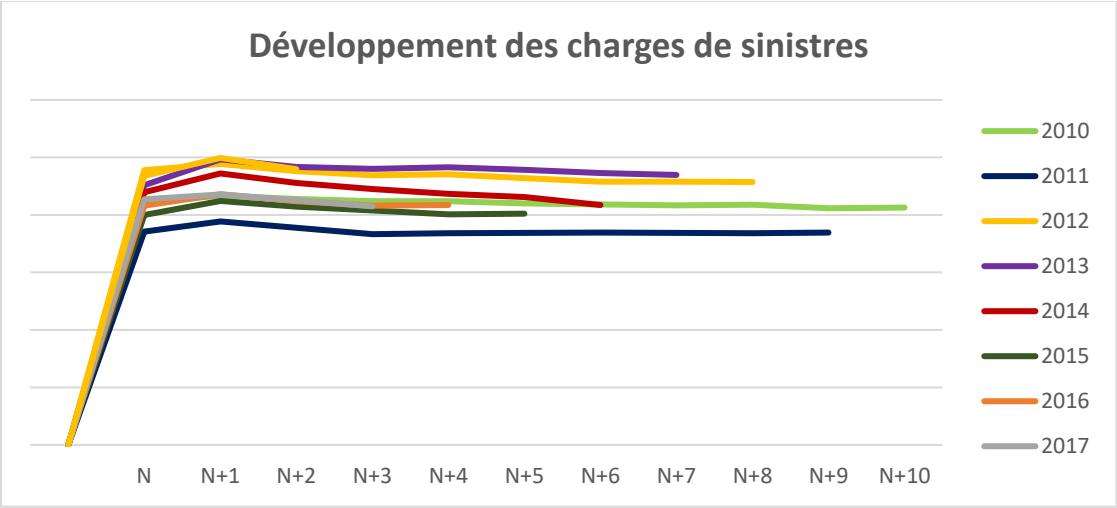


Figure 2.3.2 - Développement des charges de sinistres pour la branche RI

A partir de ce graphique, il apparaît qu'avec 2 ans de vieillissement la charge est largement surestimée. Les charges de l'année 2014 (courbe rouge) ont été particulièrement surestimées. Cette surestimation importante à l'ouverture du sinistre peut s'expliquer par la prudence de l'assureur. Celui-ci prévoit une charge considérable pour couvrir le risque mais après étude approfondie, il peut s'avérer que le sinistre coûte moins cher que le montant défini initialement ou que l'assuré a fait une fausse déclaration lors de la souscription ou alors que l'assuré n'a pas souscrit à la garantie… Ainsi, il est décidé de travailler par la suite avec la charge ultime. Pour cela, des méthodes classiques de provisionnement sont appliquées.

Tout d'abord, un sinistre se décompose en plusieurs dates clés schématisées sur la figure ci-dessous :

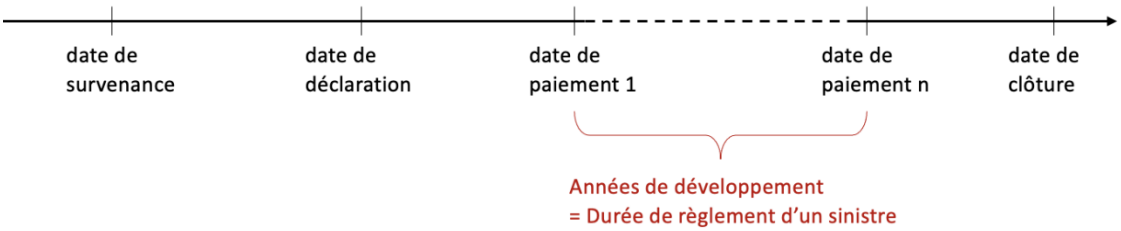


Figure 2.3.3 - Les dates clés de la vie d'un sinistre

Les notations utilisées ultérieurement :

- $i \in [0, \mathrm{n}]$ L'année de survenance d'un sinistre correspondant à l'année où le sinistre a eu lieu,
- $j \in [0, \mathrm{n}]$ L'année de développement,
- $X_{i,j}$ Paiement réglé durant l'année de développement j pour les sinistres survenus lors de l'année de i,
- $C_{i,j} = \sum_{k=0}^{j} X_{i,k}$ Les charges des j premières années de développement de l'année de survenance i.

Ci-dessous le triangle de charges pour chaque année de survenance :

Année de survenance \ Année de développement	0	…	j	…	n
0	$C_{0,0}$				$C_{0,n}$
…				…	
i	$C_{i,0}$		$C_{i,j}$		
…		…			
n	$C_{n,0}$				

Tableau 2.3.3 - Triangle de charges pour chaque année de survenance

La méthode usuelle de provisionnement est la méthode de Chain-Ladder. C'est une méthode déterministe qui suppose que les charges peuvent être connues à l'ultime à partir des données historiques. Donc son but est de donner une estimation de la charge ultime à l'aide des triangles de développement (triangles de paiements cumulés ou triangles de charges). Dans le cadre de cette étude, la méthode est appliquée au triangle des charges (seul les sinistres hors atypiques font partis de l'étude, Cf. section 2.3.1 Charge des sinistres). Le triangle est un tableau à double entrée dont seule la partie supérieure en bleue est connue, correspondant aux données historiques). Les lignes du tableau correspondent aux années de de survenance des sinistres et les colonnes aux années de développement (représentant la durée de règlement d'un sinistre). Chaque sinistre survenu l'année i, donne lieu à un ou plusieurs règlements sur plusieurs années de développement j. Par souci de simplification, les sinistres sont considérés comme complétement développés au plus tard après n années de développement. Ainsi, la charge ultime relative à une année de survenance i correspond à la valeur de $C_{i,n}$. Le but est d'estimer les paiements futurs (la partie grisée du tableau).

Cette méthode repose sur deux hypothèses :

- (H1) : Les sinistres sont clos après l'année de développement n,
- (H2) : $\forall$ i = 0,…,n et $\forall$ j = 0,…, n-1, $f_j = \frac{C_{i,j+1}}{C_{i,j}}$ Les facteurs de développement dépendent uniquement de l'année de développement j et sont indépendants de l'année de survenance i. Cette hypothèse peut se réécrire sous la forme suivante :
  $\forall$ i = 0,…,n et $\forall$ j = 0,…, n-1, $\frac{C_{0,j+1}}{C_{0,j}} = \frac{C_{1,j+1}}{C_{1,j}} = \dots = \frac{C_{n-j-1,j+1}}{C_{n-j-1,j}}$

Ainsi, l'estimateur du facteur de développement est défini comme :

$\forall$ j = 0,…, n-1, $\widehat{f_j} = \frac{\sum_{i=0}^{n-j-1} C_{i,j+1}}{\sum_{i=0}^{n-j-1} C_{i,j}}$

Ces facteurs permettent d'estimer itérativement les charges inconnues pour une année de survenance i associés à une année de développement j, dans le triangle de développement. Ainsi, il suffit de multiplier les charges associées à l'année de développement précédente par le facteur correspondant. L'estimateur des charges futures s'écrit donc en fonction des facteurs de développement :

$\widehat{C_{i,J}} = \widehat{f_{j-1}} * C_{i,j-1}$

La partie grisée du tableau peut être remplie avec les estimations des charges obtenues :

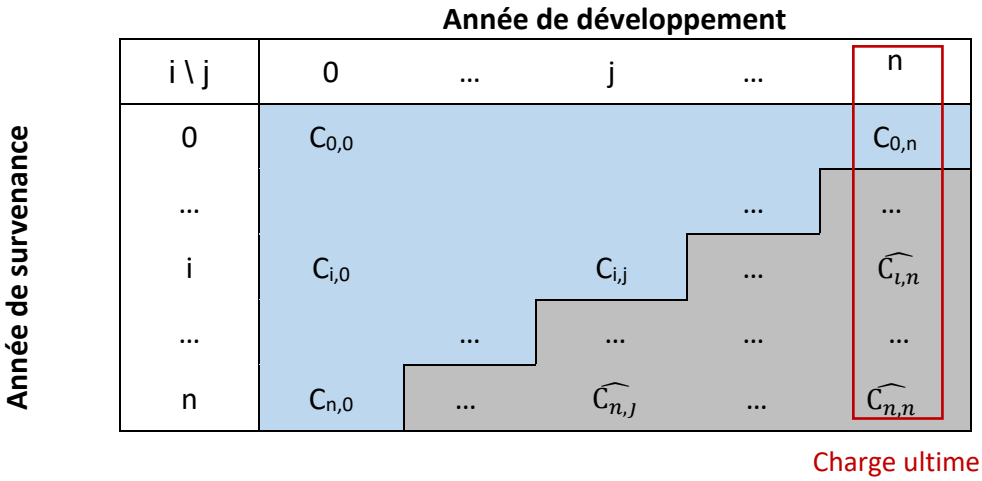


Tableau 2.3.4 - Triangle des charges pour chaque année de survenance avec leurs estimations

Une fois la charge ultime obtenue à partir de la méthode de Chain-Ladder pour chaque année de survenance, des coefficients de vieillissement sont calculés entre la charge vue à N+2 et ultime. Ces coefficients permettent de déterminer la charge ultime pour les charges vues à N+2 (choix initial).

$$\text{Coefficient de vieillissement charge N} + 2 \text{ à ultime} = \frac{\text{Charge vue à N+2 obtenue avec la base sinistre d'AXA}}{\text{Charge ultime obtenue avec la méthode Chain Ladder}}$$

Ci-dessous un tableau et un graphique des coefficients de vieillissement selon l'année de survenance et le produit :

	MRP PP	MRP EN / MPME	MRES
2010	1,04	1,13	1,06
2011	0,99	1,06	1,10
2012	0,99	1,04	1,04
2013	1,05	1,16	1,09
2014	1,12	1,16	1,06
2015	1,06	1,01	1,06
2016	1,05	1,07	1,06
2017	1,06	1,17	1,08
2018	1,05	1,09	1,05

Tableau 2.3.5 - Coefficients de vieillissement selon l'année de survenance et le produit

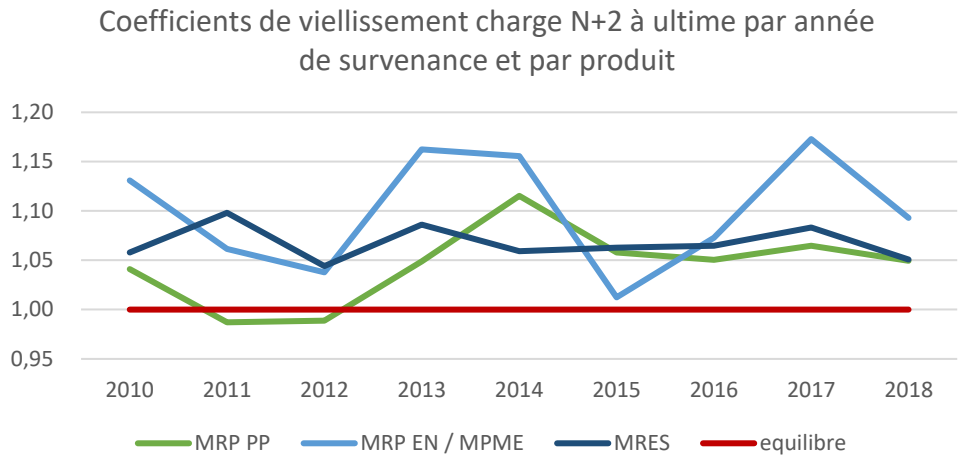


Figure 2.3.4 - Coefficients de vieillissement selon l'année de survenance et le produit

Graphiquement, il est observable que généralement la charge vue à N+2 est surestimée par rapport à la charge ultime, à l'exception entre les années 2011 et 2012 pour le produit MRP PP. Par ailleurs pour les produits MRP, MRP EN et MPME (bas de segment) les coefficients de vieillissement varient fortement. L'estimation de la charge pour le produit MRES est plus stable que les produits du bas de segment.

Pour obtenir la charge ultime pour chaque sinistre, ces coefficients sont ensuite divisés par la charge vue à N+2 de la base sinistre d'AXA selon l'année de survenance et le produit.

Par exemple :

- Année N de survenance du sinistre = 2016
- Produit = MRP PP
- Charge vue à N+2 (2018) = 5 000€

$$\text{Charge ultime} = \frac{\text{charge vue à N+2}}{\text{coefficient de passage}_{2016,\ \text{MRP PP}}} = \frac{5\,000}{1{,}05} = 4\,761{,}9€$$

La charge ultime est inférieure à la charge vieillie de 2 ans, puisque cette dernière est surestimée par rapport à la charge ultime.

Finalement, pour étudier et réévaluer le coût final, il a fallu deux coefficients :

- Coefficient de vieillissement charge N+2 à ultime
- Coefficient d'inflation

Ainsi, le coût final réévalué s'obtient :

$$\text{Montant final}_{\text{as if 2018}} = \text{Montant}_{\text{année}} * \underbrace{\frac{\text{Coefficient inflation}_{\text{année}}}{\text{Coefficient vieillissement}_{\text{année}}}}_{\text{Coefficient de passage final}}$$

Nous obtenons les coefficients de passage final par produit et par année de survenance :

	MRP PP	MRP EN / MPME	MRES
2010	1,12	1,03	1,10
2011	1,14	1,06	1,00
2012	1,11	1,05	1,03
2013	1,02	0,92	0,98
2014	0,95	0,92	1,00
2015	1,00	1,05	0,99
2016	1,00	0,98	0,99
2017	0,95	0,86	0,95
2018	0,95	0,91	0,95

Tableau 2.3.6 - Coefficients de passage final par produit et par année de survenance

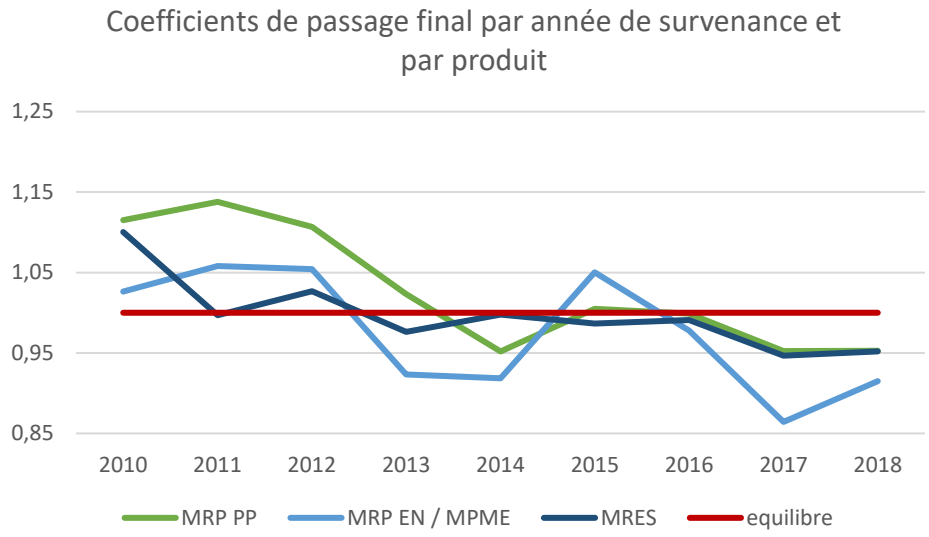


Figure 2.3.5 - Coefficients de passage final par année de survenance et par produit

La charge finale sera parfois réévaluée à la hausse (coefficients supérieurs à 1) due à l'inflation et parfois à la baisse (coefficients inférieurs à 1) due aux surestimations importantes qui camouflent l'effet d'inflation. Nous remarquons que plus l'historique est profond plus l'effet d'inflation est visible. La charge est réévaluée à la hausse pour les années 2010 à 2013 (le montant du sinistre passé sera beaucoup plus élevé vu à une vision ultérieure).

2.3.1.4 Charges graves

Au sein d'AXA France, le montant de la charge a été découpée en trois parties :

- **Charge attritionnelle**, correspondant aux sinistres ayant une charge comprise entre 0 et le seuil grave préconisé par AXA (soit 150 000€) et les sous-crêtes des sinistres graves.
- **Charge grave (sur-crête grave)**, représentant les sinistres dont la charge est supérieure au seuil grave d'AXA et ne dépassant pas le seuil atypique,
- **Charge atypique**, regroupant les sinistres dont la charge est supérieure au seuil atypique d'AXA. Ce seuil peut être différent selon le produit, il est spécifié à 500 000€ pour les produits MRP EN et MPME, 600 000€ pour le produit MRP PP et 1 200 000€ pour le produit MRES. Ce sont des sinistres atypiques, très rares qui ne reflètent pas la sinistralité réelle du contrat, ils ne seront donc pas pris en compte dans la modélisation.

Ci-dessous un schéma de la décomposition de la charge avec les différents seuils d'AXA :

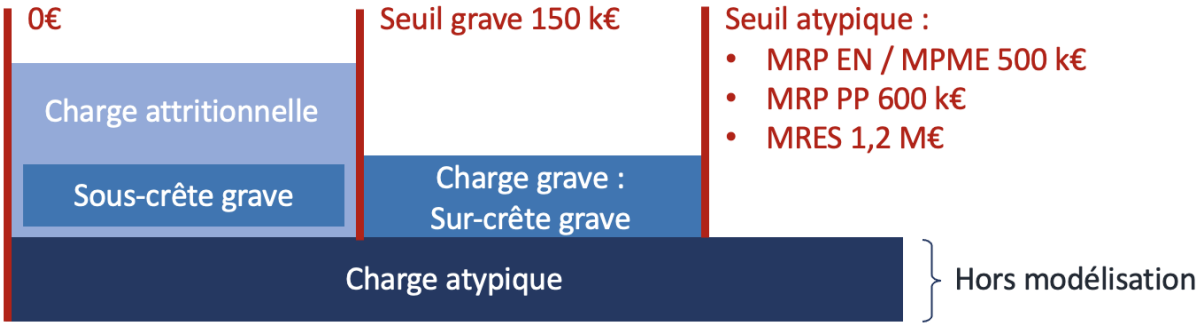


Figure 2.3.6 - Décomposition de la charge avec les différents seuils d'AXA

Le seuil qui distingue la **charge attritionnelle** de la **charge grave** est de 150 000€ pour l'ensemble des garanties de la branche RI. Or, ce seuil n'est pas forcément optimal, adapté et se relève trop important pour la garantie DDE. Ainsi, un nouveau seuil grave a été déterminé à l'aide de la théorie des valeurs extrêmes (TVE). Cette dernière permettra de définir un seuil qui séparera les sinistres fréquents et peu coûteux (sinistres attritionnels) et les sinistres rares et coûteux (sinistres graves).

Le choix de ce seuil nécessite un arbitrage entre choisir :

- un seuil suffisamment grand pour permettre une distribution optimale mais limitant le nombre de données de sinistres graves,
- un seuil assez petit pour disposer d'un nombre d'observations suffisant et d'éviter un biais trop important dans la modélisation.

Pour ce faire, la TVE sera appliquée. Dans un premier temps, l'aspect théorique sera présenté avec les domaines d'attraction des valeurs extrêmes, suivi par les différents estimateurs pour déterminer le seuil. Après avoir expliqué cette théorie, elle sera appliquée sur les données de la base.

A noter que les différents seuils atypiques d'AXA qui permettent de distinguer la **sinistralité grave** de celle **atypique** ont été laissés tels quels. Il aurait été plus judicieux de challenger également les trois seuils atypiques par la TVE pour la garantie DDE.

Aspect TVE :

Contrairement à l'approche classique qui consiste à modéliser toute la distribution, la TVE vise à modéliser les évènements rares. Elle s'intéresse donc uniquement à la queue de distribution, qui permet ensuite d'estimer le comportement asymptotique des extrêmes d'un échantillon de variables aléatoires indépendantes et identiquement distribués (iid). Cette théorie est fondée sur un théorème fondamental, appelé le théorème de Fisher-Tippett. Il existe deux approches, à savoir l'approche des blocs maxima et l'approche de dépassement de seuil ou Peak Over Threshold (POT).

Domaine d'attraction

Soit $X_1, \dots, X_n$ une suite de variables aléatoires iid, de fonction de répartition F.

On définit les maxima, $M_n = \max(X_1, \dots, X_n)$ alors,

$$\begin{aligned}\mathbb{P}(M_n \leq x) &= \mathbb{P}(\max(X_1, \dots, X_n) \leq x) \\ &= \mathbb{P}(X_1 \leq x, \dots, X_n \leq x) \\ &= \mathbb{P}(X_1 \leq x) \dots \mathbb{P}(X_n \leq x) \\ &= [F(x)]^n\end{aligned}$$

La loi de M_n dépend de la fonction de répartition F, qui est généralement inconnue. Si F n'est pas connue, cette formule est peu utile. Il est donc difficile de déterminer la loi des extrêmes à partir de la fonction de répartition. Fisher et Tippett se sont intéressés au comportement asymptotique des variables aléatoires M_n. En particulier, il est important de disposer d'une expression simple pour la loi asymptotique et il serait souhaitable que cette expression ne dépende pas de F.

Notons :

- Si $F(x) < 1$, alors $\mathbb{P}(M_n \leq x) \underset{n\to\infty}{\longrightarrow} 0$
- Si x^F le point extrémal de F, i.e. $x^F = \sup\{x : F(x) < 1\}$, alors $M_n \xrightarrow{\mathbb{P}} x^F$ quand $n \to \infty$. La distribution asymptotique de M_n est donc dégénérée. Ceci suggère qu'il faut passer par une transformation ou une normalisation.

Le premier théorème fondamental de la théorie des valeurs extrêmes permet de caractériser la loi de distribution des extrêmes.

Théorème (Fisher-Tippett)

Soit $(X_n)_{n\geq 1}$ une suite de n variables aléatoires iid, et de même loi de fonction de répartition F telle que $F(x) = \mathbb{P}(\mathrm{X} \leq \mathrm{x})$.

S'il existe des suites normalisantes réels a_n et b_n et une loi non dégénérée[6] G telle que :

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) = \lim_{n\to\infty} [F(a_n x + b_n]^n = G(x), \qquad \forall\, x \in \mathbb{R}$$

Alors G est du même type que l'une des trois distributions suivantes :

$$Fréchet\ (\alpha > 0): \qquad \phi_\alpha(x) = \begin{cases} 0 & si\ x \leq 0 \\ \exp(-x^{-\alpha}) & si\ x > 0 \end{cases}$$

$$Weibull\ (\alpha < 0): \qquad \psi_\alpha(x) = \begin{cases} \exp(-(-x)^{\alpha}) & si\ x \leq 0 \\ 1 & si\ x > 0 \end{cases}$$

$$Gumbel\ (\alpha = 0): \qquad \Lambda_\alpha(x) = \exp(-e^{-x}) \quad x \in \mathbb{R}$$

Dans le cas de problèmes statistiques, il est préférable d'avoir une seule distribution qui unifie les trois précédentes. Von Mises (1945) et Jenkinson (1955) ont proposé une famille paramétrique de distribution, appelée la distribution généralisée des valeurs extrêmes, notée $GEV(\mu, \sigma\ , \xi)$.

$$G_{\mu,\sigma,\xi}(x) = \begin{cases} \exp\left(-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-\frac{1}{\xi}}\right) si\ \xi \neq 0 \\ \exp\left(-exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right)\ si\ \xi = 0 \end{cases}$$

Où $x_+ = \max(x, 0)\ et\ \sigma > 0$.

- ξ est le paramètre de forme ou indice de queue qui donne une indication sur la forme de la distribution extrêmes,
- μ le paramètre de position,
- σ le paramètre d'échelle qui caractérise la dispersion.

$G_{\mu,\sigma,\xi}(x)$ est appelée fonction de répartition de la loi des valeurs extrêmes, Generalized Extreme Value distribution (GEV).

- Si $\xi > 0$, la GEV est dans le domaine d'attraction de **Fréchet**, caractérisé par des distributions à queues lourdes.
- Si $\xi = 0$, la GEV est dans le domaine d'attraction de **Gumbel** avec des distributions à queues légères.
- Si $\xi > 0$, la GEV est dans le domaine de **Weibull** dont les queues de distribution sont bornées.

[6] loi non dégénérée : la variance de la loi est non nulle

Le tableau ci-dessous montre les trois domaines d'attraction accompagnés de quelques exemples de lois usuelles :

Fréchet $\xi > 0$	**Gumbel $\xi = 0$**	**Weibull $\xi < 0$**
Cauchy Pareto généralisée Student Log-Gamma	Normale Exponentielle Log-normale Gamma Weibull	Uniforme Beta

Tableau 2.3.7 - Exemples de lois usuelles pour les domaines d'attraction

Estimateurs du seuil

Afin de déterminer un nouveau seuil grave spécifique à la garantie DDE trois estimateurs seront utilisés : l'estimateur de Hill, l'estimateur de Pickands et l'estimateur de Gerstengarbe.

Estimateur de Hill :

L'estimateur de Hill (1975) est l'estimateur le plus utilisé lorsque les données appartiennent au domaine d'attraction de Fréchet ($\xi > 0$), distributions à queues épaisses. Il est défini de la façon suivante :

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

Où k est la statistique d'ordre, le nombre d'observations considérées.

Le choix de k est crucial. S'il est trop grand, l'approximation par une loi de Pareto sera médiocre et l'estimateur de Hill aura un biais important. A l'inverse, s'il est trop petit, le nombre d'observations sera trop faible, donc l'estimateur aura une variance importante. En pratique, on utilise une méthode graphique : Hill Plot pour déterminer l'estimateur et le seuil optimal. Le seuil est identifié à partir de la zone de stabilité de l'estimateur.

Estimateur de Pickands :

L'estimateur de Pickands (1975) est défini par la statistique suivante :

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

L'avantage de cet estimateur est qu'il peut être utilisé quel que soit le domaine d'attraction (Gumbel, Weibull ou Fréchet). Cependant, il est très sensible à la taille de l'échantillon, ce qui peut le rendre peu robuste. La représentation graphique de cet estimateur se présente en fonction du nombre k d'observations considérées (statistiques d'ordre).

Estimateur de Gerstengarbe :

Cette méthode a été proposée par Gerstengarbe et Werner en 1989. Elle permet de déterminer un point de départ de la région extrême en donnant une estimation du seuil optimal. En effet, à partir de la zone de sinistre extrême, une modification du comportement des différences de coût entre deux sinistres Δ_i est observable. Le comportement des Δ_i pour les observations extrêmes est différent du comportement des Δ_i pour les observations non extrêmes. Nous cherchons donc à identifier un changement dans une série.

Soit $x_1, \dots, x_n$ l'échantillon de coûts de sinistres.

On considère la série des différences $\Delta_i = x_{[i]} - x_{[i-1]}, i = 1, \dots, n$ de l'échantillon ordonné $x_1 \leq \cdots \leq x_n$.

L'estimateur de Gerstangarbe est défini par la statistique :

Pour $i = 1, \dots, n-1$ nous calculons la série U_i telle que :

$$U_i = \frac{U_i^* - \frac{i(i-1)}{4}}{\sqrt{\frac{i(i-1)(2i+5)}{72}}}$$

Où $U_i^* = \sum_{k=2}^{i} n_k$ et n_k le nombre de valeurs $\Delta_2, \dots, \Delta_k$ inférieures à Δ_k. De la même manière, une autre série décroissante des différences $\Delta_n, \dots, \Delta_2$ est calculée. Le point d'intersection de ces deux séries détermine le seuil d'entrée dans la zone extrême.

Application de la TVE :

Méthodes permettant de déterminer le seuil grave

1- Étude du domaine d'attraction

Dans un premier temps, il est nécessaire d'identifier le domaine d'attraction pour ensuite appliquer les méthodes de détermination du seuil. La détermination d'un nouveau seuil est intéressante si les données appartiennent au domaine d'attraction de Fréchet, distribution à queue épaisse. Si c'est le cas, ce seuil déterminé sera pertinent car il séparera les sinistres graves qui auront un faible nombre d'observation et une charge globale importante. Autrement dit, il apparaîtra nécessaire de séparer les sinistres graves des sinistres attritionnels.

Pour identifier le domaine d'attraction, il est possible d'estimer le paramètre de forme ξ de la loi GEV ou de la loi GPD ou d'utiliser des quantiles plots généralisés.

a) Estimation du paramètre de forme

Le paramètre de forme ξ doit être strictement supérieur à 0 pour considérer que les données appartiennent au domaine de Fréchet. L'estimation du paramètre de la distribution EVD de la loi GEV par la méthode de maximum de vraisemblance donne un $\hat{\xi}$ = 0,55. Cette estimation vérifie l'appartenance des données au domaine d'attraction de Fréchet.

b) Quantile plot *généralisé*

Le Quantile plot généralisé permet de déterminer graphiquement le signe de ξ. Pour ce faire, le QQ-plot est tracé avec la loi exponentielle, où l'axe des ordonnées représente les quantiles empiriques de l'UH scores et l'axe des abscisses correspond aux quantiles théoriques de la fonction de distribution exponentielle. Ainsi, les coordonnées sont caractérisées comme : $(X; Y) = \left(\log\left(\frac{n+1}{j}\right); \log(UH_{j,n})\right)$ où $UH_{j,n} = X_{n-j,n}\hat{\xi}_{j,n}^{(H)}$, avec $\hat{\xi}_{j,n}^{(H)} = \frac{1}{j}\sum_{i=1}^{j} \log(X_{n-i+1,n}) - \log(X_{n-j,n})$ l'estimateur de Hill

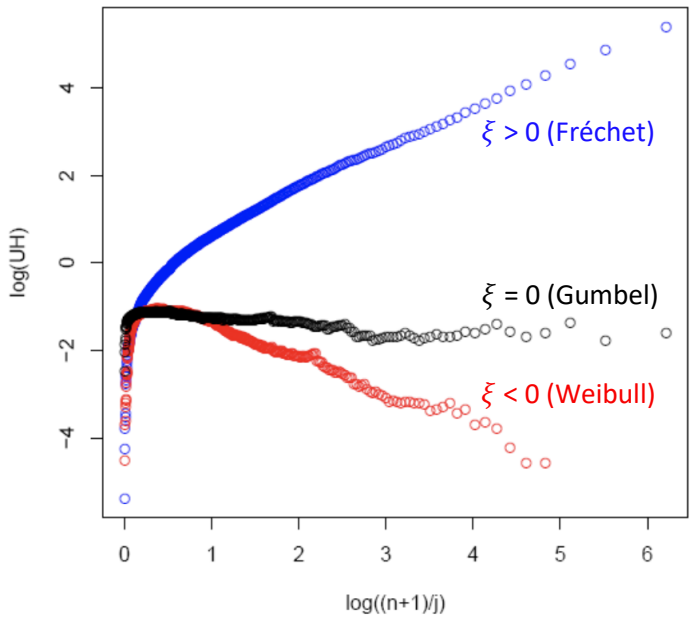


Figure 2.3.7 - Quantile plot généralisé attendu selon la valeur de ξ

Trois cas de figures sont possibles,

- $\xi > 0$ qui implique un domaine de Fréchet (en bleu). La distribution est à queue épaisse. Cela justifie la présence d'un grand nombre de valeurs extrêmes au niveau de la queue de distribution,
- $\xi = 0$ qui implique un domaine de Gumbel (en noir). Les données forment une droite de pente a, alors la distribution suit une loi exponentielle et présente une queue légère,
- $\xi < 0$ qui implique un domaine de Weibull (en rouge). La distribution a un support borné supérieurement.

Le quantile plot généralisé sur les données d'étude valide que le signe de ξ est strictement positif, donc les données appartiennent au domaine d'attraction de Fréchet. Ainsi, il y a un intérêt de déterminer un seuil grave.

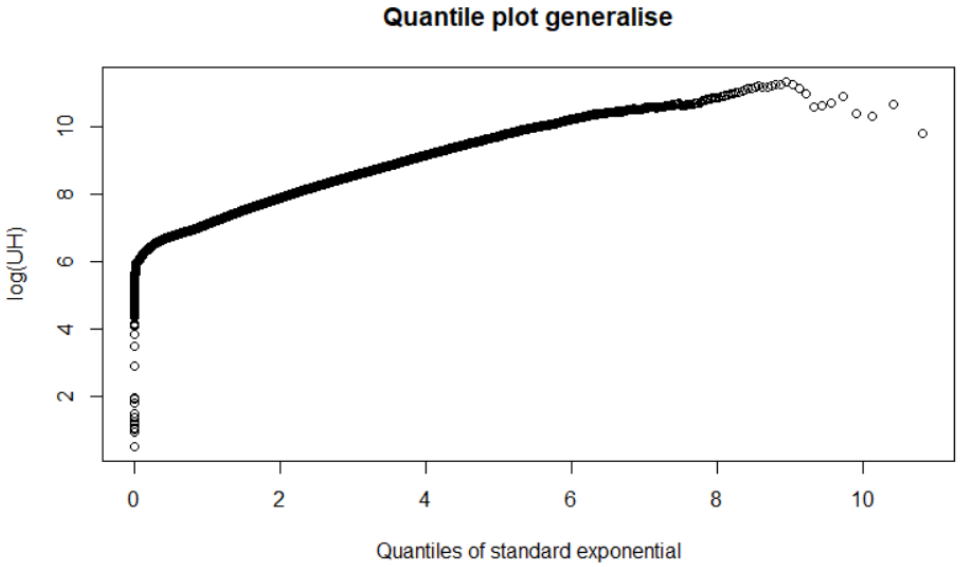


Figure 2.3.8 - Quantile plot généralisé des données RI

2- Application des estimateurs

a) Estimateur de Gerstengarbe

Cette méthode graphique de détermination de seuil, permet de déterminer le point de départ de la région extrême en donnant une estimation du seuil optimal. L'intersection des deux courbes désigne le changement de comportement de la sinistralité. Il indique une statistique d'ordre de k = 1 300, correspondant à environ 1,3% d'observations de la base totale (environ 3 millions d'observations) et le seuil estimé est de 19 083€.

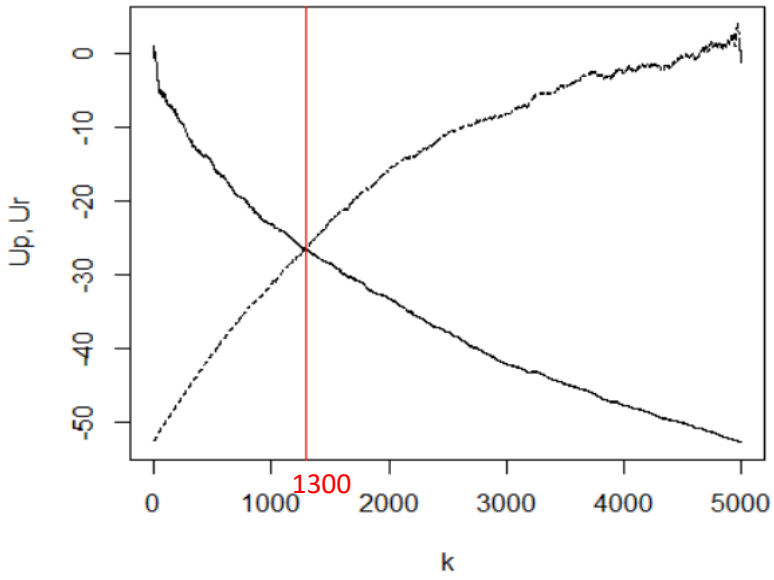


Figure 2.3.9 - Estimateur de Gerstengarbe

b) Estimateur de Hill

Le graphique de l'estimateur de Hill représente l'évolution des estimateurs en fonction du nombre de statistiques d'ordre k, ou nombre de dépassements. Cet estimateur est sensible à la taille de l'échantillon, il est volatile lorsque k est faible et se stabilise au fur et à mesure avec l'augmentation de la statistique d'ordre. Le seuil est identifié à partir de la zone de stabilité de l'estimateur. L'intervalle de confiance à 95% de l'estimateur de Hill est également tracé sur le graphique pour permettre de mieux détecter la zone de stabilité. Plus le nombre d'observations est grand plus l'intervalle est petit.

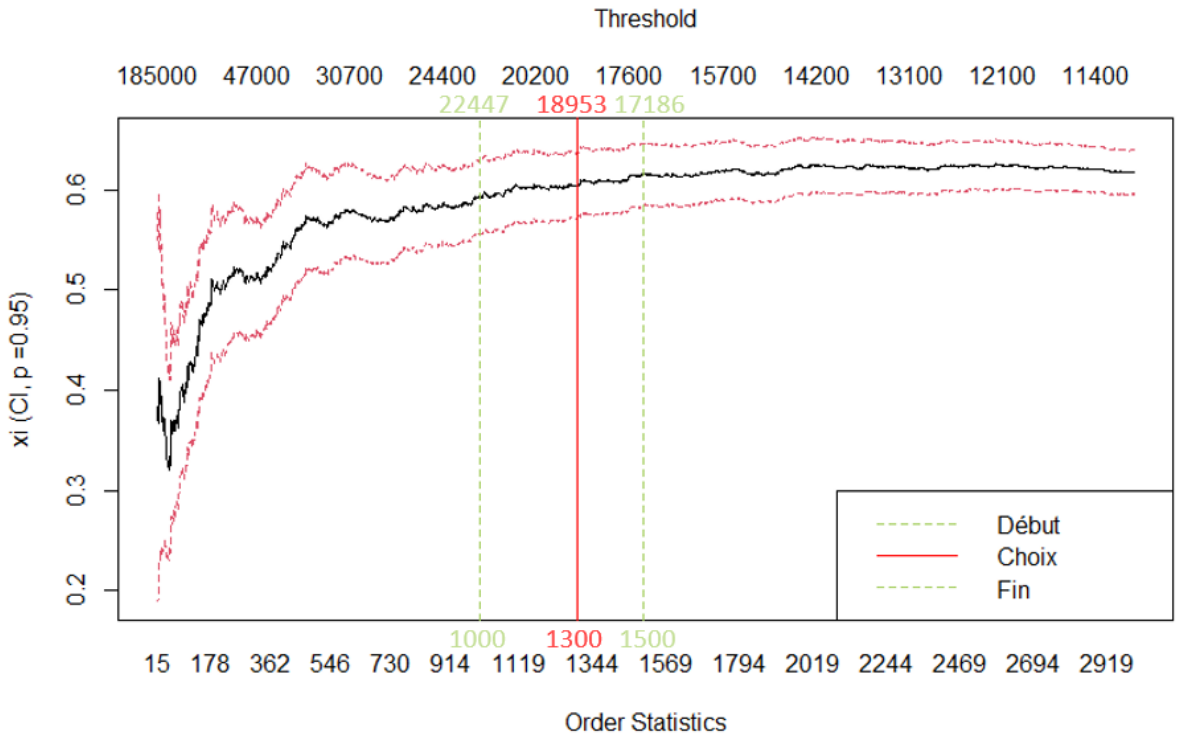


Figure 2.3.10 - Estimateur de Hill

Sur le graphique ci-dessus, l'axe des abscisses représente la statistique d'ordre du vecteur de sinistre triée par ordre croissant et l'axe des ordonnées correspond aux valeurs de l'estimateur de Hill. Trois traits verticaux sont représentés sur le graphique, les deux traits en verts représentent un premier plateau de stabilité, qui est détecté entre les statistiques d'ordre 1 000 et 1 500. Le trait vertical rouge représente la statistique d'ordre du seuil optimal obtenu par l'estimateur de Gerstengarbe. Cette dernière appartient à la zone de stabilité de l'estimateur de Hill, ce qui permettrait de valider ce seuil candidat. Nous remarquons que la statistique d'ordre k = 1 300 associées à l'estimateur Gerstengarbe donne un seuil légèrement différent de l'estimateur de Hill (19 083€ contre 18 953€) Par la suite ce seuil sera arrondi à 19 000€. Ainsi, un seuil de 19 000€ semblerait séparer les sinistres graves des sinistres attritionnels. Par ailleurs l'estimateur de Hill pour ce seuil candidat est environ égal à 0,6, ce qui est strictement supérieur à zéro, donc assure que le domaine d'attraction est bien celui de Fréchet.

c) Estimateur de Pickands

Le graphique ci-dessous, représentant l'estimateur de Pickands est un autre moyen de déterminer le seuil grave. Le principe est le même que l'estimateur de Hill, à savoir chercher une zone de stabilité. L'estimateur de Pickands paraît plus volatile que celui du Hill lorsque le nombre de statistiques d'ordre est faible. La stabilité est de nouveau observée en reprenant les mêmes statistiques d'ordre que précédemment 1 000 et 1 500. Le seuil retenu par l'estimateur de Gerstengarbe, représenté en rouge rentre dans cette zone de stabilité. De plus, l'estimateur de Pickands pour ce seuil est environ égal à 0,6, ce qui est strictement supérieur à zéro. Les résultats sont cohérents avec les analyses précédentes.

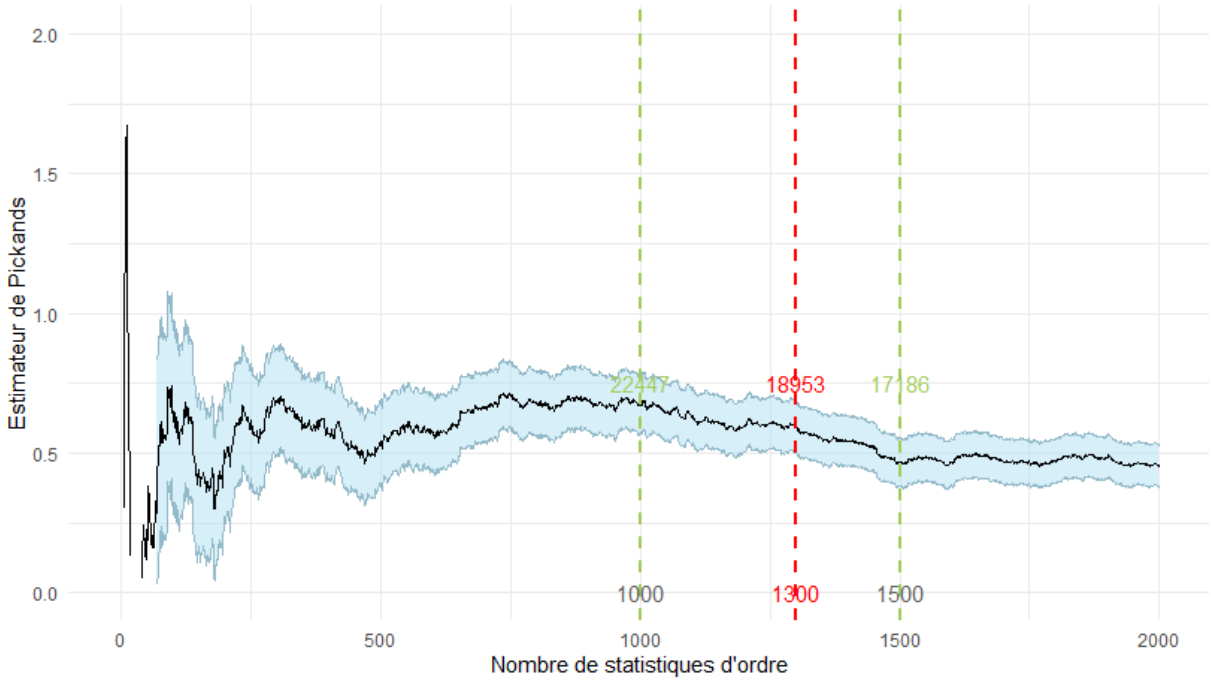


Figure 2.3.11 - Estimateur de Pickands

Finalement, le seuil 19 000€ retenu par l'estimateur de Gerstengarbe rentre dans la zone de stabilité pour les deux estimateurs étudiés. Ce seuil est donc justifié et sera utilisé dans l'écrêtement et la mutualisation des charges.

Ci-dessous la proportion de charge sous-crête, de charge sur-crête et de sinistres graves :

Seuil retenu	% charge sous-crête	% charge sur-crête	% sinistres graves
19 000	77,4%	22,6%	1,3%

Tableau 2.3.8 - Détails sur le seuil retenu

Avec le nouveau seuil déterminé, 1,3% des sinistres sont considérés comme graves lorsque le montant de la charge est supérieur à 19 000€. Les résultats montrent qu'environ 22,6% de la charge sur-crête grave sera mutualisée.

Après avoir déterminé le seuil grave, il est nécessaire d'écrêter la charge pour mutualiser uniquement la sur-crête des sinistres graves.

Écrêtement de la charge

La méthode d'écrêtement de la charge de sinistre est présentée ci-dessous :

Si la charge totale hors atypique est supérieure au seuil grave, alors la charge sous-crête est égale au seuil grave 19 000€. La charge sur-crête (charge grave) correspond à la différence entre la charge totale hors atypique et la charge sous-crête. Sinon la totalité de la charge totale hors atypique est conservée en tant que charge sous-crête.

Soit :

Si $\text{Charge}_{\text{totale hors atypique}}$ > Seuil grave (19 000€) alors,

$$\text{Charge}_{\text{sous-crête}} = \text{Seuil grave} = 19\,000€$$

$$\text{Charge}_{\text{sur-crête}} = \text{Charge}_{\text{totale hors atypique}} - \text{Charge}_{\text{sous-crête}}$$

Sinon,

$$\text{Charge}_{\text{sous-crête}} = \text{Charge}_{\text{totale hors atypique}}$$

$$\text{Charge}_{\text{sur-crête}} = 0$$

Mutualisation des charges graves

Une fois que l'écrêtement est réalisé, l'étape suivante est d'appliquer le principe de mutualisation.

La mutualisation est un principe fondamental en assurance. Il consiste à répartir le coût d'un sinistre entre les assurés soumis potentiellement au même risque. Sans la mutualisation, un assuré ayant été impacté par un sinistre grave, survenu sur la période étudiée, devra payer une prime largement supérieure à celle d'un assuré n'ayant pas eu de sinistre grave. Ces sinistres graves correspondent à des évènements rares qui peuvent conduire à une sinistralité très importante. Afin que les sinistres graves ne viennent pas modifier de manière considérable la prime calculée, nous sommes amenés à utiliser le principe de mutualisation. Ainsi, la mutualisation sur la sur-crête des sinistres graves est une méthode qui permettrait de palier le problème de sinistres extrêmes tout en prenant en compte la fréquence réelle des sinistres. Plus concrètement, il s'agit de mutualiser la sur-crête grave au prorata des primes acquises DDE des sinistrés. Cependant, cette méthode ne constitue pas une modélisation réelle de la charge observée et suppose que la prime DDE soit bien calibrée historiquement. Par ailleurs, cette mutualisation aurait pu être effective sur l'ensemble des contrats et non uniquement les sinistrés, mais cela pénaliserait les assurés qui n'ont pas eu de sinistres.

Finalement, pour chaque observation, la charge totale hors atypique y compris mutualisation est obtenue ainsi :

$$\text{Charge}_{\text{totale hors atypique yc mutualisation}} = \text{Charge}_{\text{sous-crête}} + \text{Montant à mutualiser}$$

Où $\text{Montant à mutualiser} = \mathbb{1}_{\text{contrat sinistré}} * \frac{\text{prime acquise DDE}}{\sum \text{prime acquise DDE}} * \sum \text{Charge}_{\text{sur-crête}}$

2.3.2 Le nombre de sinistres

Dans le cadre d'une modélisation fréquence x CM, le nombre de sinistres est une variable à expliquer. Compte tenu des évènements exceptionnels qui ne représentent pas la sinistralité réelle du contrat, il a été convenu d'exclure les sinistres atypiques de la modélisation. Le traitement réalisé consiste à identifier les sinistres atypiques par le montant de la charge, dès lors que la charge excède le seuil atypique, il est retiré de la modélisation. Ainsi, la fréquence modélisée sera également hors atypique.

CHAPITRE 3

3 Aspects théoriques

Ce chapitre sera consacré aux aspects théoriques de la modélisation. L'objectif étant de modéliser la prime pure pour la garantie DDE à partir des variables posées aux clients et des zoniers DDE disponibles obtenus par l'intermédiaire de l'adresse. L'inversion du cycle de production en assurance impose à l'assureur de déterminer au préalable la prime pure qui permet de couvrir totalement la charge de sinistre, pour un contrat donné et sur une période d'assurance donnée. Il est primordial que l'écart entre la prime pure et la charge de sinistre soit faible.

3.1 Les modèles linéaires généralisés

3.1.1 Définition

Les modèles linéaires généralisés ou GLM (Generalized Linear Models en anglais) sont une généralisation des modèles de régression linéaire classiques. Ils permettent de prédire ou d'expliquer une variable réponse Y à partir de p variables explicatives regroupées dans un vecteur.

Rappel des régressions linéaires classiques

La régression linéaire cherche à établir une relation linéaire entre une variable aléatoire à expliquer Y, à partir d'un vecteur de variables explicatives $\mathrm{X} = [\mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_\mathrm{p}]$.

Le modèle de régression linéaire est défini par une équation de la forme :

$$\mathrm{Y} = \mathrm{X}\beta + \epsilon$$

Le modèle se réécrit :

$$\mathrm{y_i} = \beta_0 + \beta_1 \mathrm{X_{i,1}} + \cdots + \beta_\mathrm{p} \mathrm{X_{i,p}} + \epsilon_\mathrm{i}, \qquad \mathrm{i} = 1, \dots, \mathrm{n}$$

Où :

- y_i est la variable aléatoire à expliquer,
- $\mathrm{X_{i,1}}, \dots, X_{i,p}$ variables explicatives associées à l'individu i,
- $\beta_0, \beta_1, \dots, \beta_\mathrm{p}$ paramètres inconnus du modèle à estimer,
- ϵ_i sont des termes d'erreur, non observés, indépendants et identiquement distribués, $\mathbb{E}[\epsilon] = 0$ et $\mathrm{V}(\epsilon) = \sigma^2 \mathrm{I_n}$

Les hypothèses concernant le modèle sont :

$$(\mathcal{H}) \begin{cases} (\mathcal{H}_1) : \ \mathrm{rang(X)} = \mathrm{p} \\ (\mathcal{H}_2) : \epsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{I_n}) \text{ où } \mathbb{E}[\epsilon] = 0 \text{ et } \mathrm{V}(\epsilon) = \sigma^2 \mathrm{I_n} \end{cases}$$

L'hypothèse $(\mathcal{H}_2)$ signifie que les erreurs sont centrées, de même variance (homoscédasticité) et non corrélées entre elles.

Modèles linéaires généralisés (GLM)

Le modèle linéaire généralisé sont des méthodes classiques de tarification d'assurances non-vie, il reprend le principe du modèle linéaire classique.

Dans ce mémoire, la variable à expliquer Y sera soit :

- Discrète : Modélisation du nombre de sinistres
- Continue : Modélisation du coût moyen d'un sinistre ou de la prime pure

Trois composantes peuvent être identifiées au sein des GLMs :

1. Une composante aléatoire : caractérisée par la loi de probabilité de la variable à expliquer Y, et en supposant que cette loi appartienne à la famille exponentielle,
2. Un prédicteur linéaire ou composante déterministe : $\eta = \mathrm{X}\beta$,
3. Une relation fonctionnelle : une relation entre la composante aléatoire et la composante déterministe, assuré par une fonction de lien g, où $\mathrm{g}\big(\mu(\mathrm{X})\big) = \mathrm{g}(\mathbb{E}[\mathrm{Y}|\mathrm{X}]) = \mathrm{X}\beta$

Les hypothèses du modèle deviennent alors :

- $Y|\mathrm{X} = \mathrm{x} \sim \mathcal{P}_{\theta,\phi}$ appartient à une famille exponentielle,
  avec θ le paramètre canonique, ϕ le paramètre de dispersion de la famille des lois exponentielles, souvent considéré comme un paramètre de nuisance[7].
- $\mathrm{g}(\mathbb{E}[\mathrm{Y}|\mathrm{X}]) = \mathrm{X}\beta$, avec g une fonction bijective appelée fonction de lien

La détermination des coefficients d'un modèle GLM à p variables explicatives consiste à rechercher les coefficients $[\beta_0, \beta_1, \dots, \beta_\mathrm{p}]$ tels que $\forall i \in \{1, \dots, n\}$,

$$\mathrm{g}(\mathbb{E}[\mathrm{Y_i}|\mathrm{X_i}]) = \beta_0 + \sum_{\mathrm{j}=1}^{\mathrm{p}} \beta_\mathrm{j} \mathrm{X_{i,j}}$$

Ainsi, la différence entre les GLMs et les modèles linéaires classiques est que le modèle linéaire classique modélise la variable à expliquer directement et le GLM modélise une fonction de l'espérance de la variable à expliquer, appelée fonction de lien. Ci-dessous les fonctions de liens les plus courantes :

1. La fonction identité pour une distribution Normale : $g(x) = x$
2. La fonction logarithme pour une distribution Poisson : $g(x) = \log(x)$
3. La fonction logit pour une distribution Bernoulli : $g(x) = \log\left(\frac{x}{1-x}\right)$
4. La fonction inverse pour une distribution Gamma : $g(x) = \frac{1}{x}$

[7] Paramètre de nuisance : paramètre qui n'est pas d'intérêt immédiat mais qui doit être pris en compte dans l'analyse des paramètres d'intérêt

Les résultats de l'ajustement du modèle GLM diffère selon la fonction de lien utilisée. Si la fonction de lien appliquée est la fonction identité alors le résultat sera un **modèle additif** :

$$\mathbb{E}[\mathrm{Y_i}|\mathrm{X_i}] = \beta_0 + \sum_{j=1}^{p} \beta_j \mathrm{X_{i,j}}$$

En revanche, si la fonction de lien est la fonction logarithmique, alors c'est un **modèle multiplicatif** qui est obtenu :

$$\mathrm{Log}(\mathbb{E}[\mathrm{Y_i}|\mathrm{X_i}]) = \beta_0 + \sum_{\mathrm{j}=1}^{\mathrm{p}} \beta_\mathrm{j} \mathrm{X_{i,j}}$$

$$\Leftrightarrow \mathbb{E}[\mathrm{Y_i}|\mathrm{X_i}] = \exp\left(\beta_0 + \sum_{\mathrm{j}}^{\mathrm{p}} \beta_\mathrm{j} \mathrm{X_{i,j}}\right) = \exp(\beta_0) * \exp\left(\beta_1 \mathrm{X_{i,1}}\right) * \dots * \exp\left(\beta_\mathrm{p} \mathrm{X_{i,p}}\right)$$

$$= \exp\left(\mathrm{X_i^t}\beta\right) = \prod_{\mathrm{j}=1}^{\mathrm{p}} \exp\left(\mathrm{X_{i,j}}\beta_\mathrm{j}\right)$$

La fonction logarithme est utilisée afin de bénéficier de cet aspect multiplicatif, ce qui est utile pour arbitrer et calibrer les coefficients. Par ailleurs, en appliquant la fonction exponentielle, les coefficients obtenus sont tous positifs et donc la prime pure obtenue ne sera pas négative.

Estimation des paramètres

Une fois la loi sélectionnée et la fonction de lien choisie, il reste à estimer les paramètres. Dans cette étude, les paramètres sont estimés par maximum de vraisemblance, une méthode implémentée par défaut dans la plupart des logiciels.

Soit $(\mathrm{y}_1, \dots, y_n)$ un échantillon aléatoire de taille n indépendantes et identiquement distribuées, qui suit une loi de la famille exponentielle, de densité notée f :

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

Où :

- θ le paramètre d'intérêt, inconnu,
- ϕ le paramètre de dispersion, souvent considéré comme un paramètre de nuisance, supposé connu,
- $\mathrm{a}(.)$ et $\mathrm{c}(.)$ sont des fonctions dérivables,
- $\mathrm{b}(.)$ de classe C^3 de dérivée première inversible.

La moyenne et la variance sont définies :

$$\mathbb{E}[\mathrm{Y}|\mathrm{X}] = \mu_\mathrm{i} = \mathrm{b}'(\theta) = \frac{\partial \mathrm{b}(\theta)}{\partial \theta} \text{ et } \mathrm{Var}(\mathrm{Y}|\mathrm{X}) = \mathrm{b}''(\theta) \cdot \mathrm{a}(\phi) = \frac{\partial^2 \mathrm{b}(\theta)}{\partial \theta^2}$$

Soit les notations :

$$\begin{cases} \eta_\mathrm{i} = \mathrm{X_i^t}\beta = \mathrm{g}(\mu_\mathrm{i}) \\ \mu_\mathrm{i} = \mathbb{E}[\mathrm{Y_i}|\mathrm{X_i}] = \mathrm{g}^{-1}\left(\mathrm{X_i^t}\beta\right) = \mathrm{g}^{-1}(\eta_\mathrm{i}) \end{cases}$$

Dans le cas des modèles exponentiels, la vraisemblance s'écrit :

$$\mathcal{L}_n(\theta) = \mathcal{L}(\theta_1, \dots, \theta_n, \phi, y_1, \dots, y_n) = \prod_{i=1}^{n} \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right]$$

La log-vraisemblance s'écrit :

$$l_n(\theta) = log\mathcal{L}(\theta_1, \dots, \theta_n, \phi, y_1, \dots, y_n) = \sum_{i=1}^{n} \left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right]$$

Le paramètre θ est une fonction des coefficients $\beta = (\beta_0, \beta_1, \dots, \beta_p)$, ainsi les coefficients du GLM sont estimés en cherchant les $\hat{\beta}$ qui maximisent la vraisemblance. Ainsi, pour déterminer le maximum, il suffit de déterminer la valeur du paramètre de la fonction qui annule la dérivée tout en gardant la dérivée seconde négative :

$$\begin{cases} \frac{\partial\, l_n(\theta)}{\partial \beta_j} = 0 \\ \frac{\partial^2\, l_n(\theta)}{\partial \beta_j^2} < 0 \end{cases}$$

La première équation amène à :

$$\frac{\partial\, l_n(\theta)}{\partial \beta_j} = \frac{\partial\, l_n(\theta)}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j} = 0$$

Avec :

- $\frac{\partial\, l_n(\theta)}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)} = \frac{y_i - \mu_i}{a(\phi)}$
- $\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = \frac{Var(Y_i|X_i)}{a(\phi)}$
- $\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} = (g^{-1})'(\eta_i)$
- $\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial X_i \beta}{\partial \beta_j} = x_{ij}\beta$

Donc les équations de la vraisemblance sont :

$$\frac{\partial\, l_n(\theta)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)\, x_{ij}}{Var(Y_i|X_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0 \;\; pour\; j = 1, \dots, p$$

Ce sont des équations non-linéaires en β dont la résolution requiert des méthodes itératives. L'algorithme de Newton-Raphson est la méthode de résolution itérative la plus courante pour estimer les prédicteurs du GLM.

Après avoir vu les estimations des paramètres, deux modèles seront abordés dans ce mémoire pour modéliser les sinistres :

- Modèle de prime pure : correspondant au modèle individuel, permet de modéliser de la prime pure de façon directe en utilisant une distribution de Tweedie,
- Modèle fréquence x coût moyen : correspondant au modèle collectif, permet de modéliser de manière distincte et indépendante la fréquence et le coût d'un sinistre lorsqu'il y a survenance.

3.1.2 Approche Tweedie

Pour modéliser la prime pure, il est nécessaire de choisir une distribution qui accepte les valeurs nulles. La distribution de Tweedie est un cas particulier des modèles de dispersion exponentielle, souvent utilisée comme distribution pour les modèles linéaires généralisés. Elle est très utile pour modéliser une distribution continues pour les valeurs supérieures à zéro avec une masse concentrée en zéro.

Soit Y une variable aléatoire suivant une distribution de Tweedie. Cette famille de distributions présente les caractéristiques suivantes :

- Une densité d'une loi de la famille exponentielle $f_{\theta,\phi}(y) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}, y \in \mathbb{R}$
- Une moyenne $\mathbb{E}(Y) = \mu$
- Une variance $\mathrm{Var}(Y) = \phi \cdot [\mathbb{E}(Y)]^p$

  Avec :

  - ϕ le paramètre de dispersion,
  - $p \geq 0$ le paramètre de forme de la distribution.

Selon les valeurs de p, des lois usuelles de la famille exponentielle peuvent être retrouvées car ce sont des cas particuliers de la loi de Tweedie :

- La loi Normale, lorsque p = 0,
- La loi de Poisson, lorsque p = 1,
- La loi Poisson-Gamma, lorsque 1 < p < 2,
- La loi Gamma, lorsque p = 2,
- La loi Gaussienne inverse, lorsque p = 3.

Lorsque la valeur de p tend vers 1, le comportement de la distribution de Tweedie se rapproche de celui de la loi de Poisson. Lorsqu'elle tend vers 2, le comportement se rapproche de la loi Gamma. Le cas où 1 < p < 2 implique une loi Poisson composée avec des sauts Gamma, une distribution continue pour Y > 0, avec une masse positive à Y = 0. Ce type de loi est adapté à notre étude car il permet de modéliser directement la distribution de la charge non nulle avec un nombre important de valeurs nulles et de ne pas imposer d'hypothèse d'indépendance entre la fréquence et le coût moyen.

La fonction de lien log est utilisée, elle est adéquate avec la modélisation de Tweedie et permet d'obtenir une formule de type multiplicatif pour le tarif.

3.1.3 Approche fréquence x coût moyen

L'approche fréquence x CM permet de modéliser de manière distincte et indépendante la fréquence et le coût d'un sinistre lorsqu'il y a survenance. Cette approche présente cependant un décalage entre les deux notions : le coût moyen n'est connu qu'au terme du développement final du sinistre alors que la fréquence est connue dès la déclaration. Ce modèle distingue donc la modélisation de la fréquence des sinistres par assuré (modèle de comptage) et celle de la sévérité des sinistres quel que soit l'assuré (modèle de montant).

Nous allons montrer que les approches fréquence x CM et prime pure sont équivalentes. Pour cela, nous allons supposer l'indépendance des deux variables.

Soit S la charge totale de sinistres. Elle peut se décomposer de la manière suivante :

$$\mathrm{S} = \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{X_i}$$

Où $X_1, X_2 \dots, X_N$ les montants aléatoires des sinistres et N_i le nombre aléatoire de sinistres total avec les hypothèses suivantes :

- Les X_i variables aléatoires indépendantes et identiquement distribuées (iid),
- X_i indépendante de N_i pour tout i.

En passant à l'espérance puis en appliquant la formule des espérances conditionnelles totales :

$$\begin{aligned}
\mathbb{E}[\mathrm{S}] &= \sum_{\mathrm{k}=0}^{+\infty} \mathbb{E}[\mathrm{S}|\mathrm{N}=\mathrm{k}] \times \mathbb{P}(\mathrm{N}=\mathrm{k}) \\
&= \sum_{\mathrm{k}=0}^{+\infty} \mathbb{E}\left[\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{X_i} \,\middle|\, \mathrm{N}=\mathrm{k}\right] \times \mathbb{P}(\mathrm{N}=\mathrm{k}) \\
&= \sum_{\mathrm{k}=0}^{+\infty} \mathbb{E}\left[\sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{X_i} \,\middle|\, \mathrm{N}=\mathrm{k}\right] \times \mathbb{P}(\mathrm{N}=\mathrm{k}) \\
&= \sum_{\mathrm{k}=0}^{+\infty} \mathbb{E}\left[\sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{X_i}\right] \times \mathbb{P}(\mathrm{N}=\mathrm{k}) \text{ car les } \mathrm{X_i} \text{ sont indépendants des } \mathrm{N_i} \\
&= \sum_{\mathrm{k}=0}^{+\infty} \sum_{\mathrm{i}=0}^{\mathrm{k}} \mathbb{E}[\mathrm{X}] \times \mathbb{P}(\mathrm{N}=\mathrm{k}) \text{ car les } \mathrm{X_i} \text{ sont iid} \\
&= \sum_{\mathrm{k}=0}^{+\infty} \mathrm{k}\, \mathbb{E}[\mathrm{X}] \times \mathbb{P}(\mathrm{N}=\mathrm{k}) \\
&= \mathbb{E}[\mathrm{X}] \times \sum_{\mathrm{k}=0}^{+\infty} \mathrm{k}\, \mathbb{P}(\mathrm{N}=\mathrm{k}) \\
&= \mathbb{E}[\mathrm{X}] \times \mathbb{E}[\mathrm{N}]
\end{aligned}$$

⟺ Prime pure = Fréquence × Coût moyen

L'équivalence des deux approches est ainsi démontrée sous la condition de l'indépendance entre les variables nombre de sinistres N_i et coût des sinistres X_i. Pour pouvoir utiliser l'approche fréquence-CM, il est donc nécessaire de vérifier l'indépendance entre ces deux termes. Pour évaluer le degré de dépendance de ces deux variables quantitatives, plusieurs mesures de dépendance peuvent être calculées tels que le coefficient de Pearson, le coefficient de Spearman et le tau de Kendall.

1. Coefficient de Pearson :

Le coefficient de Pearson permet de mesurer une dépendance linéaire entre deux variables quantitatives. Sa formule est la suivante :

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

Avec :

- $Cov(X,Y) = \mathbb{E}[(\mathrm{X} - \mathbb{E}[\mathrm{X}]) \cdot (Y - \mathbb{E}[\mathrm{Y}])]$
- $r_x = \sqrt{\mathbb{E}[(\mathrm{X} - \mathbb{E}[\mathrm{X}])^2]}$
- $r_y = \sqrt{\mathbb{E}[(\mathrm{Y} - \mathbb{E}[\mathrm{Y}])^2]}$

Ce coefficient varie entre -1 et 1. Son interprétation dépend de la valeur du coefficient :

- Si $r_{xy} < 0$ alors la relation entre les deux variables est linéaire et positive, c'est-à-dire qu'une variable augmente, l'autre diminue,
- Si $r_{xy} = 0$ ne signifie pas toujours que les variables sont indépendantes car l'indépendance implique que $r_{xy} = 0$. Ainsi, Si X et Y sont indépendantes alors $r_{xy} = 0$. Mais la réciproque est en générale fausse, et n'est vérifiée que si le couple (X,Y) suit une loi normale bivariée.
- Si $r_{xy} > 0$ alors la relation entre les deux variables est linéaire et négative, c'est-à-dire que les deux variables varient dans le même sens et linéairement.

2. Coefficient de Spearman :

Le coefficient de Spearman est un cas particulier du coefficient de Pearson, calculé à partir des rangs des données plutôt que leur valeur observée directement. Les données sont triées par ordre croissant et les valeurs sont remplacés par leurs rangs. Ce coefficient, appelé ρ (rhô) est calculé de la façon suivante :

$$\rho_{rg(X),rg(Y)} = \frac{Cov\big(rg(X), rg(Y)\big)}{\sigma_{rg(X)} \sigma_{rg(Y)}}$$

Avec :

- $rg(X)$ et $rg(Y)$ respectivement la variable de rang de X et Y,
- $Cov\big(rg(X), rg(Y)\big)$ covariance des variables de rang,
- $\sigma_{rg(X)}\ et\ \sigma_{rg(Y)}$ écarts-types des variables de rang.

Ce coefficient permet d'évaluer la relation monotone entre deux variables continues ou ordinales. Lorsque la tendance est affine, il se comporte de façon similaire au coefficient de Pearson. Plus la tendance monotone est marquée, plus la valeur du coefficient est proche de 1 ou -1. Ainsi, de façon similaire au coefficient de Pearson, le coefficient de Spearman aura une valeur positive lorsque la tendance est croissante, négative lorsqu'elle est décroissante et lorsqu'elle est non monotone, le coefficient sera proche de 0.

3. Tau de Kendall :

Le τ (tau) de Kendall mesure également la corrélation de rang entre deux variables.

Soit $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ un ensemble d'observation des variables jointes X et Y tel que les valeurs des (x_i) et (y_i) sont uniques. Les paires d'observations (x_i, y_i) et (x_j, y_j) sont dites concordantes si $x_i < x_j$ et $y_i < y_j$ ou si $x_i > x_j$ et $y_i > y_j$. Elles sont dites discordantes si $x_i < x_j$ et $y_i > y_j$ ou si $x_i > x_j$ et $y_i < y_j$. Lorsque $x_i = x_j$ ou $y_i = y_j$, la paire n'est ni concordante ni discordante. Ce coefficient est alors défini comme :

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

Avec :

- n_c le nombre de paires concordantes,
- n_d le nombre de paires discordantes,
- n le nombre total de paires.

Ce coefficient est compris entre -1 et 1, et s'interprète de la même façon que les deux autres coefficients de corrélation.

Après avoir vu les différents coefficients de corrélation pour évaluer le degré de dépendance, nous allons par la suite définir les variables fréquence et coût.

La fréquence

La fréquence correspond à la modélisation d'une variable de comptage. Plusieurs lois permettent de modéliser ce type de variable, par exemple la loi de Poisson ou la loi Binomiale Négative. La distribution de Poisson est souvent utilisée et sera retenue pour modéliser la fréquence des sinistres. La fonction de lien logarithmique pour une distribution de Poisson est appliquée.

La densité de la loi de Poisson de paramètre λ est :

$$\mathrm{f}_\lambda(\mathrm{y}) = \exp(-\lambda) \cdot \frac{\lambda^y}{y!} = \exp\left\{\frac{\mathrm{y}\log(\lambda) - \lambda}{1} + \log\left(\frac{1}{y!}\right)\right\}, y \in \mathbb{N}$$

Où on reconnaît les fonctions et paramètres de la forme générale de la densité de la famille exponentielle :

$$\begin{cases} \theta = \log(\lambda) \\ a(\phi) = 1 \\ b(\theta) = \lambda \\ c(y, \phi) = \log\left(\frac{1}{y!}\right) = -\log(y!) \end{cases}$$

La moyenne et la variance obtenue correspondent bien à une loi de Poisson :

$$\begin{cases} \mathbb{E}[\mathrm{Y}] = b'(\theta) = \lambda \\ Var[Y] = b''(\theta)\, a(\phi) = \lambda \end{cases}$$

Le coût moyen

Le coût moyen correspond à la modélisation d'une variable strictement positive. Pour le modéliser, les lois usuelles sont la loi Gamma et la loi log-Normale. La loi Gamma sera retenue pour modéliser le coût moyen des sinistres.

La loi Gamma de paramètres α et β présente la fonction de densité suivante :

$$f_{\alpha,\beta}(y) = \frac{1}{\Gamma(\alpha)} \left(\frac{\alpha}{\beta}\right)^{\alpha} y^{\alpha-1} \exp\left(-\frac{\alpha}{\beta}\, y\right) \; avec \; y > 0$$

$$= \exp\left\{ \frac{\mathrm{y}\dfrac{-1}{\beta} - \left(-\log\left(\dfrac{1}{\beta}\right)\right)}{\dfrac{1}{\alpha}} + \left[\alpha\log(\alpha y) - \log(y) - \log\big(\Gamma(\alpha)\big)\right] \right\}$$

Où on reconnaît les fonctions et paramètres de la forme générale de la densité de la famille exponentielle :

$$\begin{cases} \theta = -\frac{1}{\beta} \\ a(\phi) = -\frac{1}{\alpha} \\ b(\theta) = -\log(-\theta) = -\log\left(\frac{1}{\beta}\right) \\ c(y,\phi) \;\; = \alpha\log(\alpha y) - \log(y) - \log\big(\Gamma(\alpha)\big) \end{cases}$$

3.2 Critères de sélection des modèles

Après avoir vu les deux approches pour calibrer la prime pure, il reste à déterminer les variables à conserver dans le modèle. La modélisation GLM est basée sur la maximisation de la vraisemblance. Cette vraisemblance est minimale dans un modèle ne comportant qu'une constante et maximale dans le modèle saturé. Ainsi, la vraisemblance permet d'apprécier la qualité d'un modèle. Par la suite, trois parties seront traitées : une première partie décrivant le paramétrage des modèles, une deuxième expliquant la sélection des variables et une troisième partie détaillant les indicateurs de qualité de modèles.

3.2.1 Paramétrage des modèles

Avant de passer à la modélisation Tweedie ou fréquence x CM, il faut au préalable définir la ou les variable(s) à expliquer associée(s) à une exposition au risque, puis sélectionner les variables explicatives. Nous allons voir dans la section 3.2.2 Sélection de variables, l'outil utilisé permet de sélectionner les variables les plus pertinentes dans le modèle de manière automatisée, par une généralisation de l'approche forward et des critères de spread et corrélation. Dans le cadre d'une modélisation « Prime Pure » avec l'approche Tweedie, la charge est la variable à expliquer, et le paramètre p est à définir. Tandis que, dans le cadre d'un modèle fréquence * CM, le nombre et la charge de sinistres sont les variables à expliquer.

La variable année de vision est également prise en compte dans la modélisation afin de vérifier la stabilité du modèle dans le temps.

Par ailleurs, dans le paramétrage il est possible d'imposer des contraintes en amont du calibrage des modèles, en spécifiant la croissance des coefficients pour les variables ordinales telles que les tranches de surface, de contenu incendie, de chiffre d'affaires et les antécédents de sinistres.

Ensuite, pour calibrer les modèles, nous avons eu recours à la *validation croisée* ou *cross-validation* en anglais. C'est une méthode statistique qui permet d'évaluer la capacité de généralisation d'un modèle et de s'assurer qu'il n'y ait pas de sur-apprentissage dans les modélisations. Il est important d'évaluer la robustesse d'un modèle et la capacité à généraliser sur de nouvelles données. Pour ce faire, la base de données initiale est divisée en deux parties : une partie « base d'apprentissage » (80%) permettant de calibrer les modèles et l'autre partie « base de validation » (20%) qui ne sera pas utilisée pour l'ajustement des modèles mais permettra de mesurer la performance des prédictions sur des données qui ne sont pas dans la base d'apprentissage, qui n'ont donc pas servi à construire les modèles.

Par ailleurs, pour réduire au maximum le risque de sur-apprentissage, le mécanisme *k-fold cross-validation* est utilisé. Le mécanisme de cette procédure est le suivant :

- La base initiale est divisée en base d'apprentissage (80%) et base de validation (20%),
- La base d'apprentissage est elle-même découpée en k échantillons (folds) de manière aléatoire,
- Le modèle est entrainé et testé pendant k itérations. A chaque itération, il est entrainé sur k-1 folds, appelés « échantillons de train », et est testé sur le fold restant, appelé « échantillon de test ».

Pour cette étude, le processus de 4-fold a été mis en œuvre.

Ci-dessous le tableau montrant le mécanisme de 4-fold cross-validation :

	Base initiale (100%)				
	Base d'apprentissage (80%)				Base de validation (20%)
	Fold 1 (20%)	Fold 2 (20%)	Fold 3 (20%)	Fold 4 (20%)	
Itération 1	Test	Train	Train	Train	
Itération 2	Train	Test	Train	Train	
Itération 3	Train	Train	Test	Train	
Itération 4	Train	Train	Train	Test	
Validation finale					Base de validation

Tableau 3.2.1 - Mécanisme de 4-fold cross-validation

Base d'apprentissage :

La **base d'apprentissage** correspond à 80% de l'échantillon de données. Pour chaque itération, cette base est elle-même découpée en 4 échantillons (folds) de manière aléatoire, avec trois **échantillons de train** (60%) sur lequel le modèle est ajusté et un **échantillon de test** (20%), utilisé pour évaluer le modèle ajusté.

Base de validation :

La **base de validation** correspond à 20% de l'échantillon de données. Elle est utilisée pour évaluer un modèle final et elle est adaptée à l'ensemble des données de la base d'entraînement. Cela permet d'avoir une évaluation finale et fiable des modèles puisque les modèles seront entraînés sur une base indépendante. De plus, les modèles seront comparés de manière non biaisée.

Ainsi, ce mécanisme permet de créer quatre modèles différents avec quatre partitionnements distincts de l'échantillon total de données. Le choix de fixer k à quatre permet d'avoir des résultats robustes tout en étant confronté à un temps de calcul raisonnable.

Lors de la première itération, le premier fold sert d'échantillon de test, pendant qu'un modèle est entrainé sur le reste des folds (échantillon de train). Lors de la seconde itération, le modèle est entrainé cette fois-ci sur les données des folds 1, 3 et 4, le fold 2 servant d'échantillon de test, et ainsi de suite pour les itérations suivantes. Ainsi, chaque observation, aura servi au moins une fois d'échantillon d'entrainement et de test. Les performances moyennes sur les différentes itérations de modélisation sont ensuite retenues. Il sera possible d'évaluer le degré de généralisation et la stabilité du modèle en vérifiant que les indicateurs de qualité du modèle (Gini, RMSE) soient proches entre les différents échantillons de test utilisés.

3.2.2 Sélection de variables

Afin d'obtenir une combinaison optimale des variables explicatives, des techniques de sélection de variables sont utilisées. Ce sont donc des procédures itératives qui réalisent des comparaisons successives de variations du modèle initial. Ce processus permet de sélectionner les variables pertinentes pour la tarification de la garantie DDE parmi les six variables posées aux clients.

Critère de pénalisation

Des critères adaptés aux GLM, prennent en compte la pénalisation des modèles. Les critères les plus couramment utilisés sont l'Akaïke Information Criterion (AIC) et le Bayesian Information Criterion (BIC). Ce sont des critères qui pénalisent la vraisemblance du modèle par le nombre de paramètres, afin de privilégier des modèles parcimonieux.

Les deux critères sont définis par :

$$\begin{cases} AIC = -2l_n + 2p \\ BIC = -2l_n + p \cdot log(n) \end{cases}$$

Où :

- l_n est la log-vraisemblance du modèle
- p le nombre de paramètres
- n le nombre d'observations

En considérant un ensemble de modèles candidats, le modèle sélectionné est celui qui minimise ces quantités. Ces critères reposent donc sur un compromis entre la qualité de l'ajustement et la complexité du modèle, en pénalisant les modèles ayant un grand nombre de paramètres, ce qui limite

le sur-paramétrage ou le sur-ajustement. Il s'agit donc de trouver un modèle de manière parcimonieuse qui décrit les données avec le moins de paramètres possibles.

D'après la définition des deux critères de pénalisations, plus la taille de l'échantillon n augmente, plus le BIC pénalisera par rapport à l'AIC. Ainsi, la pénalisation ne dépendra pas seulement du nombre de paramètres mais aussi de la taille de l'échantillon.

Cependant, ces deux critères ne sont pas retenus car la pénalisation, avec l'outil utilisé dans ce mémoire, ne se base pas exactement sur les mêmes informations. Elle se base tout de même sur ce type de structure, mais le détail de celle-ci ne sera pas présenté pour des raisons de confidentialité. Par ailleurs, cet outil permet de déterminer de manière parcimonieuse le nombre de variables qui serait le plus adapté au modèle par des représentations graphiques des évolutions de différents critères de performance selon le nombre de variables explicatives injectées dans le modèle.

Procédure de sélection de variables

Après avoir défini le critère de pénalisation, la procédure de sélection de variables sera développée ci-dessous. Deux méthodes peuvent être appliquées :

- Sélection exhaustive : En théorie, pour trouver le meilleur modèle parcimonieux, il faudrait réaliser une recherche exhaustive, c'est-à-dire tester tous les modèles possibles et retenir celui qui minimise le critère de validation. Cependant, le nombre de modèles possibles s'élève à $\sum_{k=0}^{p} \frac{p!}{k!(p-k)!} = 2^p$. Ce nombre croit très rapidement avec le nombre de variables p à disposition. Lorsque p est grand, cette méthode peut alors s'avérer trop coûteuse en temps et en capacité de calcul pour construire tous les modèles possibles. Cette méthode ne sera pas utilisée pour la procédure de sélection de variables.

- Sélection pas à pas : En pratique, la méthode de sélection pas à pas est utilisé pour réduire le nombre de modèles à tester tout en sélectionnant celui qui minimise le critère. Il permet ainsi de sélectionner les variables les plus pertinentes et limiter le sur-paramétrage. Il existe trois types d'approche pour cette méthode : la sélection ascendante (forward), la sélection descendante (backward) et la sélection mixte (stepwise).

- La sélection ascendante (forward) : Approche où le premier modèle testé ne contient que la constante puis les autres variables sont ajoutées au fur et à mesure des résultats au test de significativité (par exemple, test de Student ou test de Wald qui sont asymptotiquement équivalents). Cela revient à sélectionner la variable la plus pertinente à chaque itération, et en arrêtant le processus dès que la performance du modèle n'augmente plus. Cette approche réduit donc sensiblement la complexité du modèle et permet une simplification de son interprétation. Cependant, elle a un inconvénient majeur : une fois la variable introduite dans le modèle, elle ne peut plus être éliminée.
- La sélection descendante (backward) : Approche qui est le procédé inverse de l'approche forward. Cette version consiste à considérer tout d'abord un modèle avec toutes les variables explicatives possibles, puis éliminer au fur et à mesure la variable la moins significative à partir des tests de significativité des variables. Cette sélection est plus économique en termes de temps et d'interprétation mais présente toutefois un inconvénient majeur : l'ordre des variables testées joue un rôle important car une fois qu'une variable est supprimée, il n'est plus possible de la réintroduire dans le modèle.
- La sélection mixte (stepwise) : Approche qui combine les deux approches précédentes. A chaque itération, une variable est ajoutée dans le modèle et on revérifie que toutes les variables

introduites précédemment dans le modèle restent significatives. Après la vérification, si des variables ne sont plus significatives, alors la moins significative des variables est retirée. Le processus continue jusqu'à ce que plus aucune variable ne puisse être ajoutée ou retirée du modèle. En effet, une variable considérée comme la plus significative à une itération, peut à une étape ultérieure devenir non significative. Cette procédure permet de palier au problème de corrélations entre les variables introduites dans le modèle.

L'outil utilisé combine la procédure stepwise et la recherche exhaustive. L'idée est d'augmenter le nombre de modèles testés à chaque itération. Ce paramétrage du nombre est spécifié au tout début dans l'outil lors du paramétrage de la modélisation. Il s'est avéré suffisant de prendre un nombre égal à cinq afin de comparer la performance de modèle avec un nombre de variables différent.

3.2.3 Indicateurs de qualité du modèle

Plusieurs indicateurs sont utilisés pour évaluer la qualité du modèle : courbe de Lorenz, Gini, courbe lift, analyse des résidus, RMSE. Le but est de construire plusieurs modèles et de comparer les indicateurs des différents modèles pour choisir le modèle le plus performant. En effet, la valeur des indicateurs seule n'a pas de sens, seule la comparaison permet de juger de la qualité du modèle. Les différents indicateurs graphiques et numériques vont être présentés ci-dessous. Ils sont construits sur les échantillons tests de la validation croisée, de sorte qu'ils reflètent les performances hors échantillon d'apprentissage du modèle. Par ailleurs, l'ordre d'importance des variables retenues dans les modèles seront également analysés via l'indicateur spread.

Indicateurs graphiques :

1. La courbe de Lorenz

La courbe de Lorenz est une représentation graphique, généralement utilisée pour déterminer le niveau de répartition des richesses au sein d'une population. Néanmoins, elle peut être intéressante dans notre contexte de tarification. Cette courbe se rapproche de la courbe ROC (Receiver Operating Characteristic), utilisée comme mesure de performance dans les régressions logistiques et représentant le Taux de Vrais Positifs (TVP) ou sensibilité en fonction du Taux de Vrais Négatifs (TVN) ou spécificité. Dans un modèle de prime pure, l'axe des ordonnées représente la part cumulée des primes pures et l'axe des abscisses représente l'exposition cumulée (la part cumulée d'assurés triés dans l'ordre croissant des prédictions). La courbe de Lorenz est figurée en bleue. La bissectrice, représentée en verte, indique la situation dans laquelle il y aurait une distribution parfaitement égalitaire de la prime pure dans l'ensemble du portefeuille. Les distributions changent en fonction du modèle choisi : prime pure, fréquence ou coût moyen.

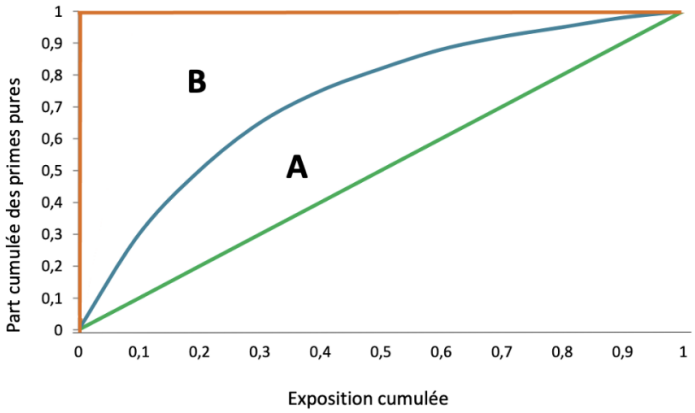


Figure 3.2.1 - Illustration de la courbe de Lorenz

Cette courbe permet de calculer l'indice de Gini, qui se définit comme suit :

$$\text{Indice de Gini} = \frac{\text{Aire A}}{\text{Aire A} + \text{Aire B}}$$

Où :

- A est l'aire entre la courbe de Lorenz et la bissectrice
- B est l'aire au-dessus de la courbe de Lorenz

2. La courbe Lift

La courbe Lift est un indicateur visuel permettant d'apprécier la qualité de prédiction des GLMs. Elle est construite en triant les prédictions par ordre croissant puis en les répartissant en vingt groupes qui représentent chacun 5% des prédictions (quantile de 5% d'exposition). La prédiction moyenne (prime pure) et l'observation moyenne (charge observée) sont représentées par quantile afin de visualiser la superposition et la tendance des deux courbes. La représentation des valeurs prédites est croissante, mais ce n'est pas une certitude pour les valeurs observées. La superposition des deux courbes correspond à un modèle parfait puisque l'observé et le prédit sont égaux en moyenne. Ainsi, cela permet d'évaluer la qualité des prédictions du modèle et en déduire si les deux valeurs sont égales en moyenne. Des barres d'erreur sont également affichées pour visualiser les écarts moyens par quantile.

3. Les résidus

Les résidus, bruit du modèle ou erreurs observées, correspond à la part non expliquée par le modèle. Ils sont communément définis comme suit :

$$e_i = y_i - \widehat{y_i}$$

L'analyse des résidus a pour objectif de valider le GLM.

Les méthodes d'analyse des résidus sont principalement des méthodes d'analyse graphique.

L'outil propose deux types de résidus : les résidus de déviance et les résidus quantiles.

- Les résidus de déviance sont calculés comme suit :

$$\mathrm{Res}_{\mathrm{Deviance}} = \mathrm{sign}(\mathrm{y} - \hat{\mathrm{y}}) * \sqrt{\mathrm{Deviance}(\hat{y},\ \mathrm{y})}$$

Où la déviance est celle liée à la distribution de Poisson, Gamma, …

$$\mathrm{Deviance}_{\mathrm{Poisson}} = 2\sum\nolimits_i \left[\mathrm{y_i}\log\left(\frac{\mathrm{y_i}}{\hat{\mathrm{y}}_\mathrm{i}}\right) - (\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i})\right]$$

$$Deviance_{Gamma} = 2\sum\nolimits_i \left[-\log\left(\frac{y_i}{\hat{y}_i}\right) + \frac{y_i - \hat{y}_i}{\hat{y}_i}\right]$$

L'interprétation graphique de ces résidus est complexe, car il n'y a aucune garantie théorique à la forme à laquelle les résidus de déviance doivent ressembler pour un modèle bien ajusté.

- Les résidus quantiles sont une alternative aux résidus de déviance. Ils sont calculés en transformant la distribution du modèle en une distribution normale. Ainsi, un modèle bien ajusté devrait présenter des résidus quantiles qui sont à peu près normalement distribués. En principe pour une distribution continue, ils sont calculés comme :

$$\mathrm{Res}_{\mathrm{quantile}} = \phi^{-1}(\mathcal{F}(\mathrm{y};\hat{\mathrm{y}}))$$

Où $\mathcal{F}$ est la fonction de répartition de la distribution choisie pour la modélisation, et ϕ est la fonction de répartition de la distribution normale standard.

Pour un modèle bien ajusté, globalement les résidus ont une forme symétrique, centrés autour de zéro et la plupart des résidus sont entre -2 et 2. Ainsi, les hypothèses de résidus seront vérifiées et le modèle validé.

Pour les deux cas, les résidus peuvent être représentés en fonction des prédictions pour vérifier qu'il n'y ait pas de structure particulière qui montrerait un biais dans la modélisation. Le graphique des résidus est en trois dimensions (x = prédiction; y = résidu; z = nombre d'observations). La troisième dimension sera visible sous forme de nuance de couleur (la couleur rouge correspond à un volume d'observations important tandis que le bleu indique un volume d'observation relativement faible).

<u>Indicateurs numériques :</u>

1. L'indice de Gini

L'indice de Gini, calculé à partir de la courbe de Lorenz, mesure l'inégalité d'une variable dans une population donnée. Dans notre cas, il permet d'évaluer la qualité de segmentation et donc la performance du modèle. Pour cela, il faut regarder la répartition des primes sur le portefeuille. La formule obtenue à partir de la courbe de Lorenz est la suivante :

$$\text{Indice de Gini} = \frac{\text{Aire A}}{\text{Aire A} + \text{Aire B}}$$

L'indice de Gini peut également être calculé à partir de la courbe ROC.

Soit Y_i les observations prenant les valeurs 0 (cas négatifs) ou 1 (cas positifs).

Le modèle prédit $\pi_i = \mathbb{P}(\mathrm{Y_i} = 1) = 1 - \mathbb{P}(\mathrm{Y_i} = 0)$. En se fixant un seuil s, si $\pi_i \leq s$, alors $\widehat{Y_i} = 0$, et si $\pi_i \geq s$, alors $\widehat{Y_i} = 1$. Le modèle aura un bon ajustement si les cas positifs sont prédits positifs et les cas négatifs sont prédits négatifs.

Le choix du seuil s permet de minimiser soit les faux positifs, soit les faux négatifs. Ainsi une matrice de confusion est obtenue :

	$Y_i = 1$	$Y_i = 0$
$\widehat{Y_i} = 1$	VP	FP
$\widehat{Y_i} = 0$	FN	VN

Tableau 3.2.2 - Matrice de confusion de la courbe ROC

Avec VP (Vrais Positifs), VN (Vrais Négatifs), FN (Faux Négatifs), FP (Faux Positifs)

Cette matrice de confusion permet de définir :

- La **sensibilité** ou TVP (Taux de Vrais Positifs) = $\frac{\mathrm{VP}}{\mathrm{VP+FN}}$
- La **spécificité** ou TVN (Taux de Vrais Négatifs = $\frac{\mathrm{VN}}{\mathrm{VN+FP}}$

L'AUC (Area Under the Curve) est l'aire sous la courbe ROC qui indique la probabilité que le modèle prédit un positif au lieu d'un négatif. Dans le meilleur des cas, l'AUC est égal à 1. En revanche, une AUC de 0,5 indique que le modèle n'apporte rien.

Ainsi l'indice de Gini peut être défini de la manière suivante avec l'AUC :

$$\text{Indice de Gini} = 2 \times \text{AUC} - 1$$

L'indice de Gini est compris entre 0 et 1. Une tarification uniforme possède un indice de Gini de 0 étant donné que la courbe de Lorenz correspondrait à la bissectrice. Ainsi, plus l'indice est proche de 1, plus le tarif est segmenté et donc plus le modèle est performant.

2. Le RMSE (Root Mean Square Error)

Le Root Mean Square Error est la racine du carré de l'erreur quadratique moyenne, il permet d'évaluer l'erreur du modèle.

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

$$\mathrm{MSE} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{y_i} - \widehat{\mathrm{y_i}})^2$$

Ainsi, le RMSE est donné par :

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{y_i} - \widehat{\mathrm{y_i}})^2}$$

Le RMSE est difficile à interpréter seul, mais permet de comparer les erreurs entre les modèles construits, par conséquent le meilleur modèle sera celui qui minimise ce critère.

Pour ce mémoire, le RMSE pondéré par l'exposition ω_i a été choisi pour prendre en compte le poids de chaque observation :

$$\mathrm{RMSE_{pondéré}} = \sqrt{\frac{1}{\sum_{\mathrm{i}=1}^{\mathrm{n}} \omega_\mathrm{i}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \omega_i * (\mathrm{y_i} - \widehat{\mathrm{y_i}})^2}$$

Importance des variables :

Le spread

Chaque variable retenue dans le modèle possède un indicateur de spread. Cet indicateur permet d'apprécier le degré d'importance des variables à travers la dispersion des coefficients calibrés. Pour cette étude, un graphique représentant l'importance des variables à travers la valeur du spread permettra de visualiser quelles variables seront incluses dans le modèle, ainsi que leur importance par rapport à leur impact sur les prédictions.

Pour les modèles multiplicatifs, le spread 100/0% d'une variable est défini comme :

$$\mathrm{Spread_{100/0\%}} = \frac{\text{Coefficient maximum} + 1}{\text{Coefficient minimum} + 1} - 1$$

Le spread 100/0% prend en compte la dispersion sur tout le portefeuille.

Afin d'éviter les résultats biaisés par les extrêmes, il est possible de calculer le spread 95/5%, où 5% de l'exposition la plus risquée (coefficients les plus élevés) et 5% de l'exposition la moins risquée (coefficients les plus faibles) sont supprimés.

CHAPITRE 4

4 Modélisation de la prime pure

Pour modéliser la prime pure, l'approche Tweedie et fréquence x CM ont été considérées avec deux méthodes d'implémentation de zonier différentes. Ces GLM ont été calibrés sur la base d'apprentissage, et seule la méthode par ajout simple de zonier qui présentait les meilleures performances a été retenue et détaillée dans cette section. Une comparaison de la prime pure calibrée par les deux approches sera effectuée, ainsi que l'impact du nombre de variables dans le modèle et le degré de lissage. Les autres modèles avec la méthode off set d'implémentation du zonier seront ajoutés en annexes (Cf annexe B).

4.1 Modélisation Tweedie

Les résultats de la modélisation Tweedie sur la base d'apprentissage seront détaillés ci-dessous. La variable à prédire est donc la prime pure observée définie par :

$$\text{prime pure observée} = \frac{\text{charge des sinistres}}{\text{exposition}}$$

Sélection de variables

Comme vu dans la section 3.2.2 Sélection de variables, cette étape implique de choisir un modèle de manière parcimonieuse en gardant uniquement les variables les plus pertinentes.

Les différents modèles proposés en fonction du nombre de variables retenues sont représentés ci-dessous :

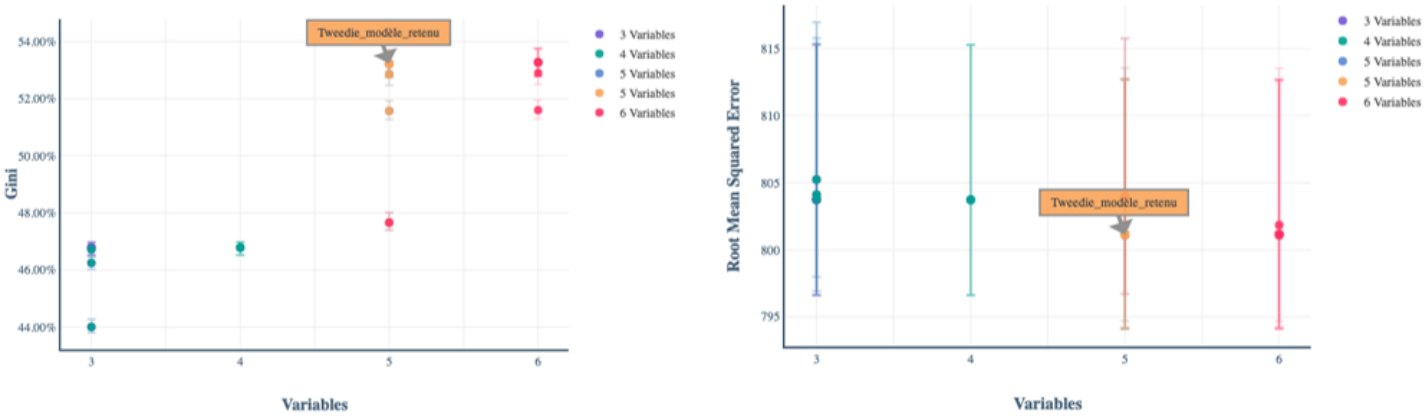


Figure 4.1.1 – Sélection du nombre de variables pour le modèle Tweedie (méthode 1 zonier)

Chaque point correspond à un modèle retenu selon les critères de sélection de variables présentées précédemment. L'axe des abscisses représente le nombre de variables explicatives retenues pour les modèles, où le maximum de variables, égal à six, correspond à toutes les questions posées au client. Pour chaque nombre de variables, cinq modèles plus ou moins lissés sont proposés. Il y a donc plusieurs modèles possibles pour un nombre de variable fixé. Le paramètre de lissage évalue la sensibilité aux signaux faibles. Un modèle lisse suivra principalement les grandes tendances des données mais ne capturera pas toutes les variations, alors qu'un modèle non lisse peut capturer des variations plus subtiles, parfois au détriment de la robustesse. Selon l'indice de Gini (graphique gauche), le modèle retenu est celui qui possède un Gini le plus élevé. Ici, le modèle à cinq variables explique tout aussi bien que le modèle à six variables. Le pouvoir explicatif du modèle n'évolue pas avec l'ajout d'une sixième variable. Selon l'indicateur de RMSE (graphique droite), le modèle retenu est celui qui minimise ce critère. Il est quasiment équivalent entre un modèle à cinq variables et six variables. Finalement, le modèle retenu est celui qui possède cinq variables.

Qualité du modèle

Pour appréhender la qualité du modèle retenu à cinq variables, plusieurs analyses graphiques seront effectuées : le spread, la courbe de Lorenz, la courbe de Lift, les résidus.

Le spread :

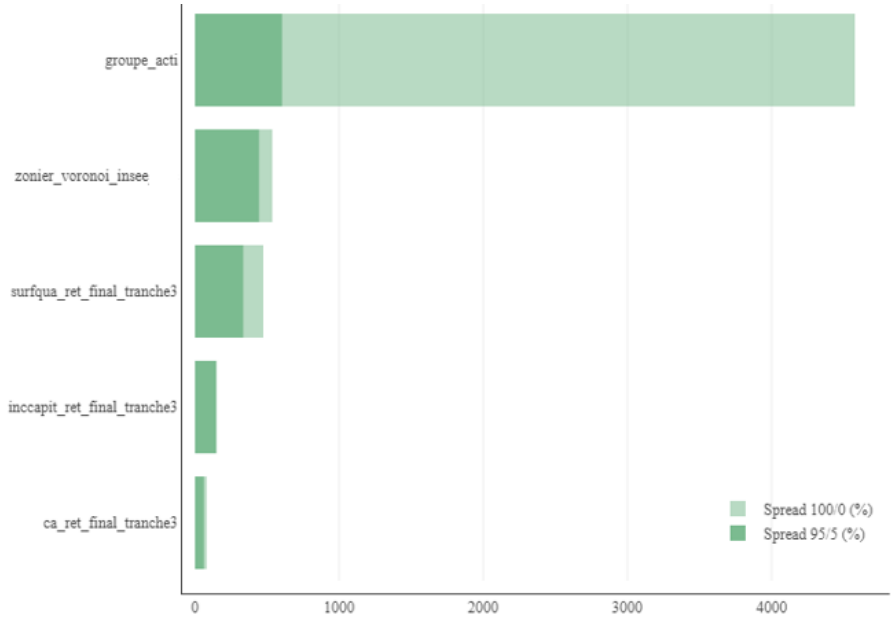


Figure 4.1.2 - Spread du modèle Tweedie méthode 1 zonier

Ce graphique représente l'importance des variables sélectionnées sur les prédictions. Les variables les plus pertinentes et discriminantes par ordre croissant sont le groupe d'activité, le zonier, la surface, le contenu incendie et le chiffre d'affaires.

La courbe de Lorenz :

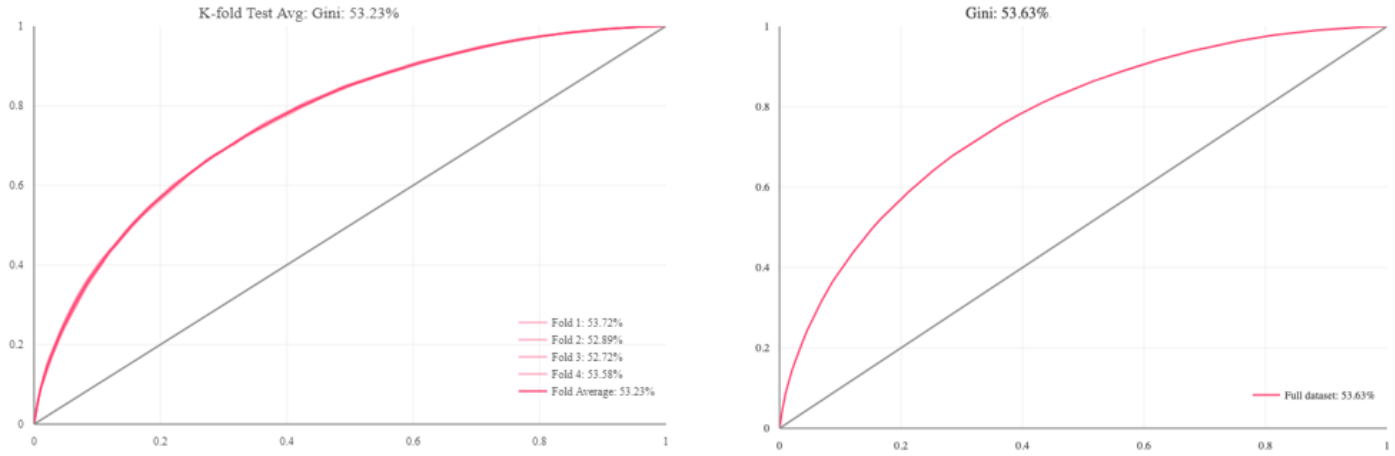


Figure 4.1.3 - Courbes de Lorenz du modèle Tweedie méthode 1 zonier

Le graphique à gauche montre la 4-fold cross-validation pour le nouveau tarif rapide, modélisé par l'approche Tweedie. Il donne des résultats concluants puisque les Ginis sont stables au cours des quatre cross-validation. La courbe de Lorenz est une mesure de segmentation du tarif. L'indice de Gini moyen sur les 4-fold est de 53,23% traduisant une segmentation favorable de notre portefeuille par le modèle. Le graphique à droite affiche un indice de 53,63% sur toute la base de données.

La courbe Lift :

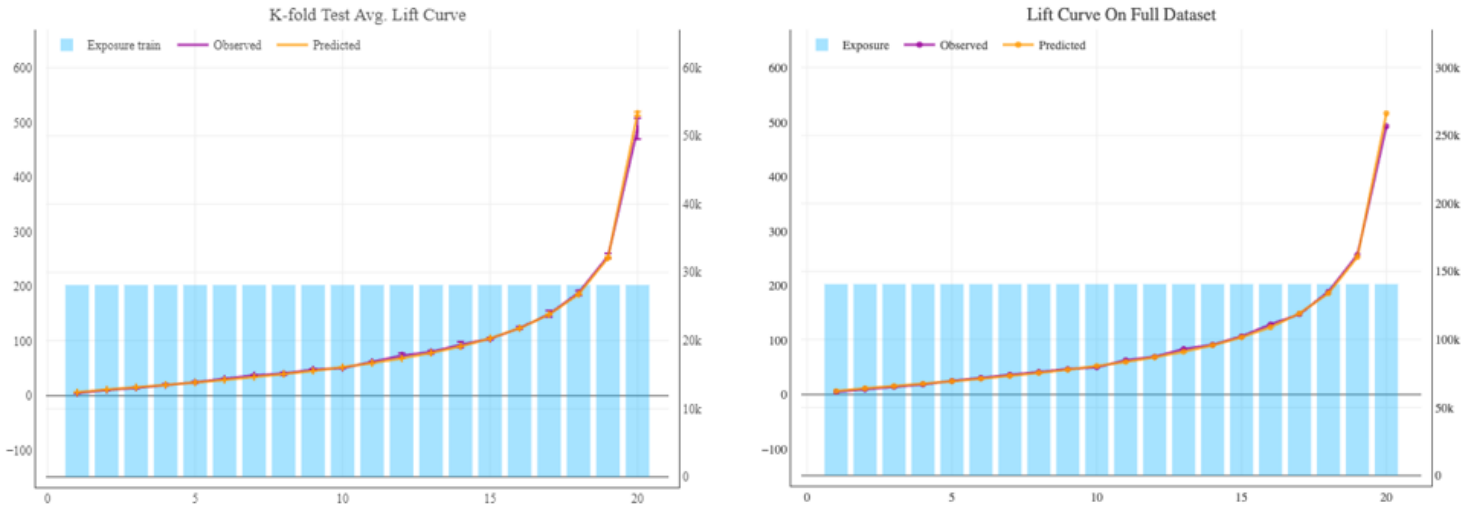


Figure 4.1.4 - Courbe Lift du modèle Tweedie méthode 1 zonier

La courbe de Lift est un graphique contenant deux courbes, la prédiction moyenne en jaune (prime pure prédite) et l'observation moyenne en violet (prime pure observée). Pour les échantillons tests de la base d'apprentissage et les échantillons de toute la base de données, les deux courbes se superposent ce qui permet de déduire que les observations et les prédictions sont égales en moyenne. La moyenne des prédictions présente alors une tendance similaire à la moyenne des observés. Le modèle ajuste donc bien les données. Néanmoins, on peut remarquer une très légère surestimation de la prime pure prédite pour les gros risques.

Les résidus :

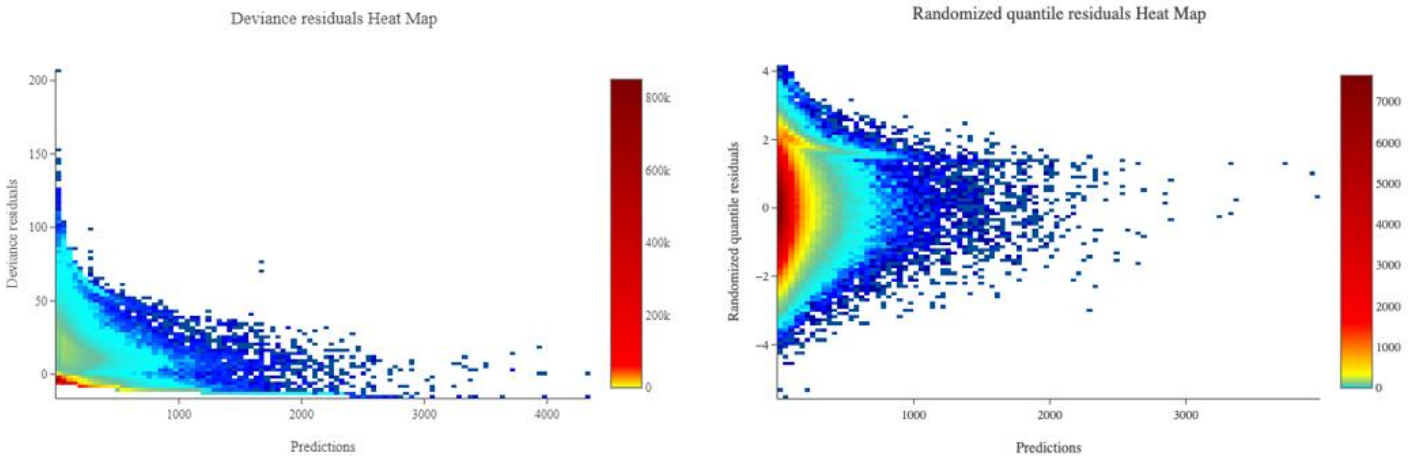


Figure 4.1.5 - Résidus du modèle Tweedie méthode 1 zonier

Pour la modélisation avec l'approche Tweedie, l'outil ne propose qu'une représentation graphique des résidus de déviance (graphique à gauche). L'interprétation de ces résidus est complexe, car il n'y a aucune garantie théorique à la forme à laquelle les résidus de déviance doivent ressembler pour un modèle bien ajusté. Toutefois, il n'y a qu'un seul groupe de résidus donc il n'y a pas une distribution bimodale derrière les observations ce qui est rassurant. La grande majorité des résidus sont autour de 0. L'écart maximal entre l'observé et le prédit est d'environ 4 500€, qui ne concerne qu'un très faible nombre d'observations. En représentant les résidus quantiles sur toute la base de données (graphique à droite), les résidus suivent une loi normale centrée en 0. La prime pure est bien calibrée.

Les indicateurs numériques RMSE et Gini sur les bases d'apprentissage et de validation sont proches. Il n'y a pas de sur-apprentissage observé, le modèle obtenu est robuste.

	Base apprentissage - Echantillon test	Base validation	Full dataset
RMSE	801	809	803
Gini	53,23%	53,47%	53,63%

Tableau 4.1.1 - Indicateurs numériques de performance pour le modèle Tweedie

4.2 Modélisation fréquence

Les résultats de la modélisation de la fréquence sur la base apprentissage sont présentés dans cette section. La variable à prédire est donc la fréquence définie par :

$$\text{Fréquence} = \frac{\text{nombre de sinistres}}{\text{exposition}}$$

Sélection de variables

Les différents modèles proposés en fonction du nombre de variables retenues sont représentés ci-dessous :

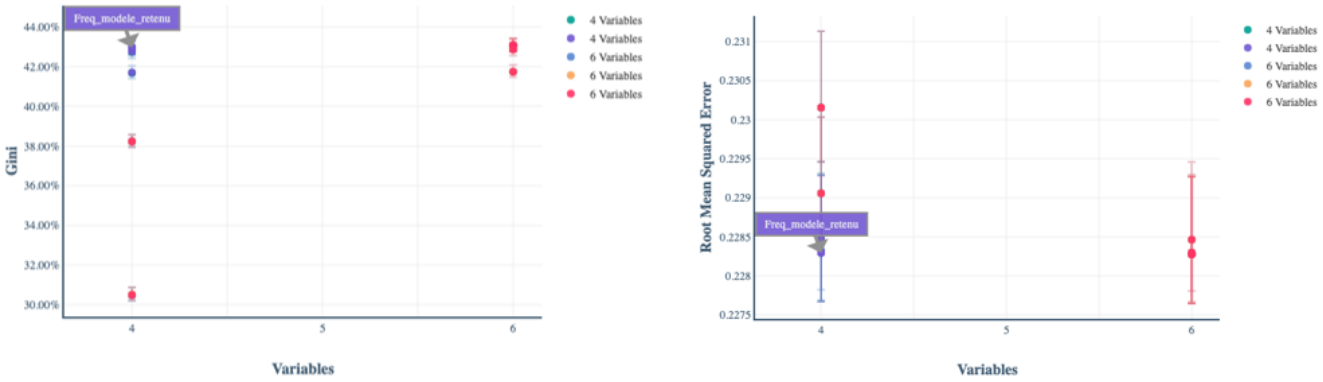


Figure 4.2.1 - Sélection du nombre de variables pour le modèle fréquence (méthode 1 zonier)

Ici, le modèle à quatre variables explique tout aussi bien que le modèle à six variables. Le pouvoir explicatif du modèle n'évolue pas avec des variables supplémentaires. De plus, le RMSE est quasiment équivalent entre un modèle à quatre variables et six variables. Finalement, le modèle de fréquence retenu est celui qui possède quatre variables.

Qualité du modèle

Comme la modélisation Tweedie, pour appréhender la qualité du modèle de fréquence retenu à quatre variables, plusieurs analyses graphiques seront effectuées : le spread, la courbe de Lorenz, la courbe de Lift, les résidus.

Le spread :

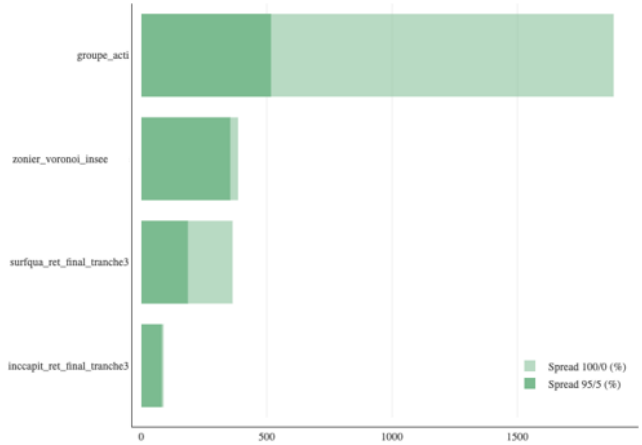


Figure 4.2.2 - Spread du modèle fréquence (méthode 1 zonier)

Le graphique ci-dessus représente l'importance des variables sélectionnées sur les prédictions par le modèle de fréquence. Les variables les plus pertinentes et discriminantes par ordre croissant sont le groupe d'activité, le zonier, la surface, le contenu incendie. Pour le modèle de fréquence, le chiffre d'affaires n'est pas une variable discriminante.

La courbe de Lorenz :

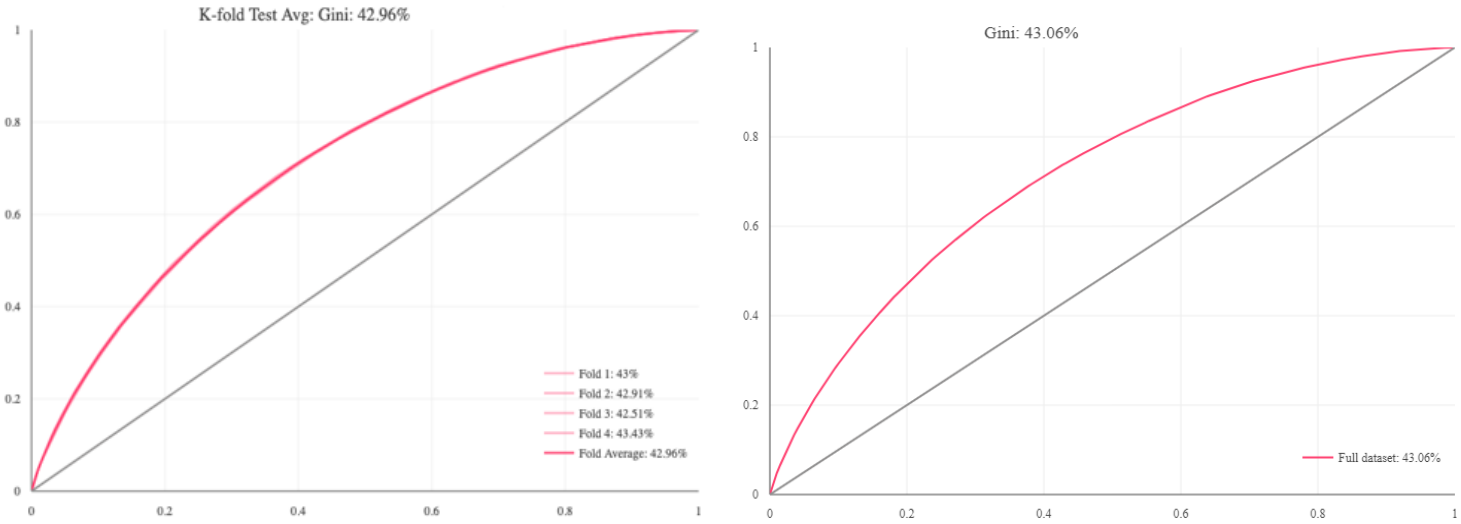


Figure 4.2.3 - Courbe de Lorenz du modèle fréquence (méthode 1 zonier)

Le modèle donne de bons résultats puisque les Ginis sont stables au cours des quatre cross-validation. Ici, dans le cadre de la modélisation de la fréquence, l'indice de Gini moyen sur les 4-fold est de 42,96% traduisant une segmentation satisfaisante de notre portefeuille par ce modèle. Le graphique à droite affiche un indice de 43,06% sur toute la base de données.

La courbe Lift :

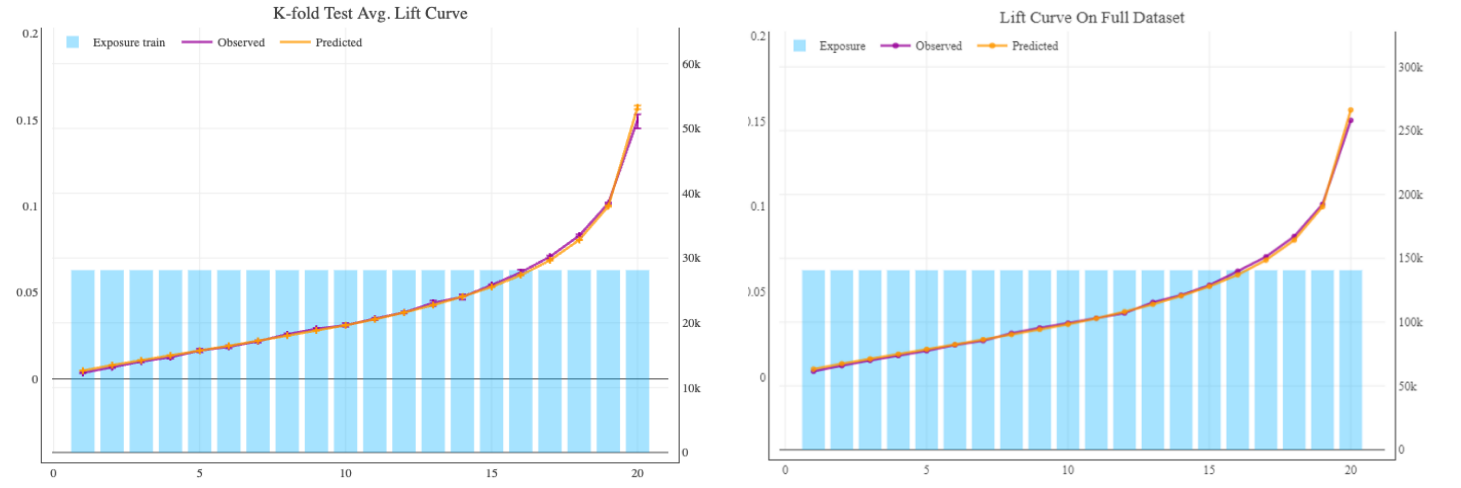


Figure 4.2.4 - Courbe Lift du modèle fréquence (méthode 1 zonier)

Les deux courbes se superposent ce qui permet de déduire que les observations et les prédictions sont égales en moyenne. La moyenne des prédictions présente alors une tendance similaire à la moyenne des observés. Le modèle ajuste donc bien les données. Néanmoins, on peut remarquer une légère surestimation et sous-estimation dans les prédictions des extrêmes, pour les petits risques et les gros risques. Même conclusion obtenue que la prime pure.

Les résidus quantiles :

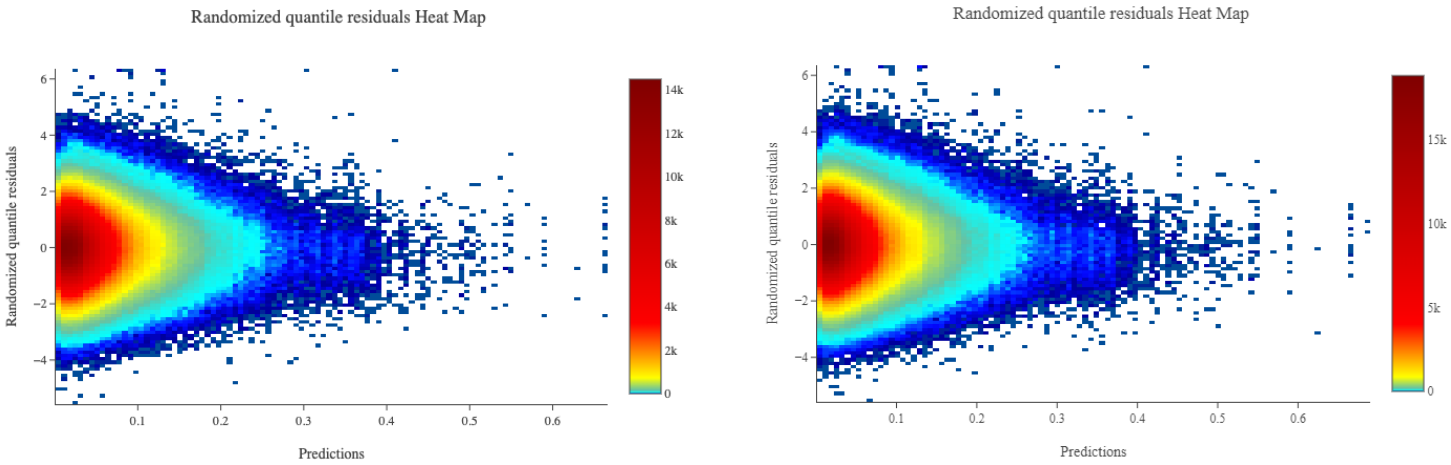


Figure 4.2.5 - Les résidus du modèle fréquence (méthode 1 zonier)

Graphiquement, le modèle de fréquence peut être validé, puisque globalement les résidus ont une forme symétrique, centrés autour de 0 et la majorité des résidus sont dans l'intervalle -2 et 2. Ainsi, les hypothèses de résidus sont vérifiées et le modèle validé.

Les indicateurs numériques RMSE et Gini sur les bases d'apprentissage et de validation sont proches. Il n'y a pas de sur-apprentissage observé, le modèle obtenu est robuste.

	Base apprentissage - Echantillons tests	Base validation	Full dataset
RMSE	0,23	0,23	0,23
Gini	42,96%	42,63%	53,63%

Tableau 4.2.1 - Les indicateurs numériques de performance du modèle fréquence (méthode 1 zonier)

4.3 Modélisation coût moyen

Les résultats de la modélisation du coût moyen sur la base apprentissage sont présentés dans cette section. La variable à prédire est donc le coût moyen défini par :

$$\text{Coût moyen} = \frac{\text{charge des sinistres}}{\text{nombre de sinistres}}$$

Sélection de variables

Les différents modèles proposés en fonction du nombre de variables retenues sont représentés ci-dessous :

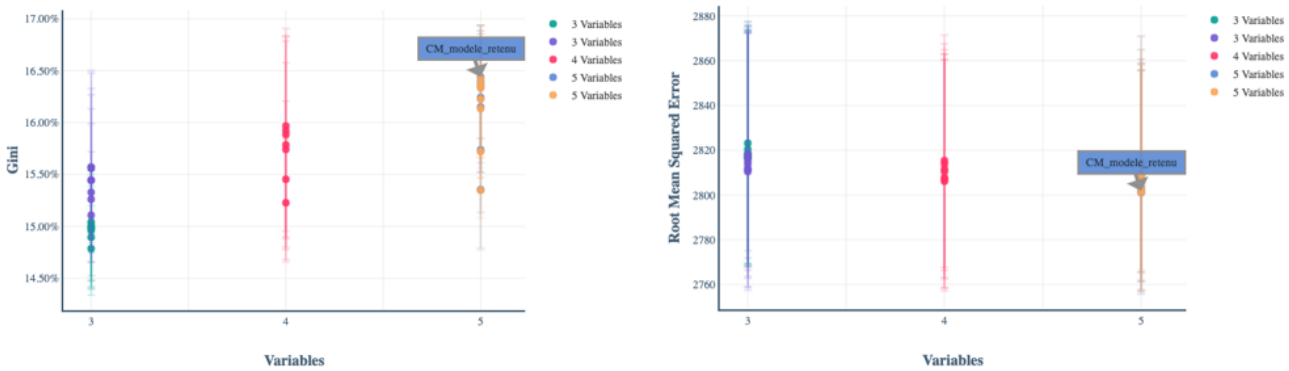


Figure 4.3.1 - Sélection du nombre de variables pour le modèle CM (méthode 1 zonier)

La variable zonier était une variable très discriminante dans le modèle de fréquence. Cependant, il apparaît de manière moins significative dans le modèle de coût moyen. De ce fait, cette variable a été retirée de la modélisation dans le cadre du modèle coût moyen, il apparaîtra tout de même dans l'agrégation des modèles fréquence-coût moyen. Le modèle à cinq variables est retenu puisqu'il maximise l'indice de Gini et minimise l'indicateur RMSE.

Qualité du modèle

Pour appréhender la qualité du modèle coût moyen retenu à cinq variables, plusieurs analyses graphiques seront effectuées : le spread, la courbe de Lorenz, la courbe de Lift, les résidus.

<u>Le spread :</u>

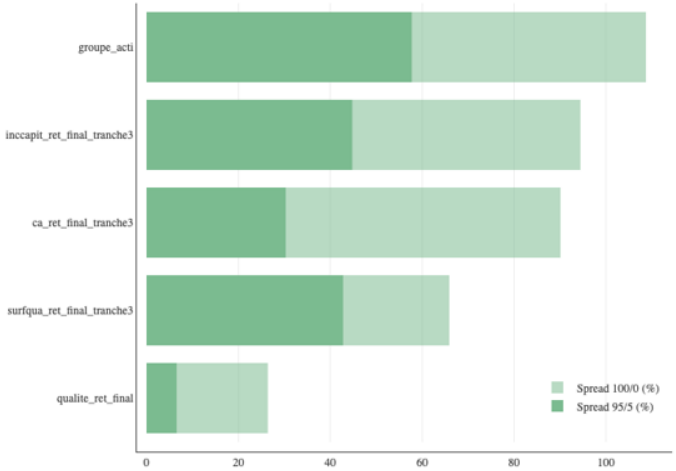


Figure 4.3.2 - Spread du modèle CM (méthode 1 zonier)

Le graphique ci-dessus représente l'importance des variables sélectionnées sur les prédictions par le modèle de coût moyen. Les variables les plus pertinentes et discriminantes par ordre croissant sont le groupe d'activité, le contenu incendie, le chiffre d'affaires, la surface et la qualité. Les variables chiffre d'affaires et qualité n'étaient pas considérées comme discriminantes pour le modèle de fréquence mais elles ressortent comme explicatives pour le modèle de coût moyen.

La courbe de Lorenz :

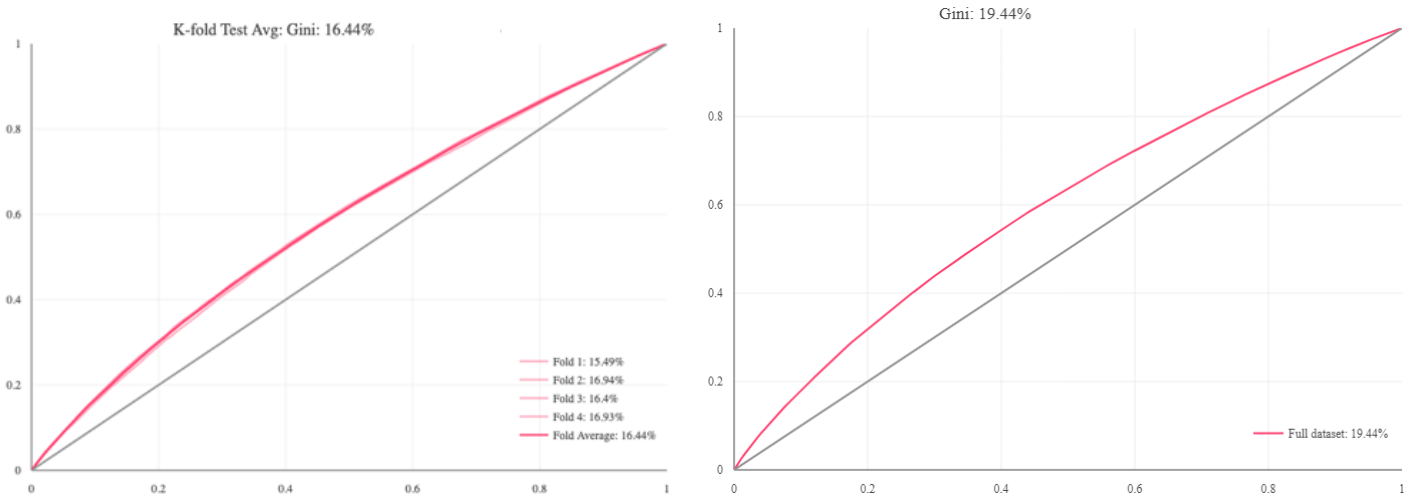


Figure 4.3.3 - Courbe de Lorenz du modèle CM (méthode 1 zonier)

Les Ginis sont stables au cours des quatre cross-validation, le modèle donne des résultats acceptables. Ici, dans le cadre de la modélisation du coût moyen, l'indice de Gini moyen sur les 4-fold est de 16,44% et l'indice de Gini sur toute la base de données est de 19,44%.

La courbe Lift :

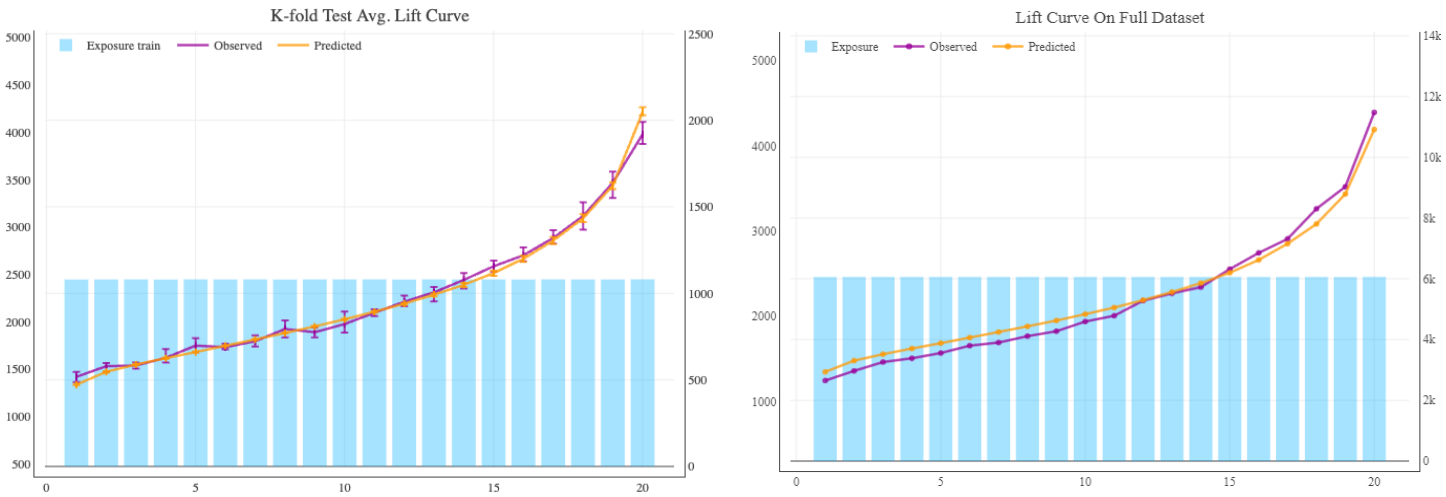


Figure 4.3.4 - Courbe Lift du modèle CM (méthode 1 zonier)

La moyenne des prédictions présente alors une tendance similaire à la moyenne des observés. Les deux courbes se superposent légèrement. Le coût moyen n'est pas assez bien expliqué par le modèle, on peut remarquer des surestimations et des sous-estimations dans les prédictions et des barres d'erreur très élevées.

Les résidus quantiles :

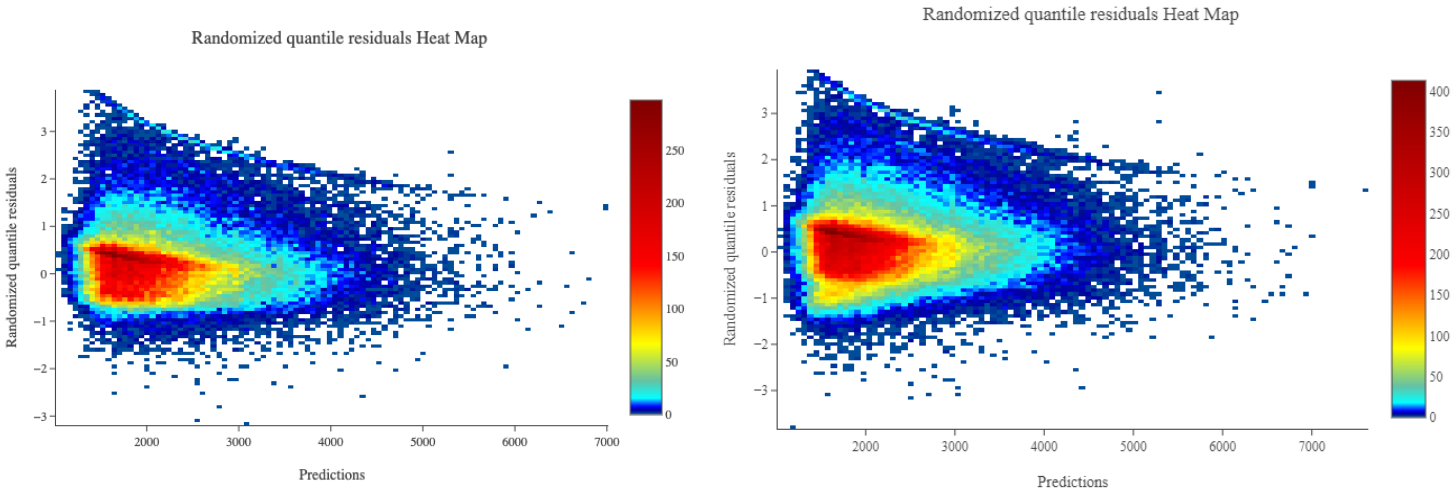


Figure 4.3.5 - Résidus du modèle CM (méthode 1 zonier)

En observant les résidus quantiles du modèle de coût moyen, une asymétrie est observée, certains sinistres sont largement sous-prédits et d'autres sur-prédits. Les résidus supérieurs à deux montrent qu'il y a beaucoup de gros risques qui sont fortement sous-prédits. Les résidus ne sont pas centrés en 0 mais plutôt à -0.5, ce qui signifie que beaucoup de sinistres sont sur-prédits. Ces interprétations coïncident avec les observations de la courbe Lift. Ainsi, pour ce modèle, l'hypothèse d'une distribution gamma ne semble pas être la loi la plus adaptée au modèle.

Les indicateurs numériques RMSE et Gini sur les bases d'apprentissage et de validation sont proches, mais le coût moyen n'est pas bien expliqué par le modèle.

	Base apprentissage - Echantillons tests	Base validation	Full dataset
RMSE	2802	2829	2751
Gini	16,44%	16,99%	19,44%

Tableau 4.3.1 – Indicateurs numériques de performance du modèle CM (méthode 1 zonier)

4.4 Agrégation fréquence * coût moyen

L'agrégation consiste à croiser les modèles de fréquence et de coût moyen pour obtenir la prime pure :

$$\text{prime pure observée} = \text{Fréquence} \times \text{Coût moyen} = \frac{\text{nombre de sinistres}}{\text{exposition}} \times \frac{\text{charge des sinistres}}{\text{nombre de sinistres}}$$

Ci-dessous les résultats sur les modèles combinés :

Qualité du modèle

Le modèle agrégé contient six variables dont quatre variables en commun aux modèles fréquence et CM et deux autres spécifiques au modèle CM (chiffre d'affaires et qualité). Pour appréhender la qualité de ce modèle, les mêmes analyses graphiques sont effectuées, à savoir le spread, la courbe de Lorenz, la courbe de Lift, les résidus.

Le spread :

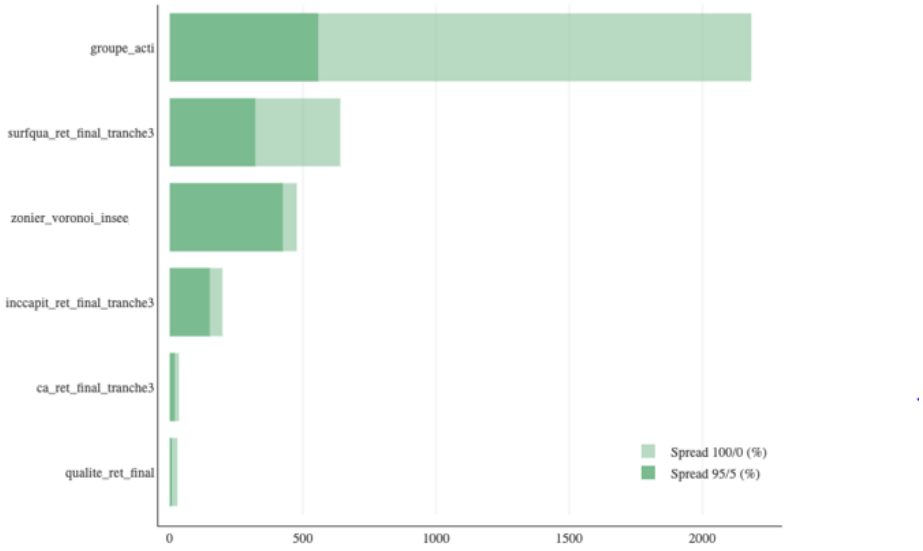


Figure 4.4.1 - Spread du modèle fréquence x CM (méthode 1 zonier)

Les variables les plus pertinentes et discriminantes par ordre croissant sont le groupe d'activité, la surface, le zonier, le contenu incendie, le chiffre d'affaires et la qualité.

La courbe de Lorenz :

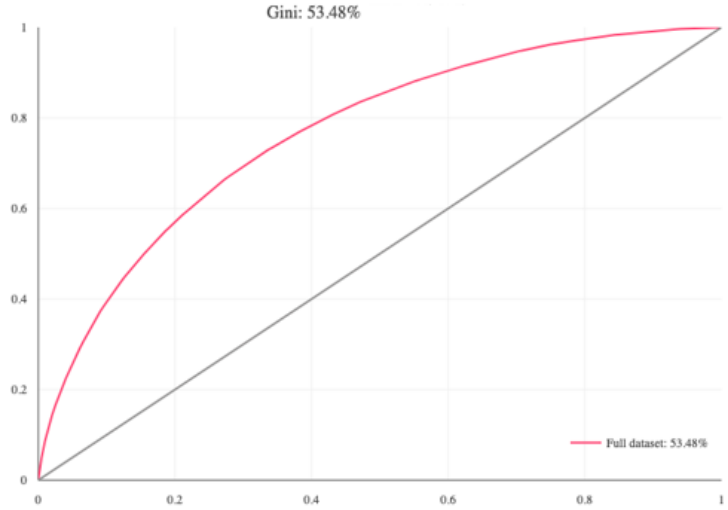


Figure 4.4.2 - Courbe de Lorenz du modèle fréquence x CM (méthode 1 zonier)

Le Gini sur toute la base de données est de 53,48% pour le modèle de fréquence-coût moyen. Il est légèrement inférieur au modèle Tweedie (53,63%).

La courbe Lift :

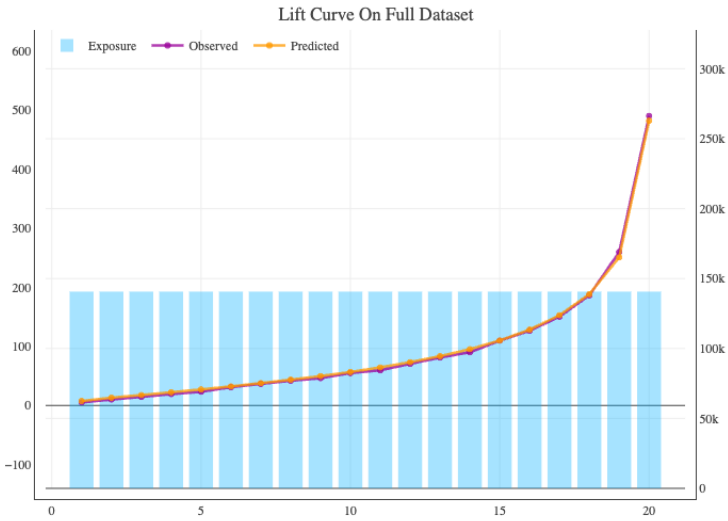


Figure 4.4.3 - Courbe Lift du modèle fréquence x CM (méthode 1 zonier)

Les deux courbes se superposent. La moyenne des prédictions présente alors une tendance similaire à la moyenne des observés. Le modèle en approche fréquence-CM ajuste parfaitement les données, avec une très légère sous-estimation pour la dernière classe.

Les résidus quantiles :

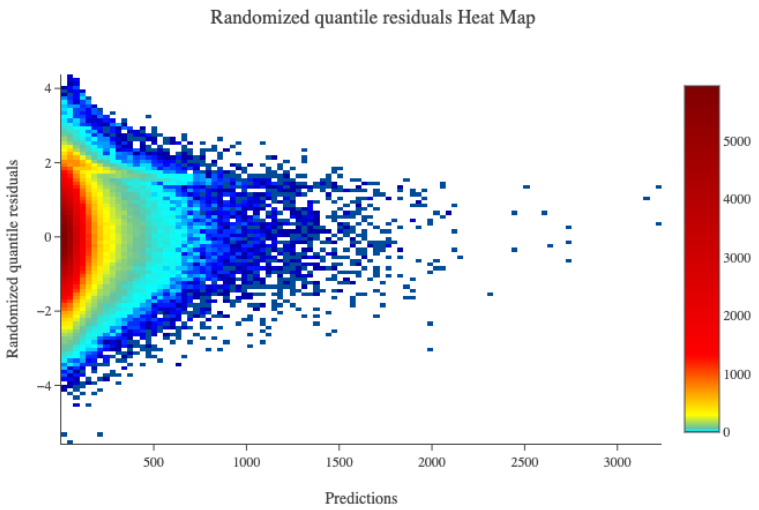


Figure 4.4.4 - Résidus du modèle fréquence x CM (méthode 1 zonier)

Aux vues des résidus quantiles, la prime pure modélisée par le modèle agrégé, fréquence-coût moyen, peut être validée.

4.5 Comparaison des modèles

Cette section sera consacrée à la comparaison des modèles Tweedie et fréquence-CM selon la méthode d'ajout simple du zonier. La comparaison est réalisée à l'aide des graphiques de spread, courbe de Lorenz et courbe Lift. Le modèle de référence correspond au modèle Tweedie (en approche prime pure) et le modèle comparé est le modèle fréquence-CM.

Spread :

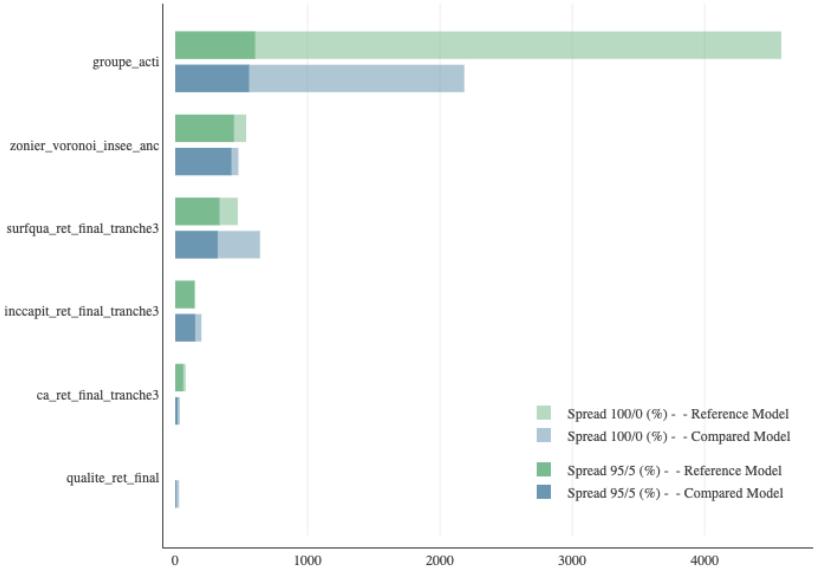


Figure 4.5.1- Comparatif spread des modèles Tweedie et fréquence x CM

Comme vu précédemment, le modèle fréquence-CM comporte plus de variables discriminantes que le modèle Tweedie. Ceci est expliqué par l'agrégation des modèles fréquence et coût moyen qui peut démultiplier le nombre de variables dès lors qu'une variable n'est pas commune sur les deux modélisations. On peut remarquer que l'ensemble des variables du modèle Tweedie sont contenues dans celui du modèle fréquence-CM. De plus l'ordre d'importance des variables reste proche. Le groupe activité est une variable très discriminante pour les deux modélisations. La surface est plus discriminante dans le modèle fréquence-CM que le modèle Tweedie. La qualité est une variable pertinente seulement pour le modèle fréquence-CM.

Courbe de Lorenz :

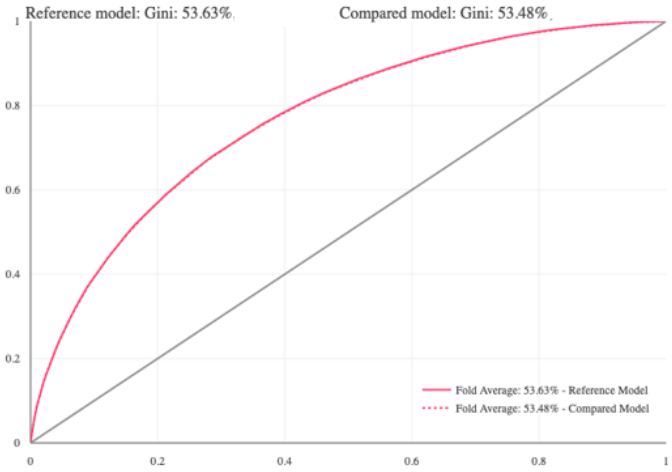


Figure 4.5.2 - Comparatif courbe de Lorenz des modèles Tweedie et fréquence x CM

Les courbes de Lorenz des deux modèles sont très proches ce qui révèlent une bonne performance, cependant la prime pure est légèrement mieux modélisée avec le modèle Tweedie.

Courbe Lift :

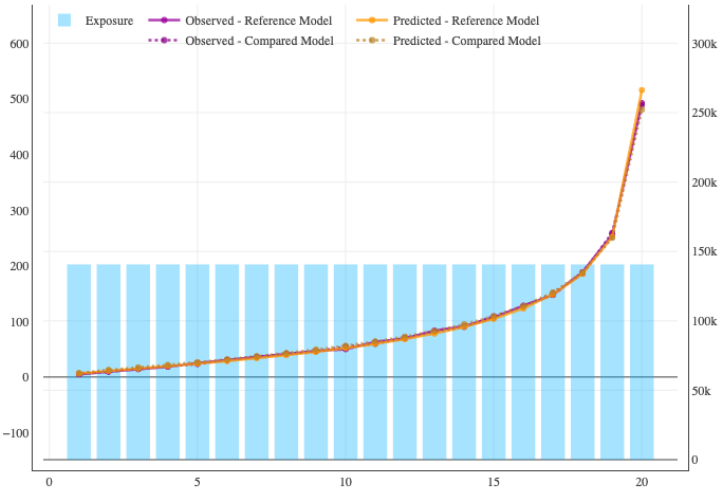


Figure 4.5.3 - Comparatif courbe Lift des modèles Tweedie et fréquence x CM

Peu de différence est observée entre les deux prédictions à part au niveau des gros risques où la prime pure est légèrement surestimée avec le modèle Tweedie et légèrement sous-estimée avec le modèle fréquence x CM. Globalement, les observations et les prédictions sont égales en moyenne et les deux modèles ajustent bien les données.

Ci-dessous un tableau des indicateurs numériques des deux modèles sur toute la base de données :

	Full dataset modèle Tweedie	Full dataset modèle fréquence - CM
RMSE	803	803
Gini	53,63%	53,48%

Tableau 4.5.1 - Comparatif des indicateurs numériques de performance des modèles Tweedie et fréquence x CM

Les indicateurs et les graphiques montrent que le modèle Tweedie ajuste mieux les données avec une meilleure performance. A la vue de l'analyse de l'ensemble des performances, le choix du modèle se tend vers le modèle Tweedie. De plus, il est préférable de surestimer légèrement la prime pure que de la sous-estimer. Par conséquent, le modèle Tweedie est retenu et privilégié par rapport au modèle fréquence x CM. Par ailleurs, en choisissant le modèle Tweedie, cela permet d'éviter de supposer l'indépendance entre la fréquence et le coût.

4.6 Détails sur le modèle retenu

Nous avons vu que le modèle Tweedie présentait une meilleure qualité d'adéquation sur la base d'apprentissage et toute la base de données. Dans cette section, les détails du modèle retenu seront présentés, c'est-à-dire les prédictions par variable avec les coefficients selon la segmentation. Elles seront présentées par ordre d'importance dans la modélisation, le groupe d'activité étant la variable la plus discriminante.

Deux graphiques par variable seront étudiés :

- Le graphique à gauche permet d'apprécier les coefficients par modalité. Sur ce graphique, sont affichés : l'exposition sur la base d'apprentissage par modalité, la valeur relative des coefficients (vert), la valeur de la charge moyenne observée (violet) et la valeur de la charge moyenne prédite (orange).
- Le graphique à droite permet d'évaluer la robustesse du modèle en affichant les coefficients par modalité et par année de survenance du sinistre (time consistency).

<u>Groupe activité :</u>

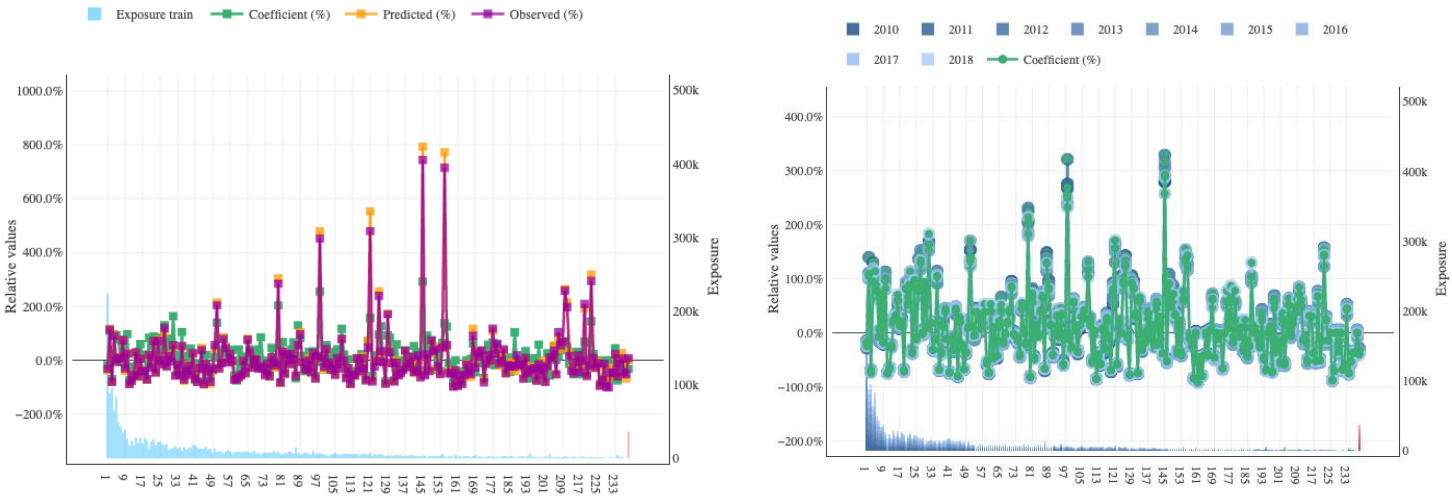


Figure 4.6.1 - Coefficients pour la variable groupe activité

Le groupe d'activité est une variable majeure dans la tarification pour la garantie DDE. Les activités ont subi un premier regroupement par type d'activité ou par un jugement à dire d'expert avant de passer à la modélisation. Puis lors de la modélisation, des coefficients sont attribués pour chaque groupe d'activité permettant de distinguer les activités à risques. Les trois groupes d'activité les plus représentés sont les bureaux (groupe 1), les restaurants (groupe 2) et les travaux de bâtiment (groupe 3). Les activités les plus risquées selon les coefficients calibrés sont les hôtels-restaurants et les moins risqués les charpentiers bois couvreur. Les coefficients de la variable groupe activité sont stables dans le temps.

Zoniers :

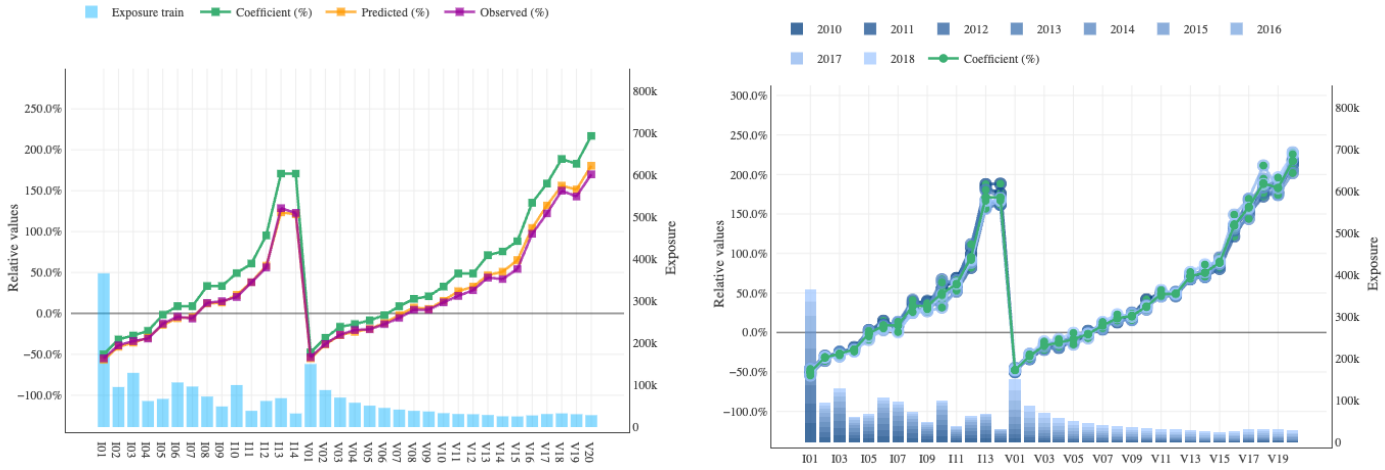


Figure 4.6.2 - Coefficients pour la variable zonier

Le zonier est la deuxième variable tarifaire la plus importante dans la tarification. La méthode d'implémentation retenue est le croisement des zoniers dans une variable, qui a été possible car le zonier INSEE est une bonne approximation du zonier Voronoi, lorsque ce dernier n'est pas renseigné. Ainsi, les zoniers INSEE et Voronoi sont calibrés dans une unique variable de 34 modalités, dont 14 proviennent du zonier INSEE et 20 du zonier Voronoi. Nous pouvons remarquer que les coefficients respectent l'évolution du risque. Plus le zonier est élevé plus le risque est élevé et donc plus le coefficient est élevé. Ces coefficients sont stables dans le temps.

Surface :

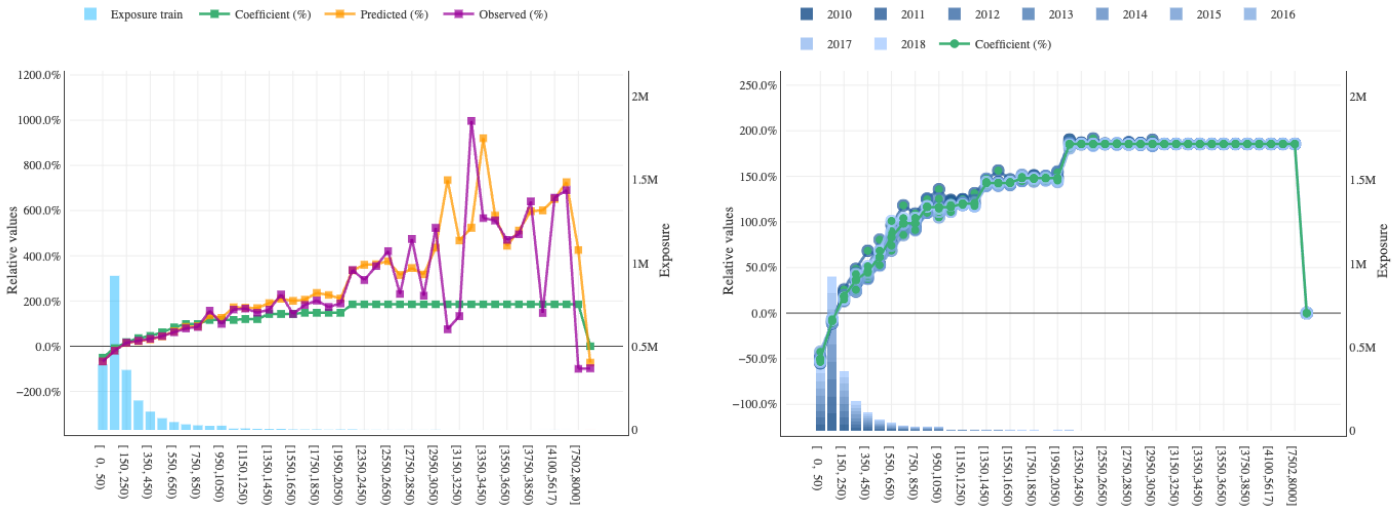


Figure 4.6.3 - Coefficients pour la variable surface

La surface est également discriminante dans la tarification. Nous supposons que plus la surface est grande plus la prime pure sera élevée avec des dégâts considérables à indemniser. Il ne serait pas cohérent qu'une entreprise de 3000 m^2 paye moins cher qu'une entreprise de 1000 m^2, même si la sinistralité observée sur cette tranche est plus faible. En suivant cette logique, cela amène à imposer une contrainte de croissance des coefficients pour cette variable ordinale. Certaines tranches ont des coefficients égaux, ce qui signifie qu'un même comportement est observé en termes de sinistralité. De plus, afin d'éviter d'avoir de fortes variations à cause d'une faible volumétrie sur les tranches les plus élevées, le modèle est plus ou moins lissé. Sur le graphique, les surfaces de plus de 2050 m^2 ont le

même coefficient, ils sont regroupés dans une même classe et correspondront à la classe la plus risquée. Par exemple, une entreprise de plus de 2050 m^2 sera dans la même case tarifaire qu'une entreprise à 7000 m^2. Ces coefficients sont stables dans le temps.

Contenu incendie :

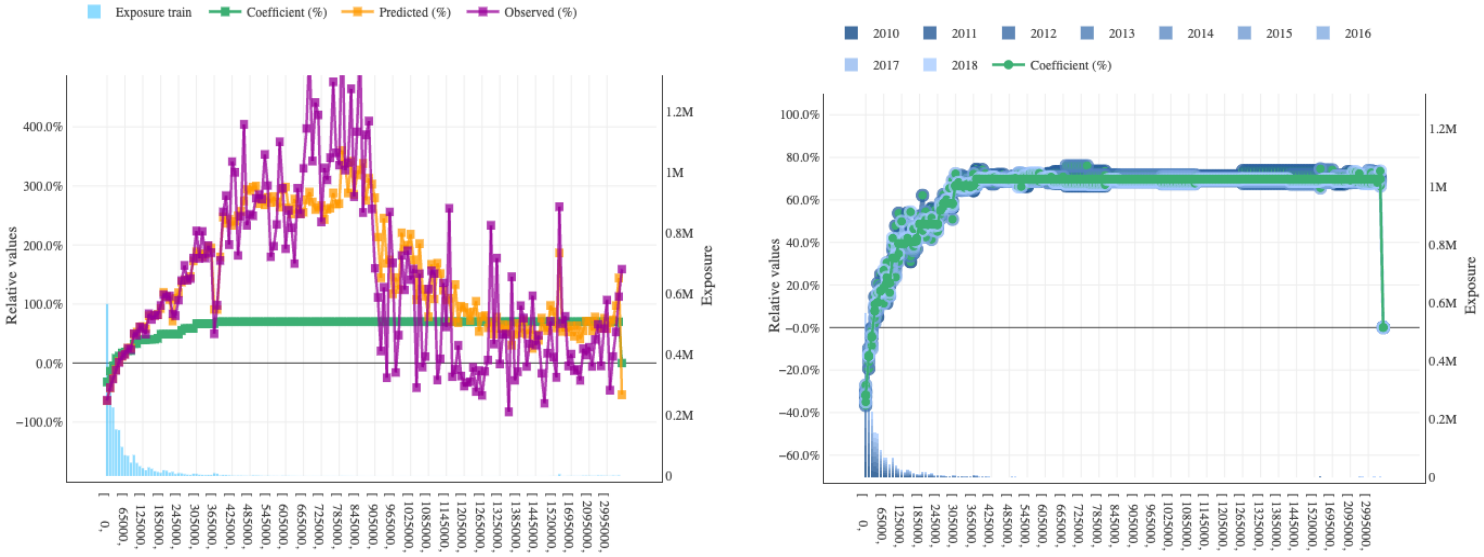


Figure 4.6.4 - Coefficients pour la variable contenu incendie

Le contenu incendie qui est égal au contenu dégât des eaux est une variable qui intervient dans le tarif rapide avec une importance moindre que le groupe d'activité, le zonier et la surface. De même que la surface, les coefficients de la variable contenu incendie doivent être strictement croissants et les tranches sont regroupées pour éviter d'avoir trop de variation entre les segmentations. La dernière classe regroupe plusieurs modalités due à la faible exposition pour les contenus très élevés. Ces coefficients sont stables dans le temps.

Chiffre d'affaires :

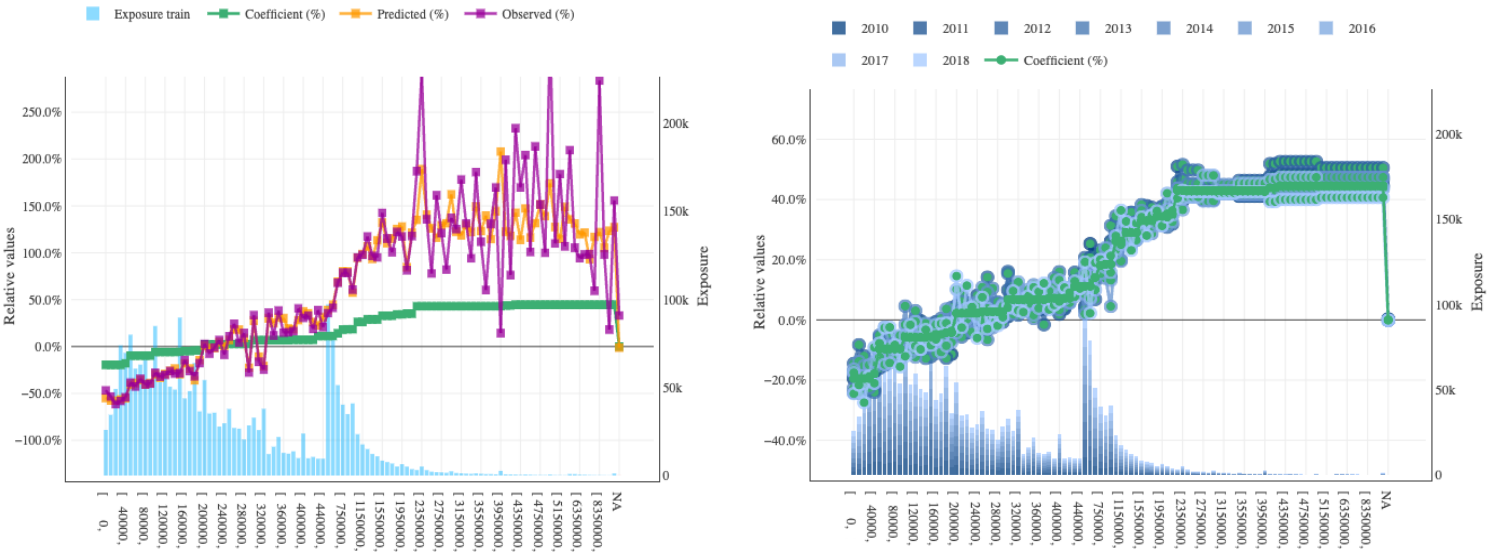


Figure 4.6.5 - Coefficients pour la variable chiffre d'affaires

Le chiffre d'affaires annuel est la dernière variable pertinente du modèle. De même, les coefficients doivent être strictement croissants et les modalités sont regroupées selon les valeurs des coefficients calibrés. Ces coefficients sont stables dans le temps.

Finalement, après avoir analysé chaque variable tarifaire du modèle, nous avons vu que certaines tranches ont des coefficients égaux (regroupement réalisé) afin d'éviter une forte variation des coefficients. Par ailleurs, les coefficients de toutes les variables sont stables dans le temps, ce qui révèle que le modèle retenu est robuste et que la charge ultime et l'indexation ont bien été prises en compte dans la modélisation.

4.7 Comparaison des primes pures

Cette section sera consacrée à la comparaison des primes pure sur la base de validation (non utilisée lors du calibrage du modèle) pour la prise de décision. En effet, les performances sur cette base permettront d'effectuer une validation finale des modèles calibrés.

Rappelons tout d'abord le contexte du mémoire : le tarif rapide actuel possède deux contraintes, la première étant les six questions posées au client et la deuxième étant la contrainte de structure tarifaire technique qui implique l'utilisation de modalité de base (pour toutes les variables intervenant dans le tarif technique mais non posée au client). Or, par définition, un tarif rapide n'est limité qu'au nombre de variables, ce tarif n'est donc pas optimal actuellement et possède un fort risque de sous-estimation de la sinistralité. Ainsi, l'objectif est de challenger le tarif rapide actuel en élaborant un nouveau tarif rapide pour la garantie DDE, sans contrainte de structure imposée tout en respectant la restriction des six questions posées au client. Son développement permet de revoir la structure tarifaire, de juger de la pertinence des six questions posées au client (définies par connaissance métier), d'évaluer le risque engendré par les modalités de base, c'est-à-dire d'étudier l'impact sur le tarif rapide de s'imposer une structure tarifaire du tarif technique et de quantifier la perte liée aux questions non posées au client (cas des antécédents pour la modélisation de la garantie DDE). Par ailleurs, sa production permet d'améliorer la compétitivité, de faciliter la souscription et d'avoir une idée brève du tarif pour répondre rapidement aux besoins des clients. La création de ce tarif rapide entraîne aussi la conception d'un nouveau tarif technique avec la structure du nouveau tarif rapide pour permettre de réaliser les différentes comparaisons. Le tarif rapide sera un « tarif à partir de », qui dépendra du risque du client et de certaines conditions, comme ne pas avoir d'antécédent. Si l'assuré possède des antécédents alors il devra indiquer le nombre et son tarif sera revu et majoré à l'aide du tarif technique.

Ainsi, les différentes primes pures à comparer entre eux sont :

- La prime pure observée $= \frac{\text{Charges}}{\text{exposition}}$;
- La prime pure modélisée du tarif rapide actuel, avec la contrainte de structure tarifaire technique : y compris les antécédents mais définis à zéro pour tout le monde (le coefficient est égal à 1 et sera neutre dans la tarification);
- La prime pure modélisée du nouveau tarif rapide, sans la contrainte de structure tarifaire technique : sans les antécédents ; ***(new)***
- La prime pure modélisée du tarif technique actuel, y compris les antécédents ;
- La prime pure modélisée du nouveau tarif technique avec la structure du nouveau tarif rapide et y compris les antécédents. ***(new)***

1. La comparaison des primes pures des deux nouveaux tarifs créés ***(cf. new)*** avec la prime pure observée permettra d'une part de valider les modèles et d'autre part de juger sur la pertinence des six questions posées au client. Pour valider ces modèles, une analyse des performances des modèles et des prédictions sur la base de validation sera présentée.
2. La comparaison entre la prime pure du tarif rapide actuel et du nouveau tarif rapide montrera l'impact sur le tarif rapide de s'imposer une structure tarifaire du tarif technique.
3. La comparaison entre la prime pure du tarif technique actuel et du nouveau tarif technique consistera à une analyse du tarif technique actuel sous contrainte des questions posées dans le cadre du tarif rapide.
4. La comparaison entre la prime pure du nouveau tarif rapide et du nouveau tarif technique permettra de quantifier et d'évaluer la perte liée aux questions non posées au client mais intervenant dans le tarif technique. Si l'écart entre les deux primes pures est faible alors la perte engendrée par les questions non posées est faible, et les six questions sont suffisantes pour établir un tarif rapide de la garantie DDE.

Les différentes comparaisons seront réalisées par des analyses graphiques : courbe de Lorenz, courbe Lift, histogramme des écarts de tarif au global et par segment de variable.

<u>Courbe de Lorenz :</u>

La courbe de Lorenz est une mesure de segmentation du tarif. L'indice de Gini, calculé à partir de cette courbe, permet d'évaluer la qualité de segmentation et donc la performance du modèle.

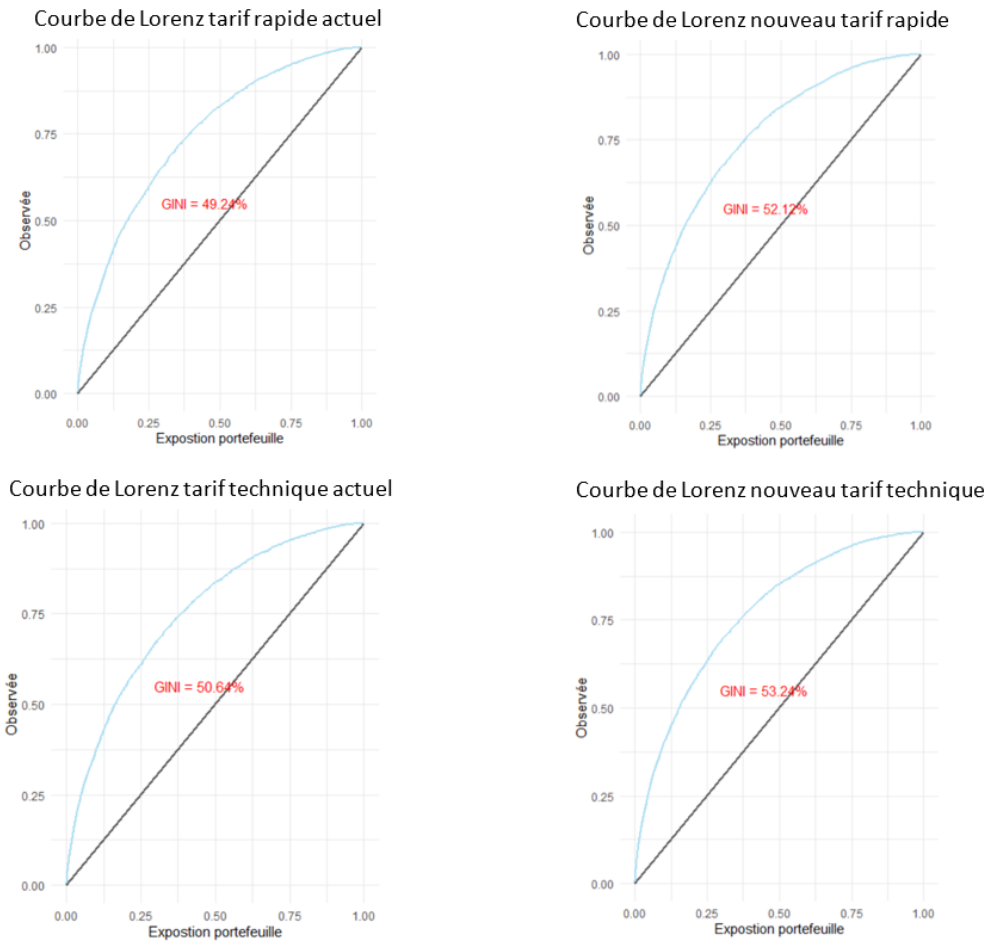


Figure 4.7.1 – Comparaison des courbes de Lorenz des modèles actuels et nouveaux sur la base de validation

Un tableau récapitulatif des indicateurs numériques des quatre modèles sur la base de validation :

	Base de validation					
	Tarif rapide actuel	**Nouveau tarif rapide**	Ecart	Tarif technique actuel	**Nouveau tarif technique**	Ecart
RMSE	752,8	**751,3**	-1,5	751,4	**751,2**	-0,2
Gini	49,24%	**52,12%**	2,88 points	50,64%	**53,24%**	2,6 points

Tableau 4.7.1 - Comparaison des indicateurs numériques de performance des modèles actuels et nouveaux

Le tarif rapide actuel possède un Gini de 49,24% et le nouveau tarif rapide possède un Gini de **52,12%**. Le tarif est donc mieux segmenté et plus performant avec la nouvelle modélisation. Il y a également une amélioration de la segmentation dans le tarif technique : le tarif technique actuel a un Gini de 50,64% contre **53,24%** pour le nouveau tarif technique.
Les nouvelles modélisations apportent donc une meilleure segmentation du tarif avec un gain de presque trois points en Gini et des RMSE légèrement plus faibles. Cet apport n'est pas négligeable et confirme l'importance du retraitement et de la qualité des données. Les analyses sur la base de validation ont montré la robustesse des modèles construits et ces modèles peuvent être validés.

Courbe Lift :

La courbe Lift est un indicateur visuel permettant d'évaluer la qualité des prédictions du modèle et de déduire si l'observé et le prédit sont égaux en moyenne. Sur les graphiques ci-dessous, en comparant la modélisation des primes pures actuelles (rapide et technique) avec la prime pure observée, les deux premiers graphiques du haut montrent une légère surestimation au niveau des petits risques et une sous-estimation considérable au niveau des gros risques. Le tarif technique actuel sous-estime la sinistralité qui implique une sous-estimation du tarif rapide actuel car ce dernier est issu de la structure du tarif technique actuel (ils ont également les mêmes coefficients). Le tarif rapide actuel est d'autant plus sous-estimé avec la modalité de base « sans antécédent ». Ainsi ces deux tarifs actuels sont critiquables. Toutefois la moyenne des prédictions présente une tendance similaire à la moyenne des observés.

En comparant les nouvelles modélisations avec la prime pure observée (deux graphiques du bas), il y a globalement une légère surestimation qui est d'autant plus visible pour les plus gros risques. L'écart entre l'observé et le prédit en moyenne est plus faible avec les nouvelles modélisations qu'avec les modélisations actuelles. Par ailleurs, il est plus judicieux de surestimer légèrement la prime pure que de la sous-estimer. Le tarif actuel nécessite de majorer plusieurs contrats dû à cette sous-estimation. Nous pouvons valider ces deux modèles et approuver la pertinence des six questions posées au client.

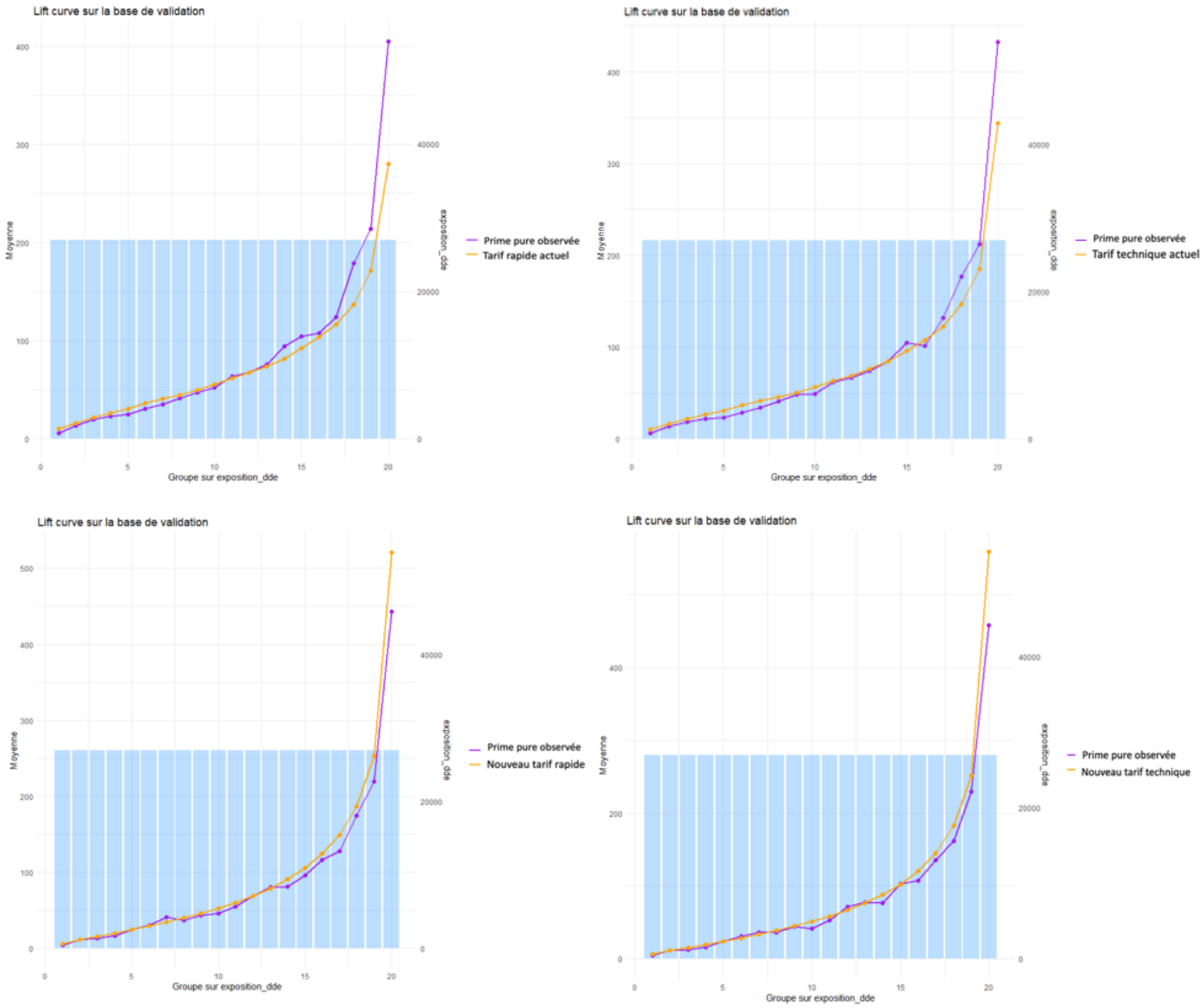


Figure 4.7.2 - Comparaison des courbes Lift des modèles actuels et nouveaux avec la prime pure observée sur la base de validation

Après avoir comparé la prime pure observée avec les modélisations de primes pures actuelles et nouvelles, il est intéressant de faire une comparaison entre l'actuel et le nouveau :

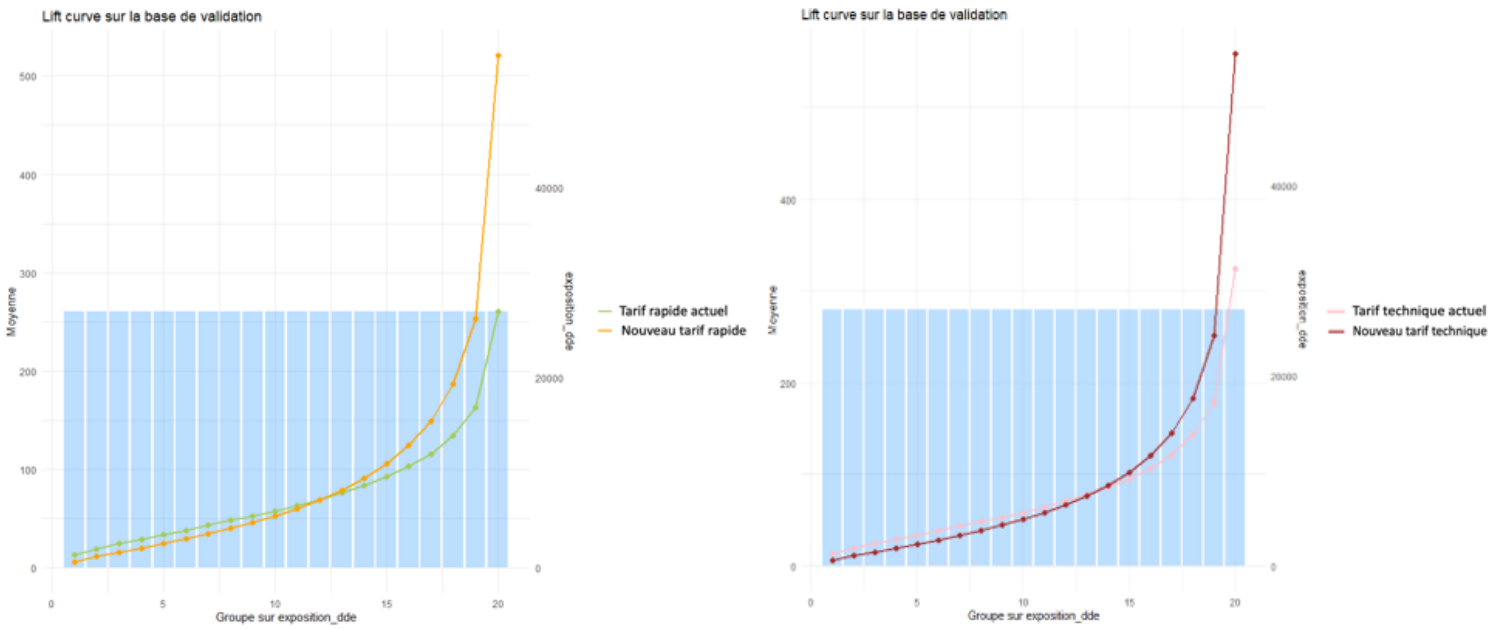


Figure 4.7.3 - Comparaison des courbes Lift des modèles actuels et nouveaux sur la base de validation

Une grosse différence est constatée entre la prime pure actuelle et la nouvelle prime pure. Ceci signifie qu'il y a un impact considérable de s'imposer une contrainte tarifaire du tarif technique. Il y a également un écart visible entre la prime pure du tarif technique actuel et du nouveau tarif technique, qui reflète qu'une revue du tarif technique actuel est nécessaire. Pour effectuer cette revue, nous nous sommes basés sur le tarif rapide modélisé auquel nous avons rajouté les antécédents pour former le nouveau tarif technique. La sous-estimation de la sinistralité peut être expliquée par les valeurs manquantes/aberrantes (valeurs nulles) non retraitées dans la modélisation actuelle. En effet, en indiquant une valeur nulle aux chiffres d'affaires, aux contenus incendie et aux surfaces, cela peut fausser la modélisation. Par exemple sans retraitement, le chiffre d'affaires n'est pas considéré comme une variable discriminante dans le tarif technique actuel alors qu'elle l'est potentiellement.

De plus, l'écart peut également être expliqué par la méthode d'implémentation du zonier. La modélisation actuelle repose sur la méthode off set et la nouvelle modélisation se démarque avec le croisement de zonier en une unique variable de modélisation. La qualité des données a donc une contribution importante dans la modélisation.

Pour une revue complète du tarif technique actuel, il aurait été judicieux de faire une refonte du tarif technique DDE et laisser le modèle sélectionner les variables à la différence de ce nouveau tarif technique basé sur les questions posées au client dans le cadre du nouveau tarif rapide.

Enfin, une dernière comparaison est possible : la comparaison entre la prime pure du nouveau tarif rapide et du nouveau tarif technique pour quantifier la perte liée aux questions non posées au client.

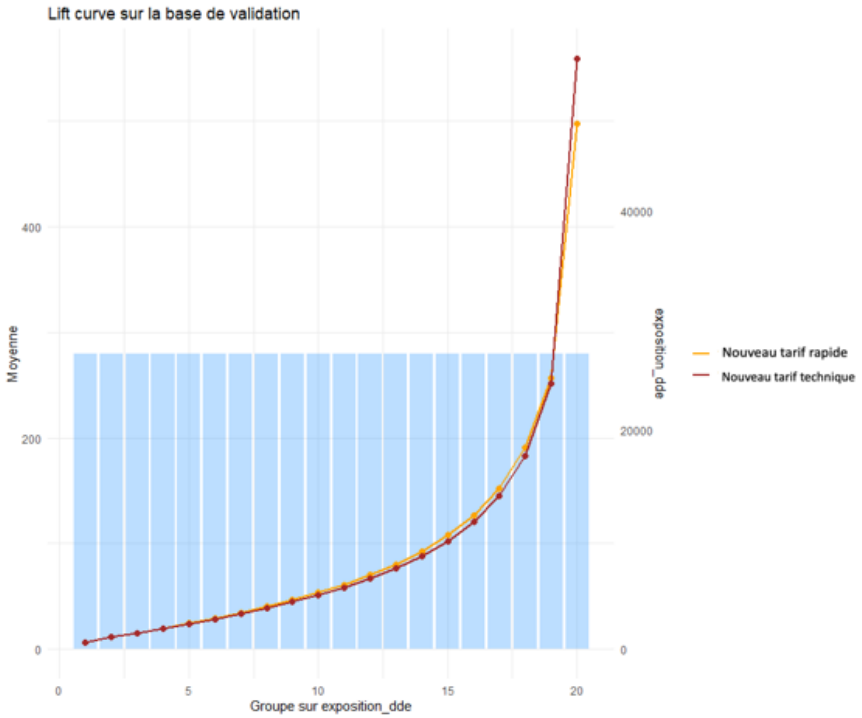


Figure 4.7.4 - Comparaison courbe Lift entre le nouveau tarif rapide et le nouveau tarif technique

Graphiquement, il y a très peu d'écart entre le nouveau tarif rapide et le nouveau tarif technique à l'exception de la dernière classe (très gros risque). Donc, la perte liée à cette question non posée est peu importante, on peut se permettre d'omettre cette question dans le cadre d'un tarif rapide. A nouveau, on peut déduire que les six questions posées au client sont suffisantes pour établir un tarif rapide de la garantie DDE.

Ci-dessous un tableau qui confirme que globalement, l'écart moyen (en euros) est faible entre le fait de ne pas considérer les antécédents (tarif rapide) et de les considérer dans le modèle (tarif technique). Cependant, en distinguant le cas des antécédents, en moyenne, nous remarquons que le tarif rapide surestime (écart moyen de 10%) la sinistralité des contrats n'ayant pas d'antécédent et sous-estime considérablement (écart moyen de -47%) la sinistralité des contrats ayant eu des antécédents.

	Au global			Sans antécédents			Antécédents		
	Nouveau tarif rapide	Nouveau tarif technique	Ecart	Nouveau tarif rapide	Nouveau tarif technique	Ecart	Nouveau tarif rapide	Nouveau tarif technique	Ecart
Moyenne	85,52	85,23	0,00	80,89	73,55	0,10	172,90	327,16	- 0,47

Tableau 4.7.2 - Ecart moyen entre le tarif rapide et le tarif technique

Le nuage de points ci-dessous confirme les observations précédentes. Le tarif rapide surestime légèrement la sinistralité des contrats n'ayant pas d'antécédent et sous-estime considérablement la sinistralité des contrats ayant eu des antécédents.

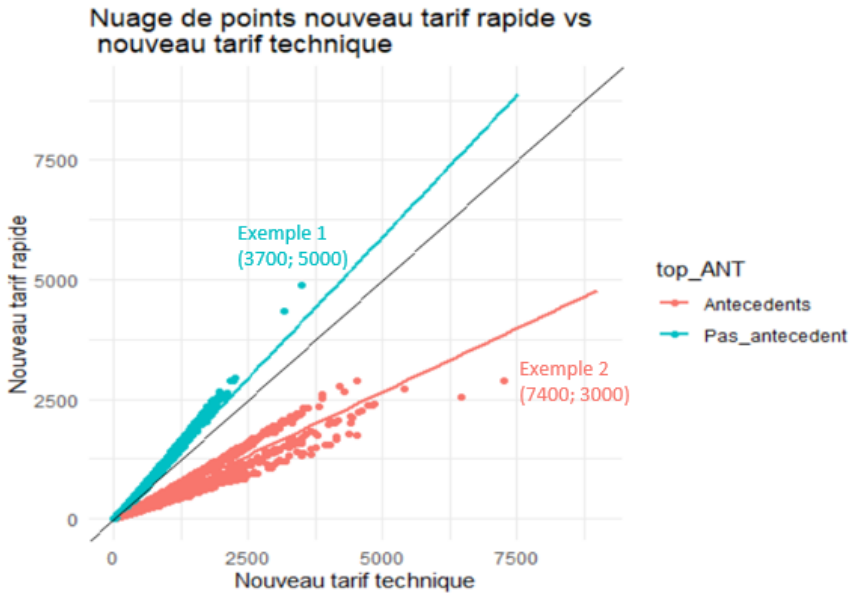


Figure 4.7.5 - Nuage de points nouveau tarif rapide vs nouveau tarif technique

Prenons des exemples de cas extrêmes sur le graphique :

- Exemple 1 point bleu : pour un contrat n'ayant pas eu d'antécédent, le tarif technique est d'environ 3 700€ mais le tarif rapide prédit 5 000€. (*Surestimation cas sans antécédent*)
- Exemple 2 point rouge : pour un contrat ayant eu au moins un antécédent, le tarif technique est d'environ 7 400€ mais le tarif rapide prédit 3 000€. (Sous-estimation cas avec antécédents)

L'écart maximal observé peut aller jusqu'à 4 400€, cela concerne qu'une faible densité (cas des contrats ayant plusieurs antécédents).

Au global, la perte engendrée par cette variable non posée au client est compensée. Cependant, ce tarif rapide est reprochable car il pénalise les assurés n'ayant pas d'antécédent et gratifie les assurés ayant eu des antécédents. Pour corriger cela, un lissage pourrait être réalisé en appliquant des majorations ou minorations pour les contrats qui ont un écart constaté.

Un algorithme CART (Classification And Regression Trees ou Arbres de classification et de régression) sur la différence des primes entre le nouveau tarif rapide et le nouveau tarif technique sera mis en œuvre. Il permettra de quantifier la perte de prime en fonction de différentes segmentations et d'établir des coefficients d'ajustement pour le nouveau tarif rapide par rapport au nouveau tarif technique. Cet algorithme découpe le portefeuille en fonction de l'ensemble des variables explicatives. Les données du portefeuille sont divisées en deux groupes de manière itérative.

Des spécifications sont à préciser pour l'application de celui-ci :

- la variable groupe activité est écartée car elle présente un nombre de modalités trop important,
- la variable zonier est retraitée en ordonnant les modalités par ordre croissant des coefficients calibrés par le GLM. Par exemple, l'ordre croissant du zonier selon les coefficients est (I01, V01, I02, V02, […], V20). Ainsi, le zonier initial (I01, V01, I02, V02, […], V20) devient le zonier ordonné et numérisé (1, 2, 3, 4, […], 34),
- comparaison des tarifs réalisées sur la base de validation,
- chaque feuille représente au minimum 2% de l'exposition
- Pour étudier la dispersion entre ces deux tarifs nous utilisons la formule suivante :

$$\text{Erreur relative (\%)} = \frac{\text{Prédiction nouveau tarif rapide} - \text{Prédiction nouveau tarif technique}}{\text{Prédiction nouveau tarif technique}}$$

  Idéalement, il faudrait avoir une erreur proche de zéro, qui signifierait que l'écart entre les deux tarifs est faible.

- un critère d'arrêt est mis en place afin de ne pas avoir un découpage trop fin. Le critère d'arrêt retenu est la complexité *cp* obtenu à partir de la fonction *rpart* sur le logiciel R. Il s'agit de la profondeur maximale de l'arbre, c'est-à-dire le nombre de niveaux de nœuds.

Ci-dessous l'arbre CART obtenu avec un *cp* égal à 0.0024 :

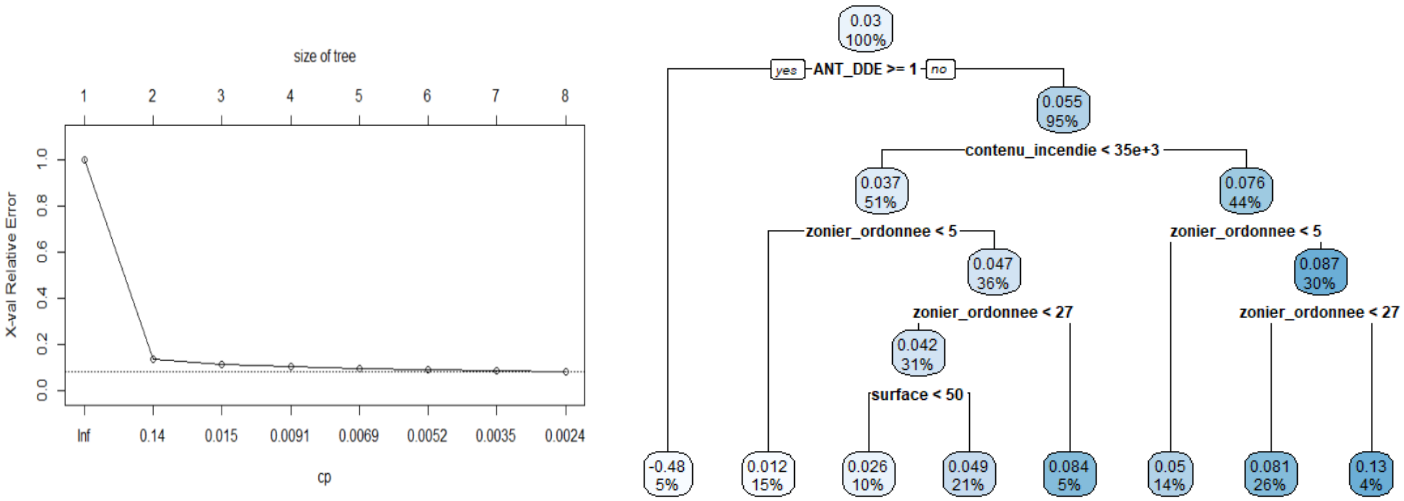


Figure 4.7.6 - Arbre CART dispersion entre le nouveau tarif rapide et le nouveau tarif technique

Cet arbre soutient les interprétations précédentes. Au global, l'écart de tarif est faible (3%) mais en analysant plus en profondeur, nous remarquons que pour les contrats ayant des antécédents, il y a une perte de presque la moitié de la prime dans le tarif rapide par rapport au tarif technique (concerne 5% du portefeuille). Cette perte est en fait compensée par les contrats n'ayant pas d'antécédent (95% du portefeuille).

Cette comparaison conclut que les six variables posées au client sont pertinentes et suffisantes pour obtenir un tarif rapide raisonnable. La perte liée à la question non posée, c'est-à-dire les antécédents de sinistres pour la garantie DDE est faible, et le concept de tarif rapide « à partir de » peuvent être corrigés par des coefficients de majoration ou de minoration par rapport au tarif technique selon si l'écart est positif ou négatif.

Après avoir comparé les modèles entre eux, une comparaison plus détaillée par segment de variable sera présentée.

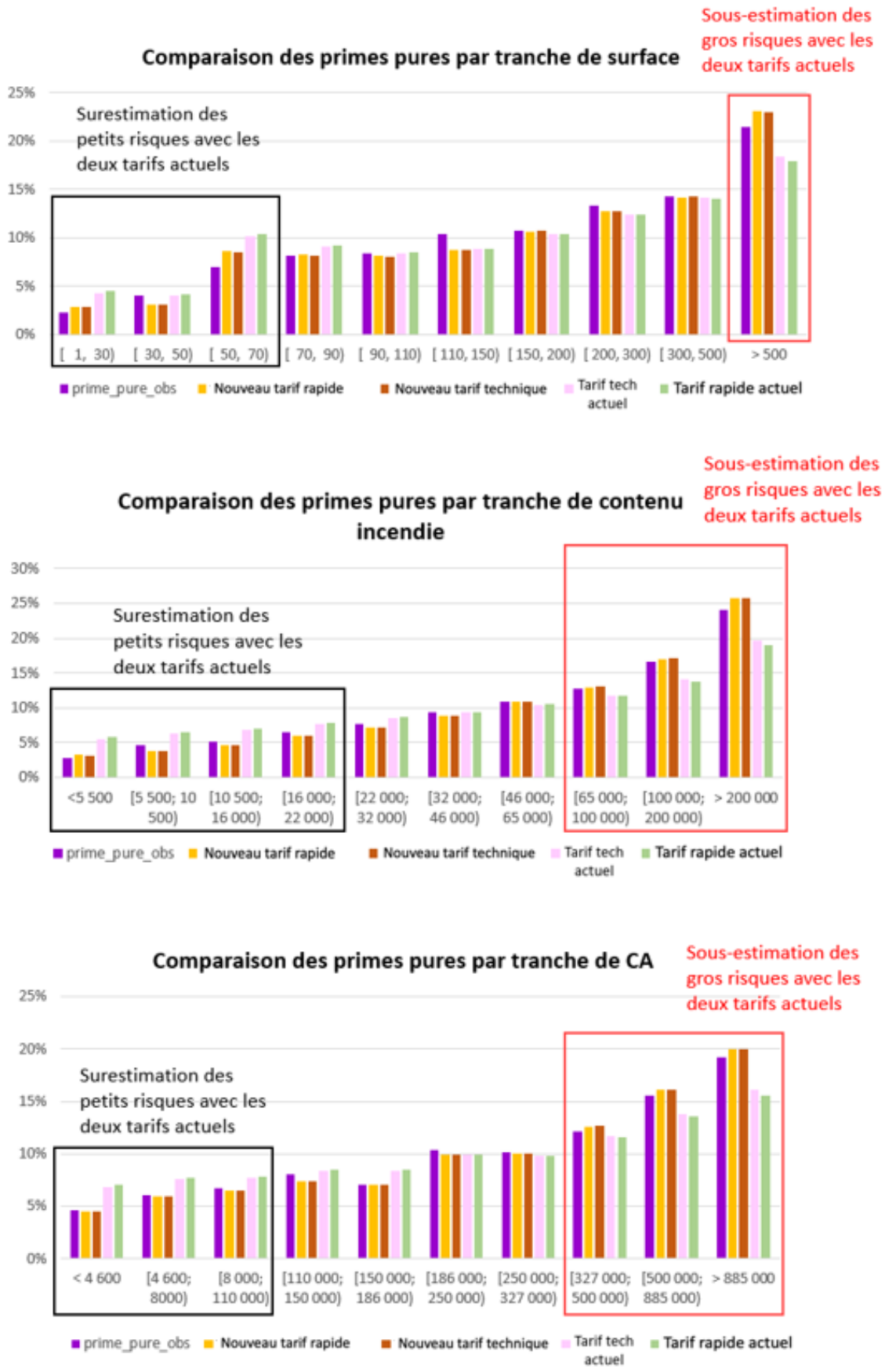


Figure 4.7.7 - Comparaison des primes pures modélisées par tranche de variable

Pour les graphiques ci-dessus, la segmentation des variables a été faite par quantile pour simplifier la visualisation et l'interprétation. En comparant les primes pures sur les trois variables explicatives ordinales, les conclusions sont identiques et confirment les précédentes interprétations. Les modèles actuels (représentés en rose pour le tarif technique actuel et en vert pour le tarif rapide actuel) surestiment la sinistralité pour les petits risques et la sous-estiment pour les gros risques. Pour les risques moyens, les prédictions des modèles sont à peu près équivalentes. Globalement, la prime pure observée (en violet) est plus proche des nouveaux modèles (représentés en jaune pour le nouveau tarif rapide et en marron pour le nouveau tarif technique).

Ainsi, à partir des analyses, les nouveaux modèles estiment mieux la sinistralité que les modèles actuels. De plus, les deux nouveaux modèles créés sont robustes sur la base de validation. Les modèles construits prédisent bien la prime pure sur des nouvelles données et il est possible de tarifier les affaires nouvelles avec ces modèles. Ces conclusions restent discutables car le tarif modélisé devra ensuite être lissé et passé à un tarif commercial. De plus, elles sont basées sur une seule garantie, un même raisonnement devrait être appliqué pour les autres garanties faisant intervenir les modalités de référence. Ainsi, en harmonisant toutes les garanties, le client pourra obtenir un tarif rapide pour toutes les garanties qu'il souhaite souscrire.

En effet, la prime pure, issue du modèle de tarification dans ce mémoire est une prime entièrement technique, correspondant à la tarification du risque, à savoir le montant du sinistre moyen auquel devra faire face l'assureur pour le risque (mathématiquement égal à l'espérance des pertes). La prime d'assurance, également appelée la prime commerciale est la prime réellement versée par l'assuré. Elle représente le prix que l'assuré doit payer pour bénéficier de la couverture d'assurance en cas de sinistre, elle est donc composée des chargements de gestion et d'acquisition, chargements de sécurité, taxes… La prime pure peut être modifiée selon la politique commerciale de la compagnie d'assurance permettant d'établir une stratégie commerciale face à la concurrence par exemple. Par conséquent, les assureurs sont amenés à revoir leur tarification régulièrement en apportant des correctifs. Le passage de la prime pure à la prime commerciale ne sera pas abordé dans ce mémoire mais pourra faire l'objet d'une étude ultérieure.

Ci-dessous un schéma synthétique des différentes comparaisons :

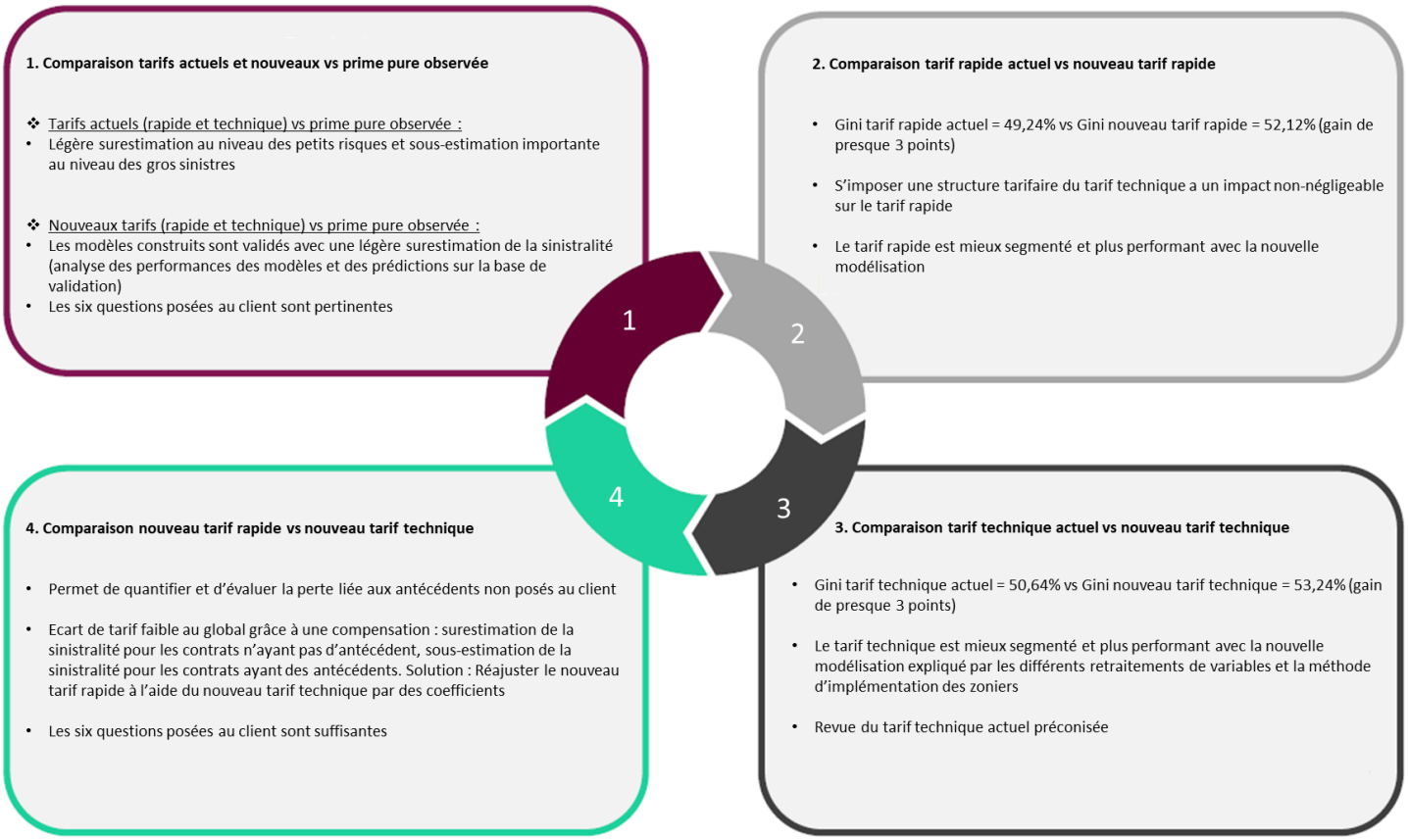


Figure 4.7.8 - Schéma synthétique des différentes comparaisons de prime pure

Conclusion

L'objectif du mémoire était de challenger le tarif rapide actuel en élaborant un nouveau tarif rapide pour la garantie DDE (garantie la plus sinistrée en termes de fréquence). En effet, le tarif rapide actuel possède deux contraintes, la première étant les six questions posées au client. La deuxième étant la contrainte de structure tarifaire technique qui implique l'utilisation de modalité de base. Une modalité de base utilisée pour toutes les variables intervenant dans le tarif technique mais non posée au client. Or, par définition, un tarif rapide n'est limité qu'au nombre de variables, ce tarif n'est donc pas optimal actuellement et possède un fort risque de sur ou sous-estimation de la sinistralité. Ainsi, le nouveau tarif rapide construit n'a pas de contrainte de structure imposée et respecte la restriction des six questions posées au client. Son développement a permis de revoir la structure tarifaire, de juger de la pertinence des six questions rapides posées au client (définies par connaissance métier), d'évaluer le risque engendré par les modalités de base, c'est-à-dire d'étudier l'impact sur le tarif rapide de s'imposer une structure tarifaire du tarif technique et de quantifier la perte liée aux questions non posées au client (cas des antécédents pour la modélisation de la garantie DDE). Par ailleurs, sa production a été utile pour améliorer la compétitivité, pour faciliter la souscription et pour obtenir une idée brève du tarif pour répondre rapidement aux besoins des clients. La création de ce tarif rapide a aussi entraîné la conception d'un nouveau tarif technique dont la structure est basée sur le nouveau tarif rapide, pour permettre de réaliser les différentes comparaisons entre l'actuel et le nouveau.

La méthode retenue pour le nouveau tarif rapide est l'approche prime pure (modèle Tweedie hors charges atypiques mais y compris franchises et mutualisation des sur-crêtes graves au prorata des primes pour les contrats sinistrés). De plus, cette modélisation est caractérisée par l'implémentation du zonier selon la méthode d'ajout simple du zonier (croisement des deux zoniers disponibles en un zonier). Les analyses graphiques sur les bases d'apprentissage et de validation ont permis d'une part de valider les modèles et d'autre part de juger de la pertinence des six questions posées au client. Ces tarifs sont considérés comme plus performants que les deux tarifs actuels. En effet, en challengeant le tarif rapide actuel et le nouveau tarif rapide, l'une des premières problématiques était d'étudier l'impact sur le tarif rapide de s'imposer la structure tarifaire du tarif technique. L'existence des modalités de base dans le tarif rapide est-elle problématique ? En comparant le tarif rapide actuel et le tarif technique actuel, l'utilisation des modalités de base entraîne une forte sous-estimation de la sinistralité. Cette sous-estimation est d'autant plus visible lorsque dès le départ le tarif technique sous-estime la sinistralité. Ainsi, les conclusions obtenues sont en faveur du nouveau tarif rapide sans contrainte de structure tarifaire. Cette sous-estimation est directement liée aux données. La modélisation actuelle ne retraite pas les valeurs manquantes ou aberrantes (valeurs saisies à zéro et correspond à une part non négligeable des données) : cas des variables chiffre d'affaires, contenus incendie et surfaces. Par exemple, un chiffre d'affaires nul est considéré comme une valeur aberrante puisque dès lors qu'une entreprise souscrit une assurance, elle exerce une activité normale et courante pour payer sa cotisation. Ne pas avoir traité les chiffres d'affaires nuls a un impact sur la modélisation actuelle car cette variable est considérée comme non tarifaire mais est réellement tarifaire si des retraitements ont été réalisés. De plus, l'écart peut également être expliqué par la méthode d'implémentation du zonier. La modélisation actuelle repose sur la méthode off set et la nouvelle modélisation se démarque avec le croisement de zonier en une unique variable et modélisation. La qualité des données a donc une contribution importante dans la modélisation. Les variables discriminantes retenues dans le nouveau tarif rapide sont l'activité, le zonier, la surface, le contenu incendie et le chiffre d'affaires.

Le deuxième enjeu lié à ce sujet était de quantifier et d'évaluer la perte liée aux questions non posées mais intervenant dans le tarif technique. Est-ce une plus-value de les ajouter parmi les questions posées au client ? Dans le cadre de la modélisation de la garantie DDE, la variable étudiée est les antécédents de sinistres. Bien que les antécédents expliquent un signal important, sa suppression permet tout de même d'obtenir un tarif raisonnable proche de la prime pure observée. La comparaison entre la prime pure du nouveau tarif rapide actuel et du nouveau tarif technique a permis de conclure que l'écart est faible, donc la perte engendrée par cette question non posée est faible. Cela permet de déduire également que les six questions sont suffisantes pour établir un tarif rapide de la garantie DDE.

Finalement, en élaborant ce tarif rapide, nous avons soulevé un problème dans le tarif rapide actuel : le chiffre d'affaires n'apparaissait pas comme tarifant dans le tarif technique mais est bien tarifaire dans le tarif nouveau tarif rapide. Ceci peut être expliqué par un travail important de qualité de données, notamment les retraitements des chiffres d'affaires nuls qui n'ont pas été traité dans la modélisation de 2019 du tarif technique. Par ailleurs, les deux tarifs actuels surestiment la sinistralité des petits sinistres et sous-estiment la sinistralité des gros risques. Ces observations amènent à revoir le tarif technique actuel. Le nouveau tarif technique modélisé pourrait être une proposition de nouveau tarif technique à implémenter (après lissage et passage de la prime pure à la prime commerciale) puisque ce modèle approchait très bien la sinistralité.

Limites de l'étude : Le tarif modélisé concerne deux marchés Particuliers-Professionnels (PP) et Entreprises (EN). Le marché Entreprise est caractérisé par une faible volumétrie de données mais un niveau de risque relativement plus élevé en termes de charge de sinistres. Or, le modèle de tarification construit est basé sur 93% de contrats PP et uniquement 7% de contrats EN. Cette différence de volumétrie peut biaiser les résultats côté Entreprise. Il est néanmoins possible d'établir des coefficients marchés pour ajuster, lisser le tarif modélisé. Cela pourrait faire l'objet d'un développement ultérieur. Une autre limite peut être mentionnée, le seuil grave a été revu dans le cadre de la modélisation de la garantie DDE et il aurait été également judicieux de revoir le seuil atypique par produit.

Améliorations futures : Plusieurs points d'améliorations peuvent être proposés, notamment de modéliser la charge hors franchise. La modélisation de la charge y compris franchise permet de mieux refléter le risque de l'assuré et de rendre les sinistres comparables, mais cette dernière sera gonflée. Des modèles de troncature à gauche peuvent être mis en place pour les petits sinistres dont la charge n'a pas atteint le montant de la franchise (correspond à une perte de données). La modélisation construite ne tient pas en compte de ce problème, donc les petits sinistres seront moins bien modélisés mais cela peut être nuancé par la modélisation de la prime pure et de la grande quantité de données à disposition (plus de trois millions de données pour neuf ans d'observation). De même, un modèle de troncature pour les antécédents de sinistres aurait été idéal car notre modèle ne prend pas en compte les antécédents des précédents assureurs. Il s'agit donc d'une perte d'information car elle est non observable. De plus, le nouveau tarif technique est basé sur la structure du nouveau tarif rapide, c'est-à-dire les six questions posées au client. Il serait plus judicieux de ne pas conditionner les variables en entrée du modèle aux questions uniquement. Enfin, des méthodes statistiques ou machine learning seraient appréciables dans la détection, implémentation des valeurs manquantes et aberrantes, segmentation des variables ou pour challenger les GLM (Random Forest, MissForest, Classification ascendante hiérarchique CAH, arbres de régression, …)

Table des figures

Liste des tableaux

Bibliographie

AXA France (2019) Les assurances des biens. *Conditions générales*

BERSON E. (2020) Refonte de la garantie Responsabilité Civile Automobile du produit Garages. *Mémoire d'actuariat*

BERTRAND F. et MAUMY Myriam (2008) Choix du modèle. *Cours IRMA, Université Louis Pasteur Strasbourg*

CHARPENTIER A. et DENUIT M. (2005) Mathématiques de l'assurance non-vie : Tome 2, Tarification et provisionnement

Fédération Française de l'Assurance (2021) L'assurance des biens de l'entreprise. https://www.ffa-assurance.fr/infos-assures/assurance-des-biens-de-entreprise

Fédération Française de l'Assurance (2021) Cotisation d'assurance : paiement et évolution du tarif. https://www.ffa-assurance.fr/infos-assures/cotisation-assurance-paiement-et-evolution-du-tarif

GUILLOU A. et You Alexandre (2011) Introduction à la théorie des valeurs extrêmes : Applications en Actuariat. *Cours université de Strasbourg*

KIBALA KUMA J. (2019) Estimation par la méthode du Maximum de Vraisemblance : Eléments de Théorie et pratiques sur Logiciel. https://hal.archives-ouvertes.fr/cel-02189969/document

KRANZLIN Sophie (2017) Modélisation du risque géographique en assurance automobile. *Mémoire d'actuariat*

KRATZ M. (2021) Extreme Value Theory. Theory and Application to Risk Management. *Cours ISUP Master 2 Actuariat*

LE TUAN A. (2017) Les méthodes de provisionnement en assurance non-vie. *Mémoire d'actuariat*

LOPEZ O. (2021) Modèles de durée. *Cours ISUP Master 2 Actuariat*

MAUD T. (2020) Econométrie de l'assurance non-vie. *Cours ISUP Master 2 Actuariat*

Sénat (2021) Les produits d'assurance. https://www.senat.fr/rap/r98-0452/r98-0452131.html

THUILLIER M. (2021) Calcul de la valeur contrat sur la branche Multirisque Immeuble comme aide opérationnelle à la relation client. *Mémoire d'actuariat*

TOESCA R. (2019) Tarification de la garantie incendie en Dommages Aux Biens – Entreprises. *Mémoire d'actuariat*

Wikistat (2021) Introduction au modèle linéaire général. https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-modlin-mlg.pdf

Annexes

A. Troncature

Il y a troncature lorsque la variable d'intérêt n'est prise en considération que sur une partie de la durée des observations.
Si la partie est connexe, trois possibilités se présentent :

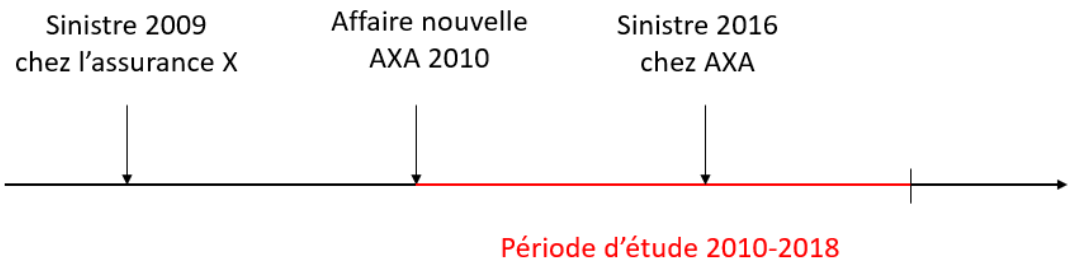
- Troncature à gauche : toutes les observations inférieures à une valeur c sont ignorées ;
- Troncature à droite : toutes les observations supérieures à une valeur C sont ignorées ;
- Troncature à gauche et à droite : toutes les observations inférieures à une valeur c ou supérieures à une valeur C sont ignorées.

Si la partie est non connexe, la troncature est par intervalles.
La fonction de survie s'écrit :

$$S(t|c \leq t < C) = \begin{cases} 1 & \text{si t} < \text{c} \\ \dfrac{\text{S(t)} - \text{S(C)}}{\text{S(c)} - \text{S(C)}} & si\ c \leq t \leq C \\ 0 & \text{si t} \geq \text{C} \end{cases}$$

Dans ce mémoire deux cas de troncatures à gauche sont remarquées :

- Cas des antécédents de sinistres, où il y a un manque d'informations dans la base sinistre d'AXA sur la sinistralité passée des affaires nouvelles (antécédents ayant eu lieu chez les précédents assureurs) ;

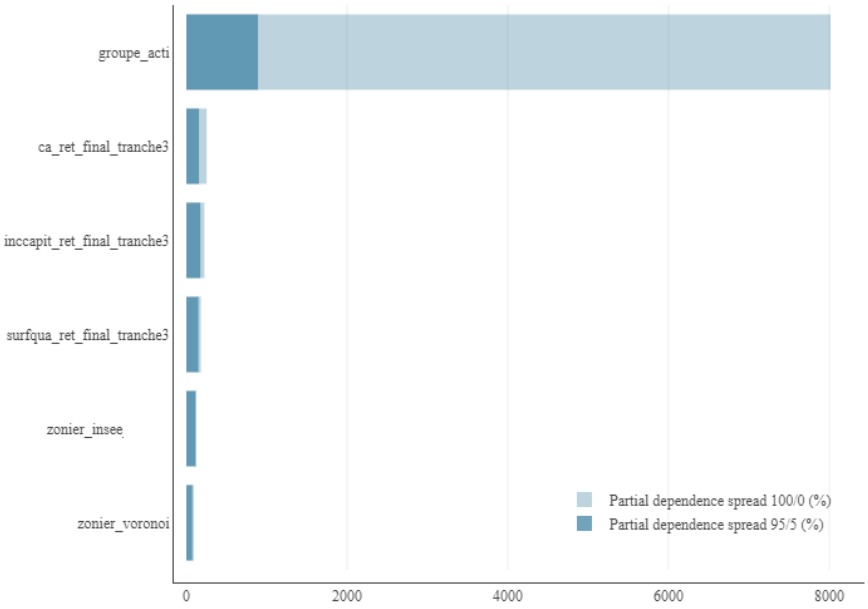


- Cas des sinistres n'excédant pas le montant de la franchise, qui correspondent aux petits sinistres à la charge de l'assuré. Les majorations dans la tarification entraînent une troncature de l'information. En effet, les assurés peuvent ne pas déclarer les sinistres inférieurs ou même supérieures à la franchise afin de ne pas voir augmenter leurs primes futures. L'assureur doit faire des hypothèses sur données qu'il n'observe pas (données tronquées à gauche de la franchise contractuelle).

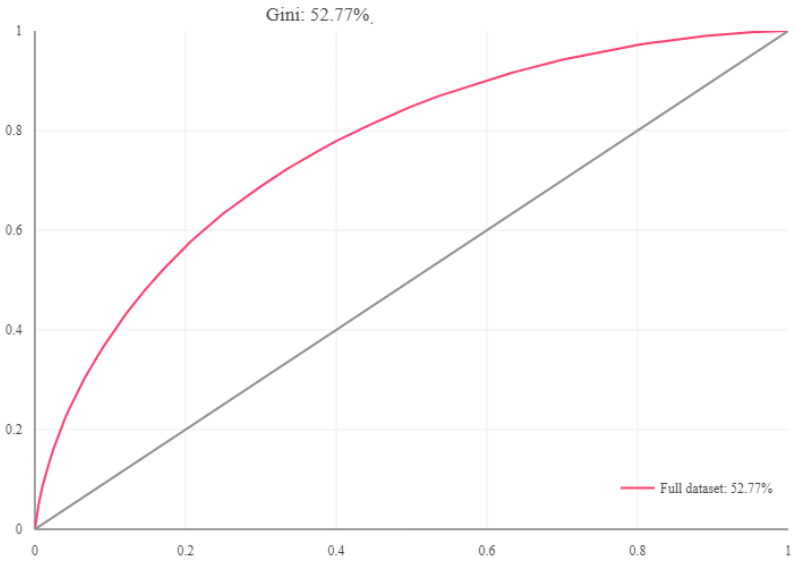
La modélisation réalisée ne tient pas en compte de ces problèmes. Une amélioration qui pourrait faire l'objet d'une étude supplémentaire serait d'appliquer des modèles de troncature.

B. Zonier méthode 2 « Off set » modèle Tweedie

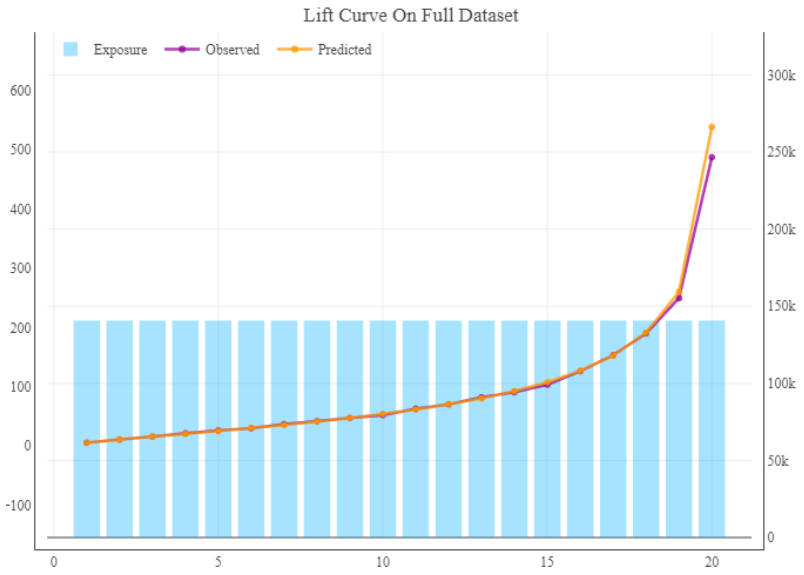
Spread :



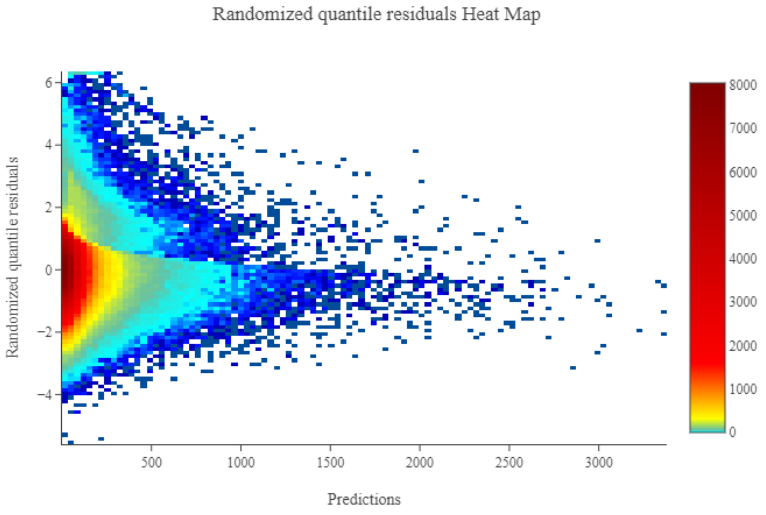
Courbe de Lorenz :



Courbe Lift :



Résidus quantiles :



Comparaison des indicateurs numériques selon la méthode d'implémentation du zonier :

	Modèle Tweedie tarif rapide	
	Zonier méthode 1	Zonier méthode 2
Gini	53,63%	52,77%
RMSE	802,7	803,2

Zonier méthode 1 : Ajout simple du zonier (Complétion du zonier Voronoi manquant par le zonier INSEE)

Zonier méthode 2 : Off set (Off set des variables tarifaires hors zonier puis calibration des zoniers séparément)