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

Par : **Monsieur Mayeul Chaput**

Titre du mémoire :
Etude de méthodes de tarification sur un portefeuille automobile

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

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

Membres présents du jury de la filière : Signature :	Entreprise : GROUPE SOLLY AZAR SAS 60 rue de la Chaussée d'Antin 75439 Paris Cedex 09 Tél. : 01 42 80 00 - Fax : 01 48 74 20 28 Siret 353 508 955 00021 APE 6727 Nom : TERKI Rezzo Signature : Directeur de mémoire en entreprise
Membres présents du jury de l'Institut des Actuaires : Signature :	Nom : KANDJI 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 :

Etude de méthodes de tarification sur un portefeuille automobile

Mayeul Chaput

27 août 2023

Table des matières

Résumé

Dans ce mémoire, nous explorons l'application des techniques d'apprentissage automatique, en particulier les réseaux de neurones en parallèle des méthodes classiques de tarification par GLM, dans le secteur de l'assurance non-vie, et plus précisément l'assurance automobile. Ce domaine est caractérisé par une forte concurrence, poussant les assureurs à proposer des tarifs économiquement viables tout en couvrant les risques de manière adéquate. Ils cherchent donc à adapter leur tarification en fonction des besoins et des caractéristiques individuelles des clients.

L'essor du numérique et la collecte d'informations génèrent une abondance de données non structurées pour les assureurs. Cependant, les méthodes classiques de tarification, telles que le GLM, ne sont pas adaptées à ce type de données en raison de leurs hypothèses statistiques complexes. Les réseaux de neurones, tels que le perceptron multicouche, sont une alternative intéressante car ils sont non-paramétriques et possèdent la capacité d'approximation universelle, permettant d'approximer n'importe quelle fonction suffisamment régulière.

Nous comparons, dans ce mémoire, les prédictions issues du GLM à celles du perceptron multicouche pour notre portefeuille d'assurance automobile. En outre, nous examinons les différentes méthodes d'optimisation disponibles pour le perceptron multicouche, afin d'améliorer davantage la précision des prédictions.

Mots-clés : Réseau de neurones, perceptron multicouche, optimisation bayésienne, optimisation PBT, Tarification, modèle linéaire généralisé, coût-fréquence, Tweedie, valeurs extrêmes, assurance automobile.

Abstract

In this thesis, we explore the application of machine learning techniques, specifically neural networks, alongside traditional GLM pricing methods in the non-life insurance sector, and more specifically, automobile insurance. This field is characterized by strong competition, pushing insurers to offer economically viable rates while adequately covering risks. They therefore seek to adapt their pricing based on the needs and individual characteristics of clients.

The rise of digital technology and information gathering generates an abundance of unstructured data for insurers. However, traditional pricing methods, such as GLM, are not suited to this type of data due to their complex statistical assumptions. Neural networks, such as the multilayer perceptron, are an interesting alternative as they are non-parametric and possess the ability to provide universal approximations, allowing them to approximate any sufficiently regular function.

In this thesis, we compare the predictions made by the GLM to those of the multilayer perceptron for our automobile insurance portfolio. Additionally, we examine the various optimization methods available for the multilayer perceptron in order to further improve the accuracy of predictions.

Key words : Neural network, Multilayer Perceptron, Bayesian optimization, PBT optimization, Pricing, Generalized Linear Model, loss-frequency, Tweedie, extreme values, car insurance.

Remerciements

Je souhaite tout d'abord remercier mon maitre d'apprentissage Mamadou Kandji pour l'encadrement effectué tout au long de l'année et plus précisement lors des différents travaux nécessaires à la réalisation de ce mémoire.

Je remercie de plus Christophe Michal, Directeur Technique chez Solly Azar pour m'avoir accueilli dans le pôle actuariat de l'entreprise.

Plus généralement je remercie toute l'équipe technique et produits avec laquelle j'ai pu travaillé durant mon alternance pour avoir su créer une atmosphère propice à la bonne humeur.

Je remercie aussi Olivier Lopez pour son cours sur les réseaux de neurones qui m'a donné envie de m'y intéresser plus profondément.

Je remercie finalement toutes les personnes qui m'ont aidé d'une manière ou d'une autre lors de cette dernière année très chargée.

1 Introduction

La tarification automobile est un secteur d'assurance très concurrentiel. Cela pousse les assureurs à proposer des primes aussi basses que possible pour constituer un portefeuille d'assurés. Néanmoins, ces primes doivent être suffisantes pour couvrir les sinistres, les commissions des courtiers, les frais de gestion, de réassurance, etc. Un autre paramètre à prendre en compte est le phénomène d'antisélection, c'est-à-dire le départ des assurés ayant peu de sinistres du portefeuille et l'arrivée de nouveaux assurés avec une sinistralité élevée. Ce problème peut survenir lorsque les primes ne sont pas suffisamment différenciées entre les profils de risque.

Plusieurs solutions existent pour atténuer ce phénomène. Une première approche consiste à proposer des tarifs bien segmentés en fonction des profils entrants, permettant de couvrir avec précision le risque du portefeuille. On parle alors de tarification à priori. Une autre méthode consiste à pénaliser progressivement les contrats ayant une forte sinistralité dans le portefeuille et, à l'inverse, à récompenser les individus non sinistrés, c'est la tarification à posteriori. Le système français Bonus-Malus en est un bon exemple.

Par souci de concision, nous nous concentrerons uniquement sur la tarification à priori. Dans cette perspective de trouver la prime adéquate pour chaque assuré, nous examinerons plusieurs méthodes de calcul de la prime pure.
Nous commencerons par les modèles linéaires généralisés (GLM), largement utilisés dans le secteur. Nous aborderons les prétraitements nécessaires, tels que l'écrêtage des sinistres graves ou la discrétisation des variables continues. Il sera également important de distinguer deux types de sinistres en fonction de leur relation avec une convention appelée IRSA.

Ensuite, nous explorerons des méthodes plus récentes, en cherchant à modéliser directement la prime pure à l'aide d'un réseau de neurones naïf, puis d'une version optimisée.

Enfin, nous comparerons les résultats des deux modèles et tirerons des conclusions sur cette étude.

2 Contexte général

2.1 Les courtiers grossistes en France

Bien que relativement méconnu du grand public, le courtier grossiste est un acteur clé du secteur de l'assurance. En effet, il est responsable de la création de produits d'assurance, qui peuvent être des produits de niche, destinés à être proposés par la suite aux distributeurs tels que les courtiers directs. Parmi ces derniers, on trouve des courtiers indépendants, des courtiers de proximité ou encore des sociétés de courtage.

Le courtier grossiste joue également un rôle d'intermédiaire entre les compagnies d'assurance et les courtiers directs qui ne trouvent pas de produits d'assurance adaptés aux besoins de leurs clients auprès des compagnies traditionnelles. Grâce à ce partenariat, le courtier grossiste obtient un mandat de souscription de la part des porteurs de risques, ce qui lui permet de proposer directement leurs offres.

Le courtier grossiste est aussi en charge de l'administration et de la gestion des contrats de ses clients. Les sinistres sont indemnisés par la compagnie d'assurance associée au produit concerné par le sinistre. Le courtier grossiste se rémunère en percevant un pourcentage de chaque prime payée par ses clients, le reste étant majoritairement reversé à la compagnie d'assurance supportant le risque.

Il est important de bien distinguer le courtier direct du courtier grossiste. En effet, le courtier direct recherche des produits d'assurance auprès des sociétés d'assurance ou des courtiers grossistes pour ensuite les proposer à ses clients, qui deviendront des assurés. De son côté, le courtier grossiste possède déjà ses propres produits ou gère ceux d'une compagnie d'assurance et les propose aux courtiers directs.

En plus de passer par un réseau de courtiers, le courtier grossiste peut également distribuer directement ses produits d'assurance sur le marché en utilisant une plateforme dédiée et spécifique à son activité. Cette approche lui permet d'élargir sa portée et de toucher directement les clients potentiels.

En plus des aspects mentionnés précédemment, il est important de souligner d'autres dimensions du métier de courtier grossiste en assurance :

1. **Expertise et innovation :** Le courtier grossiste est souvent spécialisé dans des domaines ou des niches spécifiques de l'assurance. Cette expertise lui permet de concevoir des produits innovants et adaptés aux besoins particuliers de certains segments de clientèle, qui ne sont pas toujours couverts par les offres standard des compagnies d'assurance traditionnelles.
2. **Relations avec les compagnies d'assurance :** Le courtier grossiste entretient généralement des relations étroites avec les compagnies d'assurance. Cela lui permet de négocier des conditions avantageuses pour ses clients et de bénéficier d'un accès privilégié à des produits d'assurance exclusifs. Par ailleurs, ces relations permettent également au courtier grossiste de mieux anticiper les évolutions du marché et d'adapter rapidement son offre en conséquence.

3. **Formation et conseil :** Le courtier grossiste peut également assurer un rôle de formation et de conseil auprès des courtiers directs, en les aidant à comprendre les spécificités des produits qu'il propose et en les accompagnant dans la mise en place de solutions adaptées à leurs clients. De cette manière, le courtier grossiste contribue à renforcer les compétences et l'expertise des courtiers directs, ce qui peut se traduire par une meilleure qualité de service pour les clients finaux.
4. **Gestion des risques :** Le courtier grossiste est également responsable de la gestion des risques associés à son portefeuille de produits d'assurance. Cela implique une analyse rigoureuse des profils de risque, ainsi que la mise en place de stratégies de prévention et de contrôle adaptées. Cette gestion des risques permet au courtier grossiste de garantir la rentabilité et la pérennité de son activité, tout en protégeant les intérêts des compagnies d'assurance partenaires et des clients.
5. **Adaptabilité et évolution du marché :** Le courtier grossiste doit constamment surveiller l'évolution du marché de l'assurance et des besoins des clients afin d'adapter son offre et de rester compétitif. Les nouvelles technologies, la réglementation en constante évolution et les tendances sociétales sont autant de facteurs qui peuvent impacter le marché de l'assurance et influencer la conception de nouveaux produits ou la modification de produits existants.

En somme, le courtier grossiste en assurance est un acteur clé du secteur qui apporte son expertise et sa capacité d'innovation pour concevoir des produits d'assurance adaptés aux besoins spécifiques des clients et faciliter l'accès à ces produits grâce à son réseau de courtiers directs partenaires.

Société	Siège social	Effectif salarié	Chiffre d'affaire 2019
Groupe April	69	2422	575 500 000
Entoria	92	402	147 200 000
Alptis Assurances	69	553	112 200 000
SPVie Assurances	92	228	55 200 000
Apivia Courtage	37	185	53 500 000
Solly Azar	75	325	50 000 000

TABLE 1 – Principaux courtiers grossistes en France, Argus de l'assurance

2.2 Présentation du groupe de Solly Azar

Filiale du groupe Vespieren et fondée en 1977, Solly Azar est un courtier grossiste qui gère et crée des produits atypiques en complément des offres classiques des sociétés d'assurances. En 2016, Solly Azar compte plus de 600 000 clients et se positionne comme le 2ème courtier grossiste multi-spécialiste français. Le réseau de distribution de Solly Azar est actuellement composé de plus de 400 collaborateurs ainsi que de 8 000 courtiers répartis sur l'ensemble du territoire.

Initialement spécialisé dans les produits d'assurance pour la chasse, le groupe propose désormais des offres destinées aussi bien aux particuliers qu'aux professionnels, couvrant de nombreux secteurs, allant de l'automobile à la moto, en passant par la santé, la prévoyance, les risques professionnels, l'habitation, le prêt, le voyage et même les animaux.

Les principales garanties commercialisées par le groupe sont :

– l'auto : les produits auto de Solly Azar ciblent principalement les jeunes conducteurs, les personnes malussées ou résiliées, les artisans taxi ou encore les voitures d'occasion. En 2016, Solly Azar gère plus de 118 000 contrats auto. Le produit dont est tiré notre base de données cible les personnes malussées.
– la moto : point fort de la société, le nombre de contrats moto de Solly Azar dépasse les 150 000, avec des garanties spécifiques telles que la garantie conducteur de deux-roues en utilisation saisonnière.
– l'habitation : en plus de la gamme classique des risques habitation, on peut également retrouver des offres moins ordinaires, comme par exemple l'assurance loyers impayés ou des contrats multi-risques habitation haut de gamme, incluant des garanties spécifiques pour les propriétaires de châteaux.
– enfin, les derniers produits proposés par le groupe sont des complémentaires et surcomplémentaires santé, ainsi que plusieurs produits de niche tels que l'assurance chien/chat, l'assurance emprunteur, l'assurance voyage, etc.

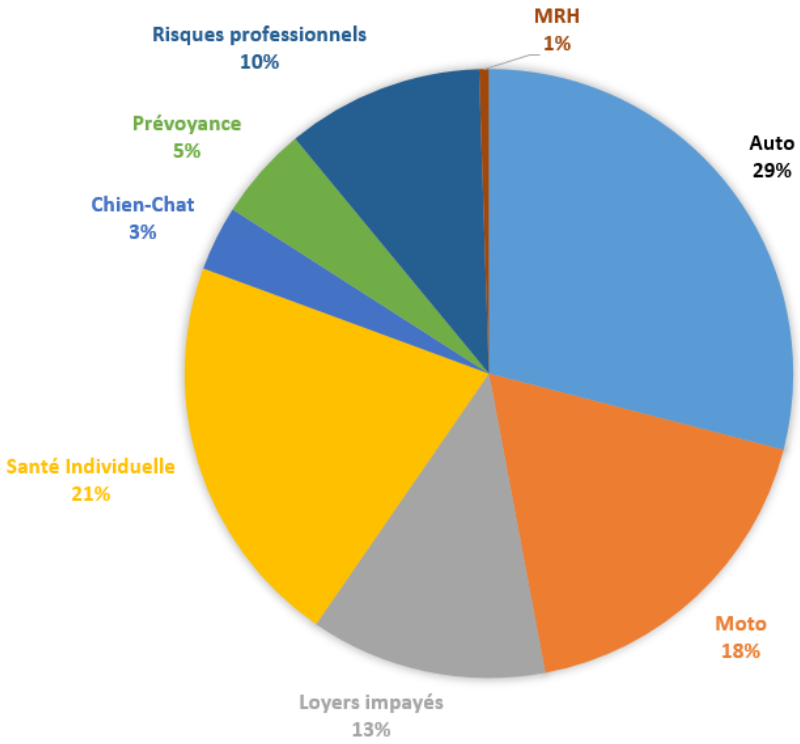

FIGURE 1 – Principaux produits gérés par Solly Azar par chiffre d'affaire

2.3 Le produit Auto 4 roues R.A

Au cours de ce mémoire, nous nous appuierons sur la base de données issue du produit Terminus de Solly Azar. Ce produit, créé en 2009, s'adresse aux conducteurs malussés. La compagnie en charge du risque associé à ce produit est l'Équité, filiale d'assurance dommages de Generali France.

Pour être éligible au produit Terminus, le conducteur doit présenter au moins l'un des critères d'aggravation suivants :

— un CRM supérieur à 0,85 ;

— une résiliation d'assurance non amiable survenue au cours des 36 derniers mois ;

— 3 sinistres ou plus survenus au cours des 36 derniers mois ;

— 2 sinistres matériels responsables ou plus survenus au cours des 36 derniers mois ;

— 1 sinistre corporel responsable survenu au cours des 36 derniers mois ;

— un antécédent d'aggravation survenu au cours des 60 derniers mois (suspension ou annulation de permis, alcoolémie, stupéfiants).

Année	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Exposition	1 565	9 412	17 742	27 117	30 353	30 268	28 721	27 848	29 841	35 931	41 306

TABLE 2 – Exposition du produit Auto 4 roues R.A depuis sa création, données anonymisées

Comme l'indique le tableau ci-dessus, le produit Terminus a connu un vif succès avec une croissance globale significative du portefeuille entre 2009 et 2019, malgré son caractère de niche et ses conditions d'éligibilité restrictives.

Pour la suite de notre mémoire, nous nous concentrerons sur la période allant de 2012 à 2019, durant laquelle la variation de l'exposition n'est pas considérable.

3 Mise en place du cadre tarifaire

3.1 Création de la base de données

Nous avons commencé avec une base de données contenant tous les sinistres de 2009 à 2019, ainsi qu'une autre base de données contenant toutes les informations sur les assurés. Nous avons d'abord fusionné ces deux bases pour obtenir une base de données unique, contenant à la fois les informations sur les assurés et les informations sur leurs sinistres potentiels.

Pour créer la base de données qui servira à modéliser notre prime pure, nous avons fait plusieurs choix arbitraires, motivés par des raisons de simplicité ou de précision dans la prédiction de notre prime pure :

— **Choix 1** : Comme mentionné précédemment, nous avons seulement pris en compte les années 2012 à 2019, car nous considérons que les trois premières années du produit Auto 4 roues R.A peuvent ne pas être significatives en termes d'année-risque et du profil des assurés entrants.

— **Choix 2** : Nous avons fait une hypothèse forte d'indépendance entre les différentes années, dans le but de regrouper toutes les années en une seule et ainsi augmenter "artificiellement" notre base de données. Grâce à cela, nous avons obtenu une base de données de 318 000 lignes.

— **Choix 3** : Nous avons choisi de ne prendre en compte qu'une seule garantie dans notre étude. En effet, la plupart des garanties, comme la garantie bris de glace, nécessitent chacune leur propre modèle, et pour des raisons de concision, nous ne voulons pas nous disperser dans trop de calculs. Nous avons donc naturellement

choisi la garantie responsabilité civile, car elle est la plus présente dans notre base. Cette garantie se compose de deux parties : la RC corporelle et la RC matérielle, que nous avons fusionnées en une seule.

3.2 Convention IRSA

Avant d'entamer l'indexation du coût des sinistres, il est important de se pencher sur une convention essentielle qui entre en jeu lors d'un sinistre auto : la convention IRSA (Indemnisation Directe de l'Assuré et Recours entre Sociétés d'Assurance Automobile).

Cette convention permet de faciliter et fluidifier l'indemnisation des dommages matériels en cas d'accident. Elle concerne les accidents impliquant au moins deux véhicules terrestres à moteur assurés et peut s'appliquer en France comme à l'étranger, mais seulement si la compagnie d'assurance a accepté cette convention dans sa politique de remboursement. Cela est généralement le cas, car elle permet de réduire considérablement les frais de gestion des sinistres.

Une fois qu'une expertise a été effectuée sur les dommages du sinistre, l'assuré est indemnisé par son assureur en fonction de la responsabilité de l'automobiliste adverse, fixée par les règles de droit commun du constat amiable et des pièces explicatives de l'accident. Il est important de noter que l'assuré est toujours indemnisé par son assureur, quelle que soit la responsabilité du sinistre, et que l'assureur se retourne ensuite vers l'assureur adverse pour recevoir son indemnisation forfaitaire ou réelle. Plusieurs cas de figure sont possibles dans ce contexte :

— Soit le coût du sinistre est inférieur à 6 500€, et dans ce cas, le recours est forfaitaire jusqu'à 1 678€ (montant IRSA 2021).

— Soit le coût du sinistre est supérieur à 6 500€, et dans ce cas, le recours est réel, ce qui signifie que l'assureur adverse rembourse le coût exact du sinistre.

La convention IRSA permet donc une gestion plus efficace et rapide des indemnisations, en simplifiant les procédures et en réduisant les délais de règlement des sinistres.

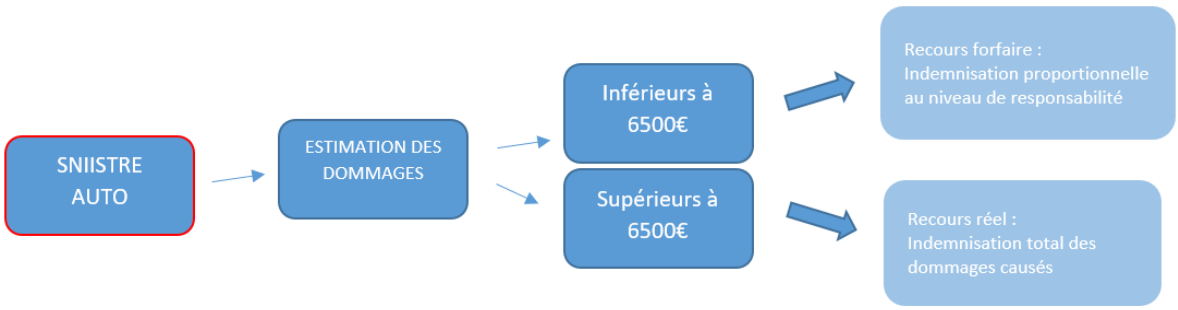

FIGURE 2 – Principe de la convention IRSA

En examinant notre base de données et en observant le nombre de sinistres en fonction du coût des sinistres (Figure 3), nous remarquons effectivement une accumulation d'un grand nombre de sinistres sur certaines valeurs. Ceci traduit bien l'application de cette convention à recours forfaitaire. Les différents pics correspondent aux différentes valeurs du forfait, qui changent chaque année.

Nous verrons ultérieurement que cette convention induit une forme anormale dans la répartition des coûts des sinistres, pouvant perturber la modélisation de la prime pure. Il est donc nécessaire de traiter ces sinistres à part. Il est également important de noter que les sinistres ayant fait fonctionner la convention IRSA et ayant généré une plus-value pour l'assureur ont été supprimés de la base (cas où le sinistre est non responsable et où le coût est inférieur au forfait IRSA).

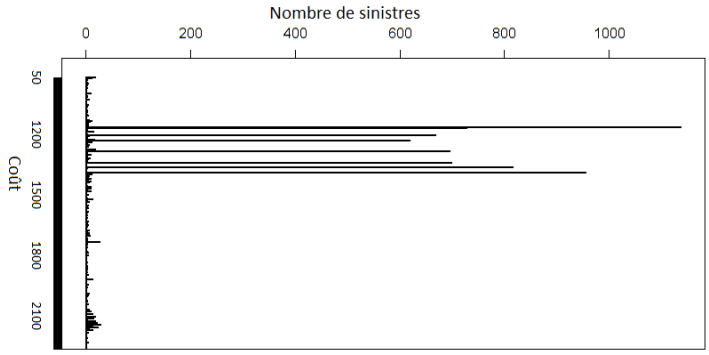

FIGURE 3 – Nombre de sinistres en fonction du coût

3.3 Indexation des coûts des sinistres en tenant compte des variations annuelles

L'indexation des coûts des sinistres est une étape cruciale de la modélisation car elle permet de rendre cohérente notre base de données en combinant les sinistres de toutes les années choisies et en fournissant une prime pure ajustée aux sinistres de l'année en cours. Elle permet de plus de ne pas laiser les sinistres plus anciens avec des coûts moyens plus faibles au profil Cette indexation prend en compte les variations annuelles des montants forfaitaires et des taux d'inflation. Nous avons choisi pour indexer nos coûts sinistres non forfaitaires l'inflation des prix à la consommation.

Dans notre étude, nous cherchons à ramener tous nos sinistres à l'année de tarification 2021. Pour ce faire, nous devons séparer nos sinistres en deux parties, en fonction de leur appartenance ou non au forfait IRSA, et indexer chacune de ces subdivisions.

3.3.1 Indexation des sinistres forfaitaires IRSA

Pour les sinistres relevant du forfait IRSA, nous nous basons sur l'historique des montants forfaitaires de la convention en fonction des années, comme indiqué dans le tableau ci-dessous :

Année	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Montant	1236€	1242€	1276€	1308€	1354€	1420€	1446€	1482€	1568€	1678€

Table 3 – Historique des montants forfaitaires IRSA, Argus de l'assurance

Ces montants nous permettent de calculer les augmentations annuelles suivantes :

Année	2012	2013	2014	2015	2016	2017	2018	2019	2020
Augmentation	35,8%	35,1%	31,5%	28,3%	23,9%	18,1%	16,0%	13,2%	7,0%

Table 4 – Augmentation du forfait IRSA en pourcentage par année

Pour les sinistres non concernés par la convention IRSA, l'indexation est basée sur le taux d'inflation en France depuis 2012 jusqu'à 2021. Cette méthode permet d'obtenir les augmentations de coûts des sinistres suivantes :

Année	2012	2013	2014	2015	2016	2017	2018	2019	2020
Augmentation	11,0%	7,5%	6,1%	5,3%	5,3%	4,9%	3,6%	1,5%	0,5%

Table 5 – Taux d'inflation par rapport à l'année 2021, INSEE

Cette indexation des sinistres permet d'obtenir une répartition plus homogène des coûts, avec un seul pic de concentration à 1678€. Ce montant correspond à la valeur regroupée des sinistres relevant de la convention IRSA, facilitant ainsi leur isolation pour une modélisation ultérieure de la prime pure. L'homogénéisation des données de sinistres contribue à une meilleure estimation des coûts futurs et permet d'adapter la tarification en conséquence.

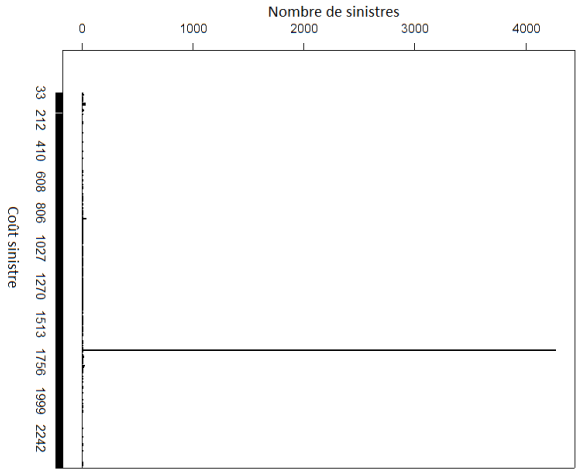

FIGURE 4 – Nombre de sinistres en fonction du coût

3.4 Présentation des variables

Avant de commencer la modélisation coût × fréquence, il est important de bien connaître la base sur laquelle on travaille. Prendre le temps de décrire en détail les différentes variables tarifaires et analyser la composition du portefeuille est donc obligatoire.

L'estimation de la prime pure est souvent liée aux caractéristiques du véhicule assuré, à son usage, à la situation géographique et aux caractéristiques du conducteur. La particularité du produit Auto 4 roues R.A est qu'il comporte, en plus des critères classiques, des variables originales, comme la résiliation antérieure pour pièces manquantes. Ci-dessous se trouve la liste des variables explicatives à notre disposition :

Variables catégorielles :

1. **ANT_ASSURANCE** : si le conducteur a déjà eu une assurance par le passé ;
2. **FORMULE** : quelle couverture de risque le conducteur a-t-il choisi pour s'assurer ;
3. **FRACTIONNEMENT** : si le conducteur paie mensuellement, annuellement, etc ;
4. **VEH_TARIF_ZONE** : la zone dans laquelle se trouve le conducteur ;
5. **RESIL_N_PAIEMENT** : si le conducteur a déjà été résilié pour non-paiement ;
6. **RESIL_PMQ** : si le conducteur a déjà été résilié pour pièces manquantes ;
7. **RESIL_FREQ_SIN** : si le conducteur a déjà été résilié pour fréquence sinistre trop élevée ;
8. **USAGE** : quelle utilisation de son véhicule le conducteur prévoit de faire ;
9. **CLASSE** : dans quelle classe est situé le véhicule ;
10. **GROUPE** : dans quel groupe est situé le véhicule ;
11. **CRM** : quel est le CRM du conducteur au moment de la souscription ;
12. **NBRE_SIN_DECLAR** : le nombre de sinistres déclarés du conducteur à la souscription ;

13. **DUREE_INT_ASS** : la durée d'interruption d'assurance du conducteur à la souscription ;
14. **EXPOSITION** : le nombre de jours ramené sur une année où le contrat a été en vigueur ;

Variables continues :

1. **AGE_COND** : L'âge du conducteur au moment de la souscription du contrat d'assurance.
2. **AGE_VEH** : L'âge du véhicule au moment de la souscription du contrat d'assurance.
3. **ANC_PERMIS** : L'ancienneté de détention du permis de conduire par le conducteur au moment de la souscription.

De plus, notre étude se concentre sur les variables cibles suivantes :

1. **COUT_TOT** : Le coût total du sinistre concerné.
2. **NBRE_SIN** : Le nombre de sinistres qu'a eu l'assuré durant l'année d'observation.

Il est important de noter que nous n'avons pas besoin de créer un zonier et un véhicule dans cette étude, car les variables **CLASSE**, **GROUPE** et **VEH_TARIF_ZONE** sont déjà établies et fournissent les informations nécessaires sur les zones géographiques et les catégories de véhicules.

3.4.1 Discrétisation des variables continues

Les trois variables continues de notre base de données, **AGE_COND**, **AGE_VEH** et **ANC_PERMIS**, doivent être transformées en variables catégorielles pour pouvoir être exploitées dans les modèles linéaires généralisés (GLM) que nous utiliserons par la suite.

Il existe dans la littérature scientifique un grand nombre de méthodes de discrétisation. Parmi les méthodes les plus connues, on peut citer l'algorithme de classification ascendante hiérarchique, l'algorithme des k-means, et l'algorithme descendant avec utilisation d'un arbre de régression.
Dans cette étude, nous avons choisi la méthode des seuils naturels de Jenks, également connue sous le nom de méthode des ruptures naturelles. Cette méthode est classiquement utilisée en géographie pour trouver la meilleure répartition en catégories d'une densité de population à présenter sur une carte.
L'algorithme de Jenks est particulièrement adapté à notre contexte car il permet de minimiser la variance intra-classe et de maximiser la variance inter-classes, ce qui facilite l'interprétation des résultats et améliore la performance des modèles GLM.

Principe de l'algorithme :
L'algorithme de Jenks est un processus itératif qui cherche à minimiser la variance intra-classes et à maximiser la variance inter-classes pour un nombre donné de classes m. La base de données est divisée en m classes à l'aide de $m-1$ seuils choisis parmi les données.

Deux étapes sont répétées jusqu'à trouver des seuils optimaux :

1. Calcul de la somme des variances intra-classes (SV_IC) :

$$SV\_IC = \sum_{i,j \in L, i<j} \sum_{k=i}^{j} \left(Base[k] - \frac{\sum_{k=i}^{j} Base[k]}{j-i+1} \right)^2 \quad (1)$$

où L est la liste des seuils et $Base$ est la base de données contenant nos valeurs.

2. Modification des $m-1$ seuils. Il y a plusieurs façons de modifier ces seuils. La méthode la plus simple consiste à tester toutes les combinaisons possibles de seuils, mais cela peut être très coûteux en temps de calcul. Une autre méthode consiste à modifier les seuils de manière à déplacer un élément de la classe ayant la plus grande variance intra-classe vers la classe ayant la plus petite variance intra-classe.

À l'issue de ce processus, on choisit les seuils qui ont donné la plus petite variance intra-classe. On peut ensuite calculer la mesure finale, appelée "Goodness of Variance Fit" (GVF) :

$$GVF = \frac{SV\_ET - SV\_IC}{SV\_ET} \quad (2)$$

avec SV_ET la variance totale de la base :

$$SV\_ET = \sum_{k=1}^{n} \left(Base[k] - \frac{\sum_{k=1}^{n} Base[k]}{n} \right)^2 \quad (3)$$

et n la taille de la base de données.

GVF prend des valeurs entre 0 (le moins bon clustering) et 1 (le meilleur clustering possible).

Le principal inconvénient de la méthode de Jenks est qu'elle ne donne pas le nombre de classes optimal. En effet, le GVF augmente avec le nombre de classes jusqu'au point où il y a autant de classes que d'éléments dans l'échantillon.

On peut se référer aux formules usuelles pour obtenir une indication sur le nombre de classes à considérer :

— Brooks-Carruthers : $k = 5 \times \log_{10}(n)$

— Huntsberger : $k = 1 + 3.332 \times \log_{10}(n)$

— Sturges : $k = \log_2(n+1)$

— Scott : $k = \frac{\max - \min}{3.5 \times \sigma \times n^{-\frac{1}{3}}}$

— Freedman-Diaconis : $k = \frac{\max - \min}{2 \times IQ \times n^{-\frac{1}{3}}}$

On constate que, peu importe la formule, le nombre de classes est dans tous les cas trop élevé. Nous choisissons donc arbitrairement un nombre de classes égal à 5, qui correspond à la moyenne du nombre de modalités de nos variables explicatives. On applique ensuite l'algorithme de Jenks sur nos variables continues et on obtient nos variables discrétisées.

Les graphiques de répartition montrent que la tendance générale est respectée malgré la discrétisation, ce qui indique que notre choix de 5 classes est adapté pour cette analyse.

Nom	Nombre de classes k
Brooks-Carruthers	27
Huntsberger	19
Sturges	18
Scott	116
Freedman-Diaconis	144

TABLE 6 – Nombre de classes suggéré par différentes formules

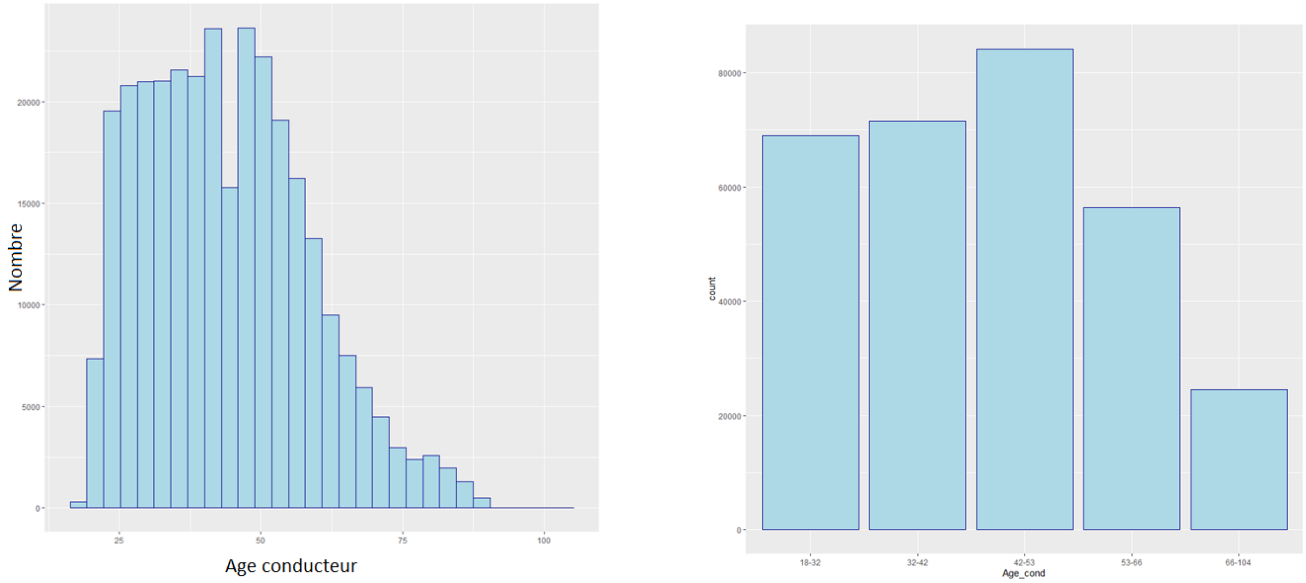

FIGURE 5 – Age conducteur non discrétisé (gauche) et discrétisé (droite)

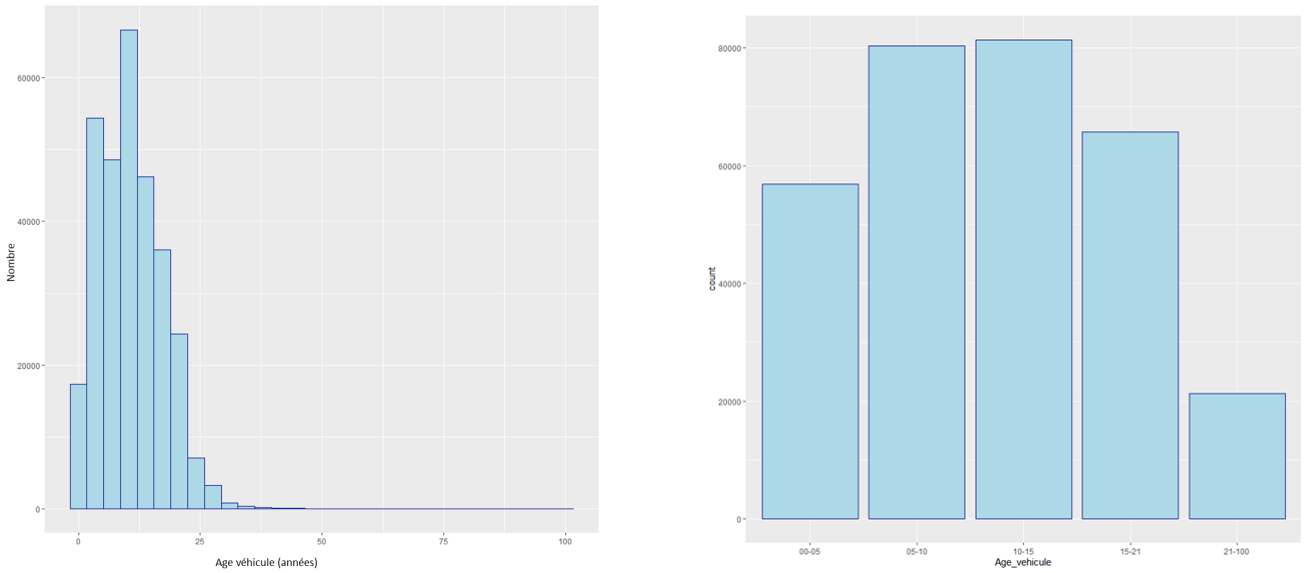

FIGURE 6 – Age véhicule non discrétisé (gauche) et discrétisé (droite)

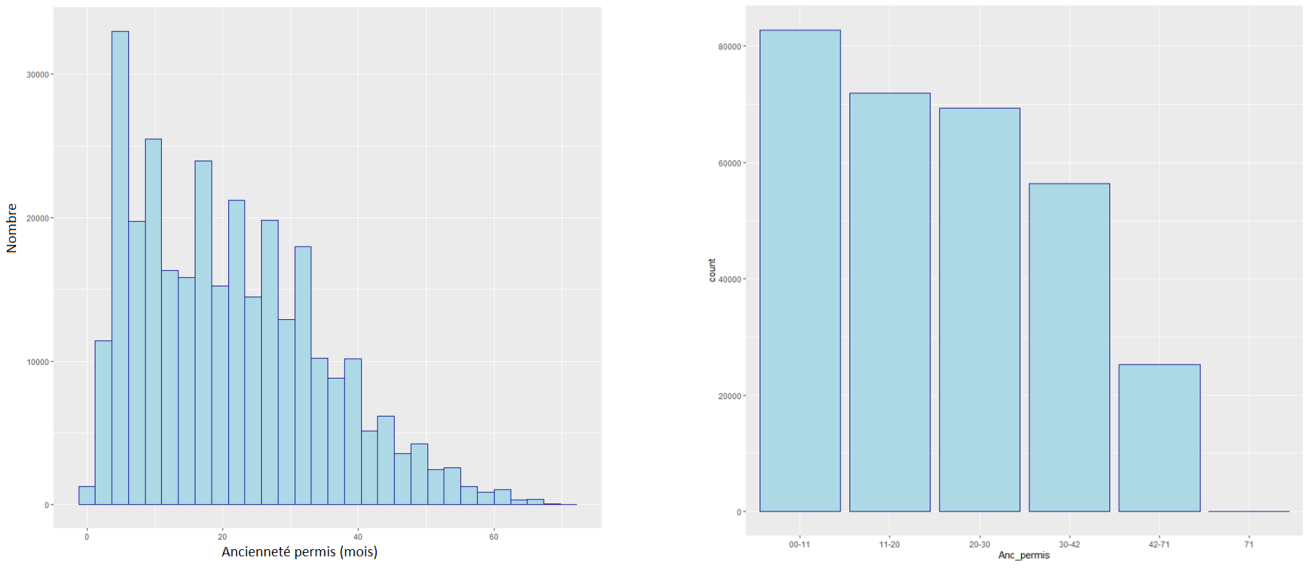

FIGURE 7 – Ancienneté permis non discrétisé (gauche) et discrétisé (droite)

3.4.2 Premières analyses graphiques de nos variables

Nous allons maintenant projeter nos variables et examiner la fréquence sinistre ainsi que le coût moyen de ces variables pour chacune de leurs modalités. Cette analyse permettra d'effectuer un premier tri sur les variables ayant peu d'influence sur la fréquence sinistre et les coûts moyens.
Il est important de considérer les graphiques des coûts moyens avec prudence, car la présence de sinistres à coûts élevés peut biaiser les résultats. Nous nous concentrerons donc principalement sur les graphiques de fréquence sinistre pour évaluer l'impact de chaque variable sur la sinistralité.

Formule : Cette variable représente le nombre de garanties souscrites par l'assuré pour se couvrir, allant de la plus basique (formule 1) à la plus complète (formule 3). Deux types de comportements sont possibles :

1. La théorie de l'aléa moral s'applique et on observe davantage de sinistres de la part des individus ayant choisi la formule 3, car ils sont confiants quant à leur niveau de couverture. En revanche, ceux ayant opté pour la formule 1 l'ont fait en se considérant moins risqués. Dans ce cas, on devrait observer une fréquence de sinistres plus faible pour la formule 1 et plus élevée pour la formule 3.
2. La souscription de garanties plus nombreuses est un signe d'aversion au risque. Ainsi, la formule 3 devrait être la moins sinistrée.

Il est clairement observable que le premier cas s'applique, avec une formule 3 présentant une fréquence de sinistres bien plus élevée que les autres formules et un coût moyen plus élevé.

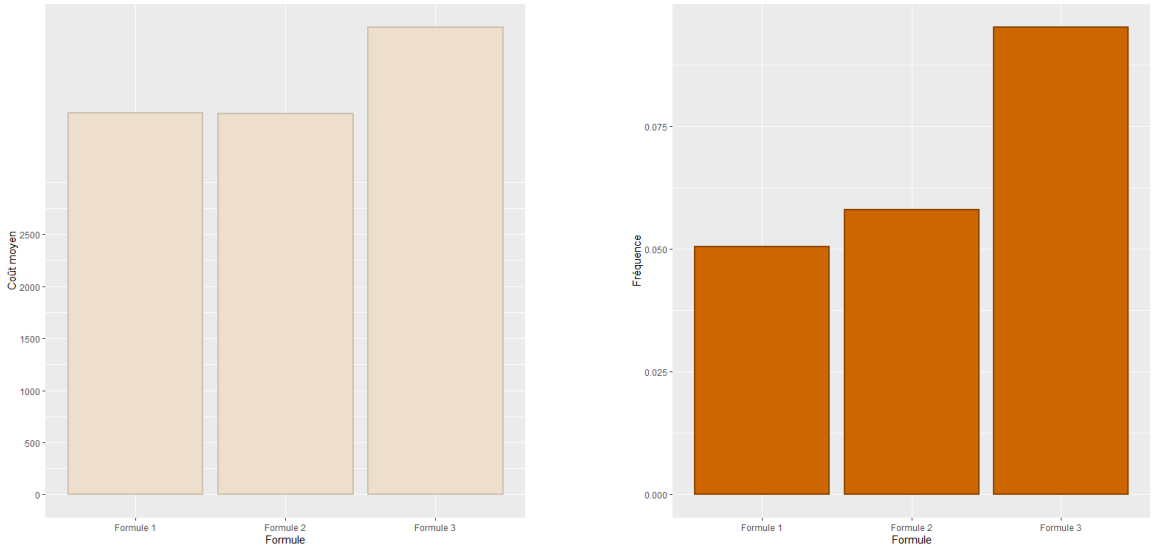

FIGURE 8 – Coût moyen et fréquence des modalités de la variable Formule

Fractionnement Le fractionnement désigne la fréquence de paiement. On peut se douter de la pertinence de cette variable à première vue mais elle peut en réalité traduire une position sociale fragilisée qui amenerait à moins entretenir son véhicule. Il faut cependant remarquer que certaines compagnies proposent de faire payer la prime mensuellement sans frais ce qui peut biaiser la variable.
On voit dans notre exemple que les paiements mensuels ont une fréquence sinitre plus élevée que les paiement annuels.

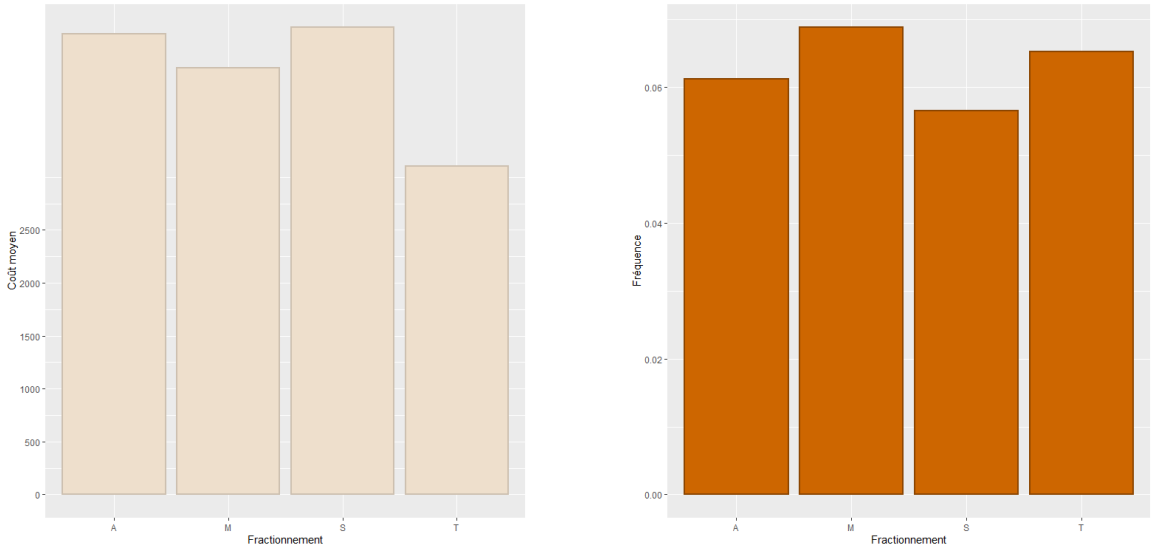

FIGURE 9 – Coût moyen et fréquence des modalités de la variable Fractionnement

Groupe Le groupe du véhicule est une variable tirée du véhiculier de la compagnie d'assurance. Elle traduit les caractéristiques du véhicule de l'assuré comme la vitesse ou encore le volume. Les premiers groupes concernent des caractéristiques "basses" (faible vitesse, puissance etc) tandis que les groupse plus élevés traduisent des caractéristiques "élevées". En théorie plus le véhicule est puissant plus il est difficile à contrôler et donc la sinsitralité sera plus élevée.
On remarque ici que la fréquence sinistre croît à mesure que les caractéristiques du véhicule augmentent ce qui confirme la théorie.

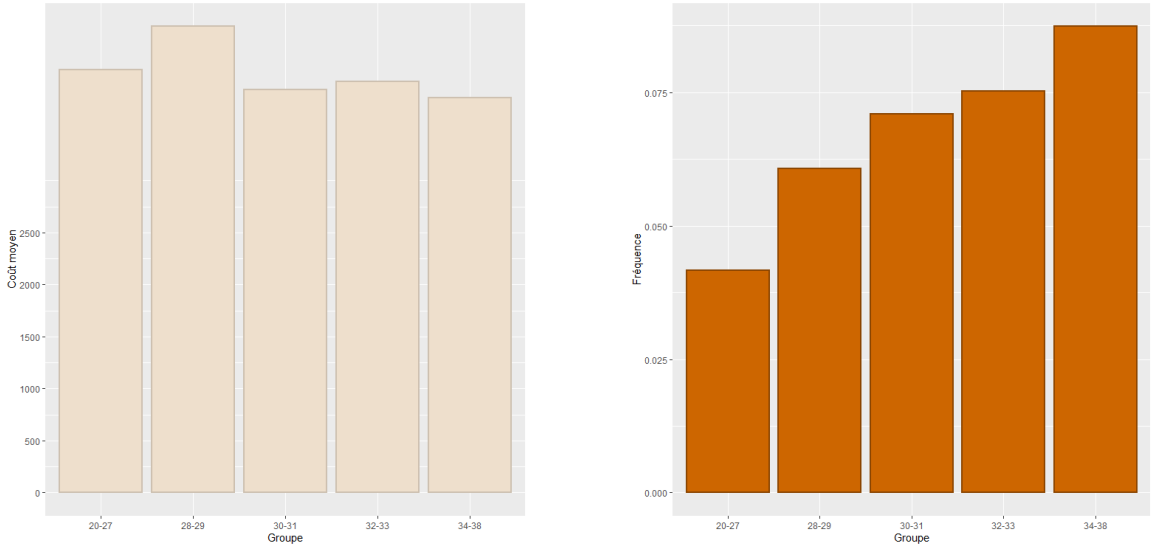

FIGURE 10 – Coût moyen et fréquence des modalités de la variable Groupe

Interruption d'assurance L'interruption d'assurance traduit le nombre de mois où l'assuré n'a plus été couvert par une assurance. On remarque qu'aucune tendance n'apparait dans la fréquence sinistre.

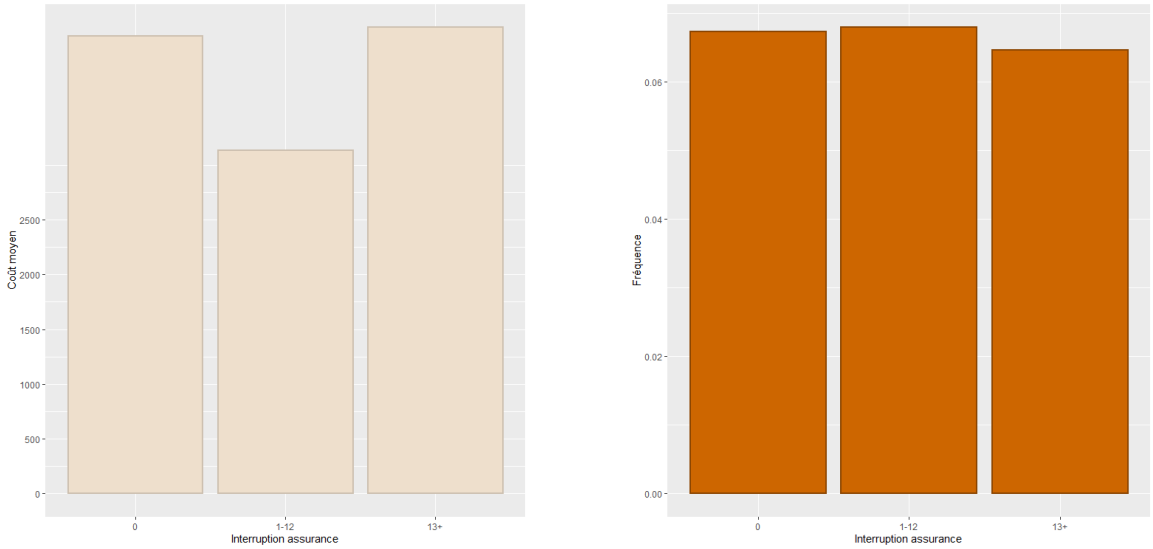

FIGURE 11 – Coût moyen et fréquence des modalités de la variable Interruption d'assurance

Antécédent assurance L'antécédent assurance est une variable bimodale qui indique si l'assuré a déjà eu une assurance auparavant. On voit que la fréquence sinistre est quasiment identique sur chaque modalité.

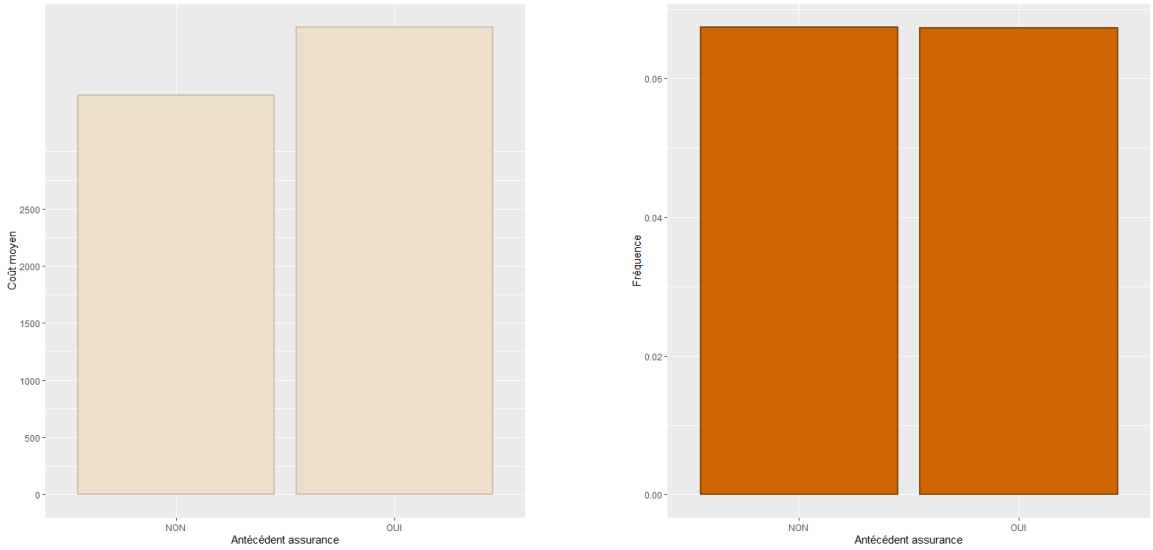

FIGURE 12 – Coût moyen et fréquence des modalités de la variable Antécédent assurance

Classe La variable classe est une variable qui trie les véhicules en fonction de leur prix à neuf. Plus un véhicule est cher à l'achat, plus sa classe sera élevée (ici en lettres alphabétiques).
On constate graphiquement que plus le véhicule est cher à l'achat plus il a une fréquence sinistre élevée. Une explication peut être proposée : plus le prix d'achat du véhicule augmente plus il a des caractéristiques élevées (comme la puissance de son moteur ou sa vitesse) et donc on retrouve l'explication sur le groupe du véhicule.

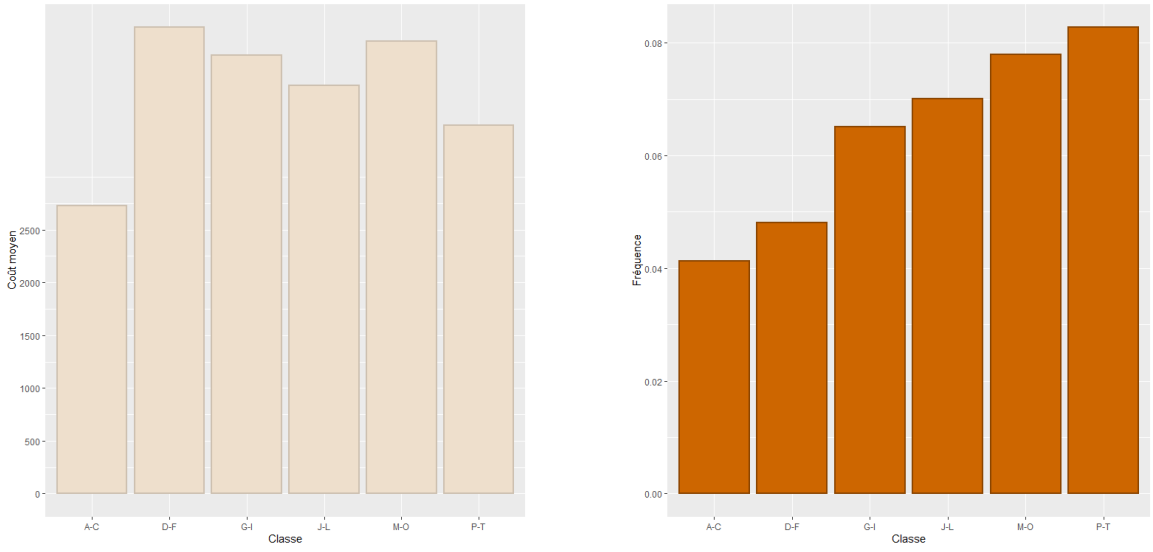

FIGURE 13 – Coût moyen et fréquence des modalités de la variable Classe assurance

CRM Le CRM est une variable qui mêle niveau d'ancienneté et sinistralité du conducteur. Plus le CRM est élevé plus le conducteur est potentiellement peu expérimenté ou plus sujets aux accidents.
Cette tendance se retrouve très bien dans le graphique de la fréquence où la fréquence sinistre augmente linéairement en fonction du CRM.

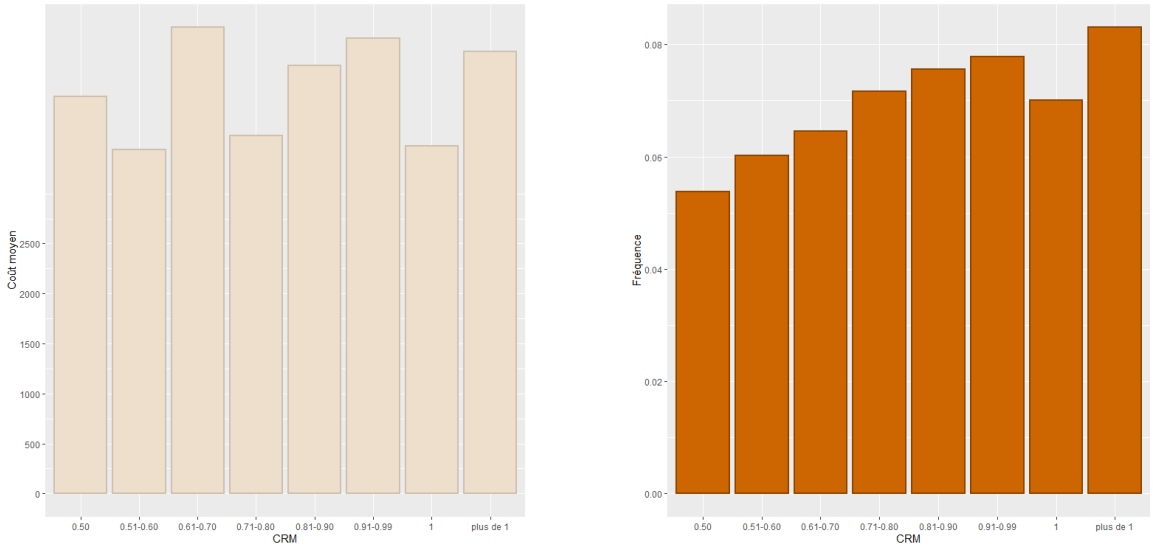

FIGURE 14 – Coût moyen et fréquence des modalités de la variable CRM

Dernier sinistre La variable Dernier sinistre compte le nombre de mois qui s'est écoulé depuis le dernier sinistre de l'assuré, avec 0 étant le fait de ne pas avoir eu de sinistre antérieur. On pourrait s'attendre à une sinistralité plus faible pour les assurés ayant eu un récent sinistre car on constante généralement que la probabilité d'avoir un sinistre quand on en a déjà eu un est faible.
On remarque qu'il n'y a pas vraiment de tendance mise en relief sur le graphique de fréquence.

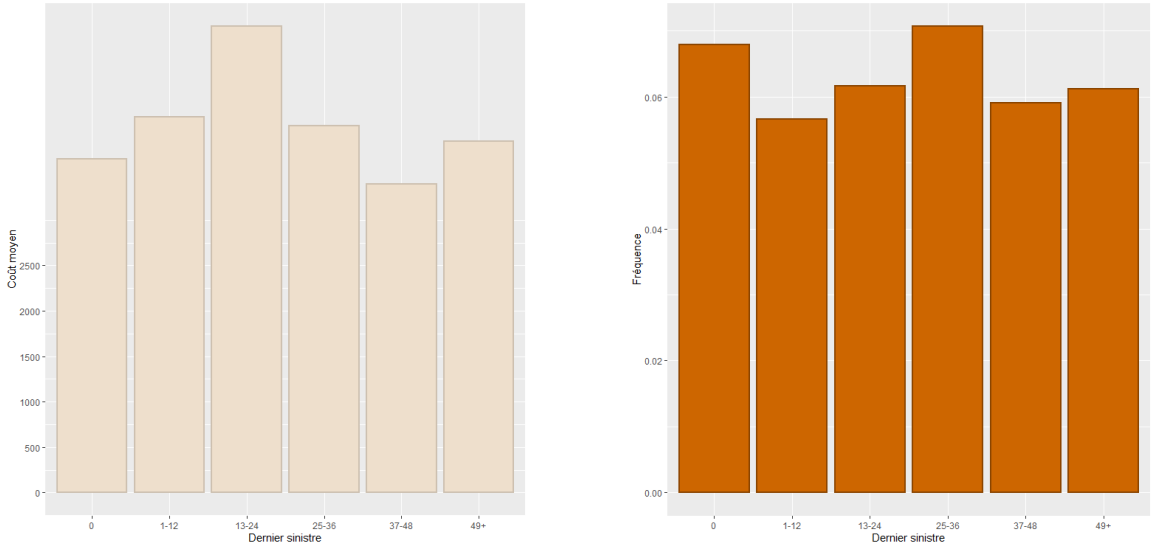

FIGURE 15 – Coût moyen et fréquence des modalités de la variable Dernier sinistre assurance

Nombre de sinistres déclarés Le nombre de sinistres déclarés est comme son nom l'indique une variable qui compte le nombre de sinistres que l'assuré a déclaré avant de souscrire. Il faut cependant remarquer que le nombre de sinistres déclarés peut être différent du nombre de sinistres réels si l'assuré choisit de ne pas déclarer un sinistre. On peut s'attendre à ce que les individus avec un nombre de sinistres antérieur élevé sont des individus naturellement propices à avoir plus d'accidents.
On remarque sur le graphique que la tendance se confirme car la fréquence sinistre augmente linéairement en fonction du nombre de sinistres passés.

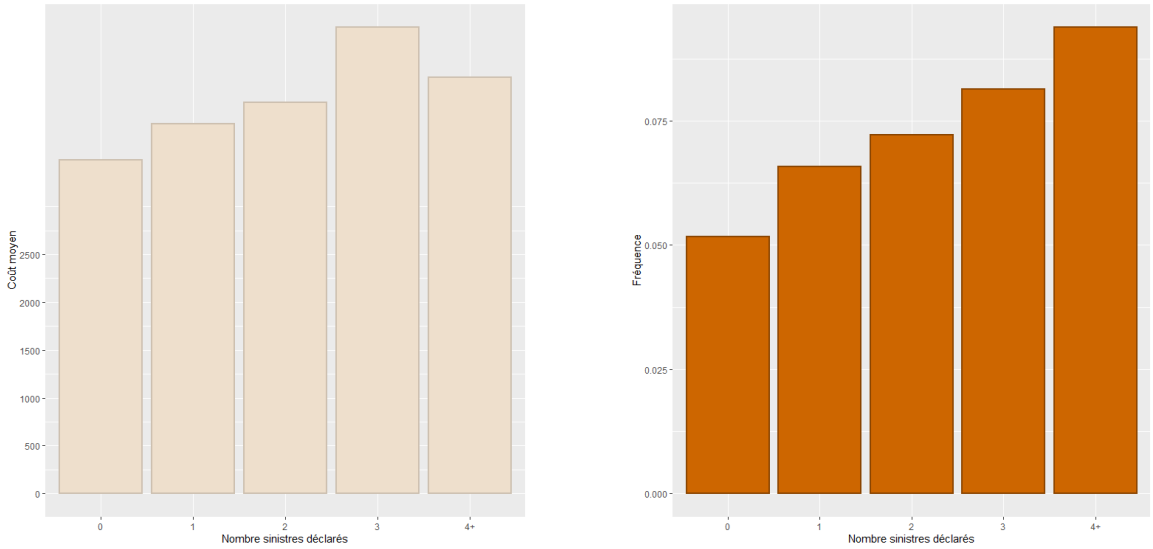

FIGURE 16 – Coût moyen et fréquence des modalités de la variable Nombre sinistres déclarés

Résiliation fréquence sinistre La variable résiliation fréquence sinistre est une variable spécifique au produit Auto 4 roues R.A. Elle indique si l'assuré s'est fait résilier chez son précédent assureur à cause d'une fréquence sinistre trop élevée.
On remarque sans surprise que ceux qui présentent la modalité "OUI" ont une fréquence sinistre plus élevée que ceux ne s'étant pas fait résilier pour fréquence sinistre.

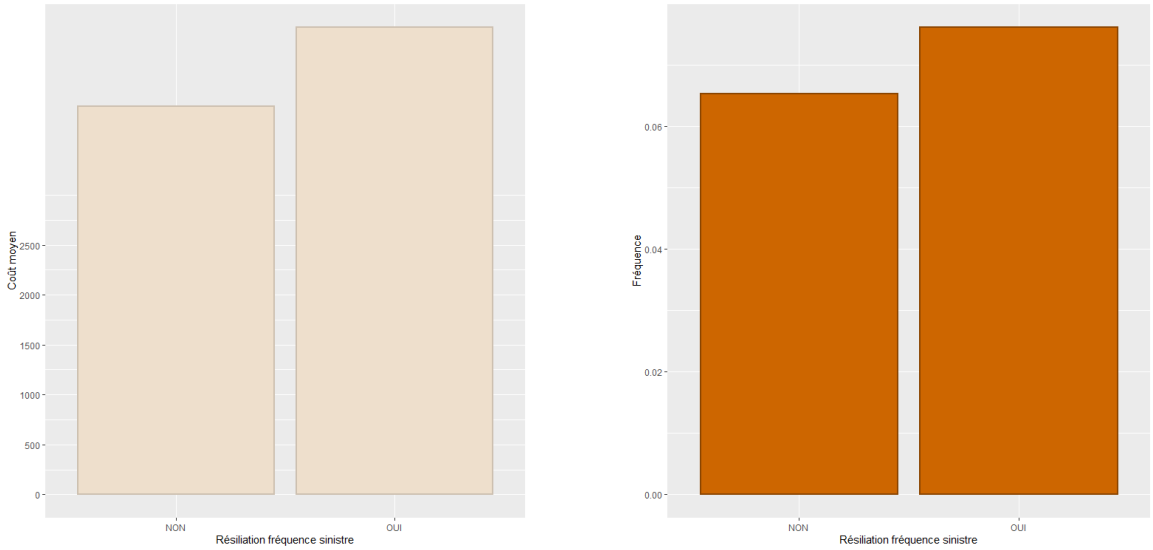

FIGURE 17 – Coût moyen et fréquence des modalités de la variable Résiliation fréquence sinistre

Résiliation non paiement Autre variable spécifique au produit Terminus, résiliation pour non paiement indique si l'assuré s'est fait résilier son précédent contrat pour non paiement de ses primes.
On remarque que la fréquence sinistre est bien plus faible sur la modalité "OUI".

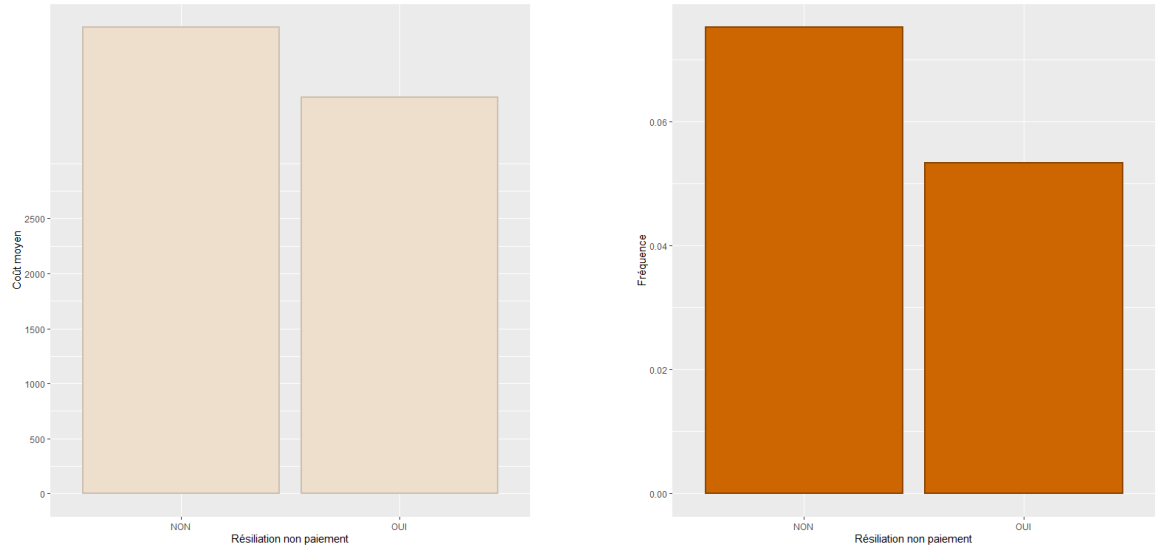

FIGURE 18 – Coût moyen et fréquence des modalités de la variable Résiliation non paiement

Résiliation pièces manquantes Résiliation pièces manquantes indique si l'assuré s'est fait résilier son précédent contrat pour non présentation de pièces nécessaires à la validation du contrat.
Aucune tendance significative ne ressort du graphique de fréquence sinistre.

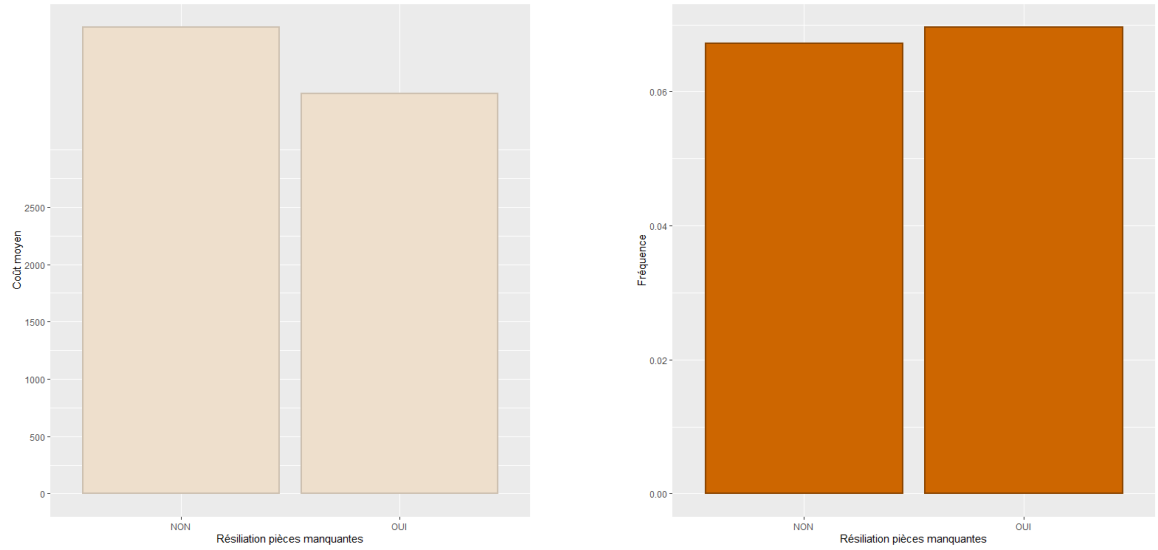

FIGURE 19 – Coût moyen et fréquence des modalités de la variable Résiliation pièces manquantes

Zone La zone est une variable qui découle du zonier établi par la compagnie d'assurance. Elle regroupe les endroits géographiques de la moins risquée (zone 1) à la plus risquée (zone 6). On retrouve généralement dans les zones très sinistrées les communes fortement urbanisées où la probablité d'avoir un accident est plus élevée bien que le coût moyen soit plus faible du fait de la vitesse limitée et au contraire dans les zones peu sinistrées qui sont généralement rurales la probablité d'avoir un accident diminue

mais le coût moyen est plus élevée du fait des vitesses de circulation plus importantes. Cette tendance est bien mise en relief par la figure 20 où la fréquence sinistre croit en même temps que la zone.

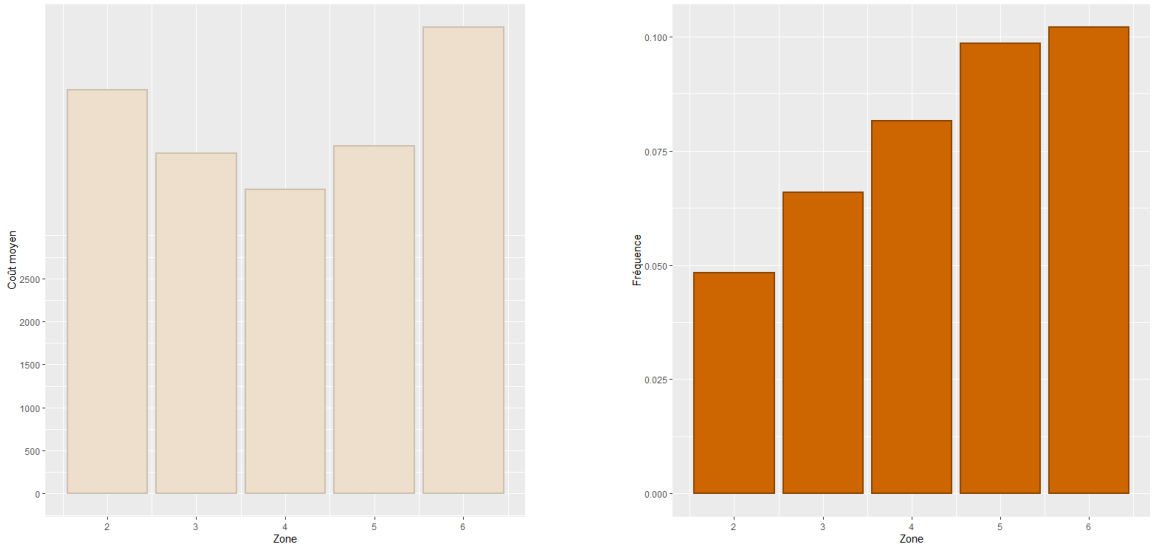

FIGURE 20 – Coût moyen et fréquence des modalités de la variable Zone

Usage La variable usage est une variable multimodale qui décrit l'usage du véhicule. Cette variable permet de jauger indirectement la fréquence d'utilisation du véhicule qui est généralement à la fréquence d'accidents.
On remarque que la modalité affaires a une fréquence sinistre plus élevée que les autres, ce qui est cohérent car un véhicule professionnel va plus souvent être utilisé qu'un véhicule privé.

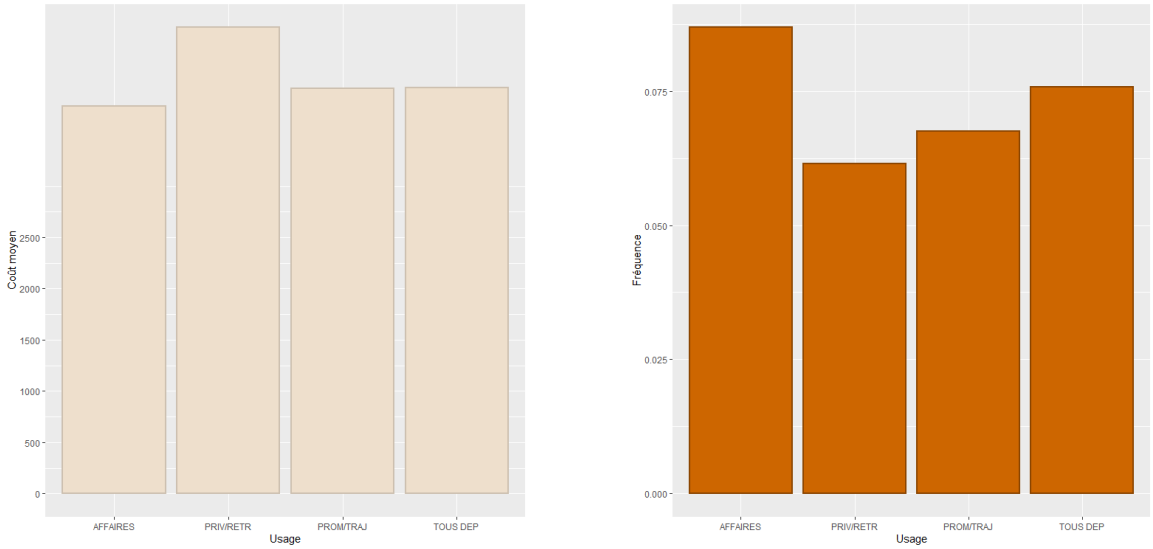

FIGURE 21 – Coût moyen et fréquence des modalités de la variable Usage

3.5 Théorie des valeurs extrêmes

Créer des classes de risque en assurance automobile est obligatoire pour que le principe de mutualisation soit efficace dans un portefeuille.
Cependant les sinistres graves présents dans une classe viennent biaiser l'hypothèse d'homogénéité de ces classes ainsi que déstabiliser le calcul de la prime pure.
Les assureurs écrêtent généralement ces sinistres graves et font soit appel à la réassurance qui prennent en charge les coûts excédentaires graves en contrepartie d'un paiement d'une prime de réassurance soit répartissent la charge grave sur l'ensemble du portefeuille.

Le problème revient donc au fait de déterminer le seuil à partir duquel on considère un sinistre comme grave pour conserver la stabilité des indicateurs de risques et donc avoir une prime pure cohérente.

La théorie des valeurs extrêmes et un outil qui permet de trouver approximativement ce seuil. Il existe plusieurs méthodes pour obtenir ce seuil : les valeurs records, la moyenne des excès et l'approximation par la loi de Pareto généralisée.
Nous rappelons tout d'abord quelques notions théoriques avant d'aborder les 3 méthodes.

3.5.1 Distribution des extrema

Soit M_n la variable aléatoire du maximum d'un échantillon n d'une suite $i.i.d$ de variables aléatoires $(X_i)_{1\leq i\leq n}$. On a donc :

$$M_n = max(X_i)_{1\leq i\leq n}$$

On cherche à calculer la loi limite de M_n en fonction de la variable aléatoire X.
Soit F_X la fonction de répartition de X, on a :

$$\begin{aligned} F_{M_n}(x) = P(M_n < x) = P(X < x) \dots P(X_n < x) \\ = [F_x(x)]^n \end{aligned} \tag{4}$$

On a donc $F_{M_n} = F_X^n$.

On ne connait cependant pas F_X dans la plupart des cas. Il est alors possible de regarder le comportement asymptotique de M_n pour étudier la fin de la distribution qui renseignera sur les valeurs extrêmes de X.

Soit $x_F = sup\{x \in R : F_X(x) < 1\}$, on regarde la limite de F_{M_n} en l'infini :

$$\lim_{n\to+\infty} F_{M_n}(x) = \lim_{n\to+\infty} [F_X(x)]^n = \begin{cases} 0 & \text{si } x < x_F \\ 1 & \text{si } x \geq x_F \end{cases}$$

On obtient donc une loi dégénérée, ce qui ne nous aide pas à déduire les valeurs extrêmes de X. On peut donc pour pallier à ce problème réaliser une transformation comme par exemple une standardisation.

3.5.2 Distribution asymptotiques du maximum

Le but est à l'issue de la dernière partie de trouver une loi non dégénérée pour caractériser M_n. L'utilisation des théorèmes issus de la théorie des valeurs extrêmes est une bonne solution.

Théorème de Fisher, Tippett et Gnedenko
Soit $X_1 \dots X_n$ une suite de variables aléatoires réelles iid de loi continue P et $M_n = max(X_1, \dots, X_n)$.
S'il existe deux suites réelles $(a_n)_{n\geq 1}$ et $(b_n)_{n\geq 1}$ avec $b_n > 0$, et une fonction de répartition non-dégénérée G telle que $\left(\frac{M_n - a_n}{b_n}\right) \xrightarrow{d} G$ alors G est nécessairement de l'un des trois types suivants :

$$G_0(x) = exp(-exp(-x)) \qquad -\infty < x <= \infty$$

$$G_{1,\alpha}(x) = \begin{cases} 0 & \text{si } x < 0 \\ exp(-x^{-\alpha}) & \text{si } x \geq 0, \alpha > 0 \end{cases}$$

$$G_{2,\alpha}(x) = \begin{cases} exp(-(-x)^{-\alpha}) & \text{si } x < 0, \alpha < 0 \\ 1 & \text{si } x \geq 0 \end{cases}$$

On appelle G_0 loi de Gumbel, $G_{1,\alpha}$ loi de Fréchet et $G_{2,\alpha}$ loi de Weibull.
On dit alors que X appartient au domaine d'attraction maximum de G et on note $X \in MDA(G)$

Une forme plus générale de la distribution des valeurs extrêmes peut être exprimée, en ajoutant μ et σ respectivement les paramètres de localisation et de dispersion. Cette forme est appelée en anglais *Generalized Extreme Value Distribution* ou plus simplement GEV.

$$G_{\mu,\sigma,\xi}(x) = exp\{-[1 + \xi(\tfrac{x-\mu}{\sigma})]^{-1/\xi}\}$$

avec ξ un paramètre de forme qui ne dépend que de X. Si ξ est élevé on parle de distributions à queue lourde.

On dit que la fonction $G_{\mu,\sigma,\xi}$ est dans le domaine d'attraction maximum de Fréchet si $\xi > 0$, de Gumbel si $\xi = 0$ ou de Weibull si $\xi < 0$.

3.5.3 Distribution conditionnelle des excès

Dans cette seconde partie nous voyons la méthode à dépassement de seuil qui consiste à fixer un seuil et à étudier la différence entre les observations qui dépassent ce seuil et le seuil en lui-même.
Cette deuxième méthode est préférable car pour estimer les paramètres de la GEV dans la 1ère approche on extrait des blocs de valeurs des X_i et on considère la distribution engendrée par les maxima de chacun de ces blocs.
Le problème de ce procédé est qu'il est destructeur en information car les blocs ainsi crées peuvent soit ne pas contenir de valeur extrême soit en contenir plusieurs.

Avant de citer le théormèe central de cette partie, il est bon de rappeler quelqules définitions :

Définition
Soit $X_1 \dots X_n$ des variables aléatoires réelles iid de fonction de répartition F et $u < x_F$, avec x_F un certain seuil.
Soit $E_u = \{j \in \{1, 2 \dots n\}/X_j > u\}$.
On appelle excès au-delà du seuil u les variables aléatoires Y_j tels que : $Y_j = X_j - u$ pour $j \in E_u$.

Définition
La distribution conditionnelle F_u des excès au-delà du seuil u est définie par :

$$F_u(x) = P(X < x/X > u) = \frac{F(x)-F(u)}{1-F(u)} \text{ pour } x \geq u.$$

Le but est donc de trouver la loi de probablité qui permet d'approcher cette distribution conditionnelle.

Théorème de Pickands, Balkema et Haan
Soit F_u la distribution conditionnelle des excès au-delà d'un seuil u et F sa fonction de répartition.
F appartient au domaine d'attraction maximum de G_ξ si il existe σ fonction positive telle que :

$$\lim_{u \to x_F} \sup_{0<y<x_F-u} |F_u(y) - F^{GPD}_{\xi,\sigma(u)}(y)| = 0,$$

avec F^{GPD} la fonction de répartition de la loi de pareto généralisée , exprimée ci-dessous :

$$F^{GPD}_{\xi,\sigma}(y) = \begin{cases} 1 - (1 + \xi y/\sigma)^{-1/\xi} & \text{si } \xi \neq 0, \\ 1 - exp(-y/\sigma) & \text{si } \xi = 0, \end{cases}$$

avec $y \in [0, (x_F - u)]$ si $\xi \geq 0$ et $y \in [0, Min(-\sigma/\xi, x_F - u)]$ si $\xi < 0$.

Grâce à ce théorème, on peut lier le paramètre de la loi du domaine d'attraction maximum avec le comportement limite des excès passé un certain seuil, notamment l'indice de queue ξ qui est à la fois un paramètre de la loi de Pareto généralisée et qui est donné par l'étude du maximum.

On peut grâce à cela distinguer les lois à queues lourdes des lois à queues légères qui appartiennent respectivement aux domaines d'attraction de Fréchet et de Gumbel. On parle de distribution à queue lourde lorsque la densité de la loi s'étale plus lentement que celle de la loi normale.

On peut définir aussi définir la loi de Pareto généralisée de cette façon :

$$F^{GPD}_{\xi,\sigma}(y) = 1 + logG(y)$$

avec $G(y)$ la loi GEV (Generalized Extreme Value) avec $\nu = 0$.

3.5.4 Estimation du seuil extrême

Il existe plusieurs méthodes pour estimer un seuil à partir d'une distribution empirique de sinistres. Nous en présenterons 2 dans ce qui va suivre, à savoir la méthode des

valeurs records et la méthode de la fonction moyenne des excès. On recherche un seuil qui soit à la fois assez grand pour pouvoir employer les outils décrits précedemment et à la fois assez petit pour pouvoir avoir un bon nombre de valeurs extrêmes.

Méthode des valeurs records
Cette méthode permet d'obtenir directement un nombre de valeurs extrêmes en fonction du nombre de sinistres dans notre échantillon.

On définit le processus de comptage :

$$N_1 = 1, N_n = 1 + \sum_{k=2}^{n} I_{\{X_k > M_{k-1}\}} \text{ avec } n \geq 2$$

avec

$$I_{\{X_k > M_{k-1}\}} = \begin{cases} 1 & \text{si } X_i > M_{k-1} \\ 0 & \text{sinon} \end{cases}$$

Pour obtenir le seuil de notre distribution, il suffit de calculer l'espérance $E(N_n)$ de notre processus de comptage en fonction du nombre de sinistres. Cette espérance est égale au nombre de valeurs extrêmes. On en déduit alors trivialement le seuil extrême qui est la $E(N_n)^{ime}$ valeur de notre distribution ordonnée du sinistre le plus coûteux au sinistre moins coûteux.

n	1000	2000	5000	10000	20000	50000	10000	200000
$E(N_n)$	7,47	8,16	9,09	9,78	10,48	11,37	12,09	12,78

TABLE 7 – Espérance du processus de comptage en fonction du nombre de sinistres

Notre portefeuille comporte 18 148 sinistres, on admet alors d'après le tableau que notre échantillon de sinistres comporte 10 valeurs extrêmes.
Notre 10^{ime} sinistre le plus coûteux étant à 273 381€, nous choisissons un seuil de sinistres graves grâce à cette méthode de 273 000€.

Fonction moyenne des excès

Soit e_n la fonction moyenne des excès définie de façon suivante :

$$e_n(u) = E[X - u/X > u]$$

Cette fonction donne l'espérance des dépassements du seuil u fixé.
On peut estimer cette fonction moyenne des excès par l'approximation empirique suivante :

$$\hat{e}_n(u) = \frac{\sum_{i=1}^{n}(X_i - u)^+}{\sum_{i=1}^{n}(I_{\{X_i > u\}})}$$

Dans la pratique on trace généralement l'estimateur empirique $\hat{e}_n$ et on choisit u de façon à ce que $\forall x \geq u$ l'estimation empirique $\hat{e}_n(u)$ soit approximativement linéaire. Cela est dû au fait que la fonction d'espérance résiduelle d'une loi GPD de paramètre $\xi < 1$ est affine.
Sous forme affine, la fonction moyenne des excès empirique est définie par :

$$\hat{e}_n(u) = \frac{\hat{\sigma}+\hat{\xi}u}{1-\hat{\xi}}, \text{ où } \hat{\sigma} + \hat{\xi}u > 0$$

On regarde donc graphiquement le moment où la fonction moyenne des excès empirique se stabilise de façon affine et on en déduit le seuil u.

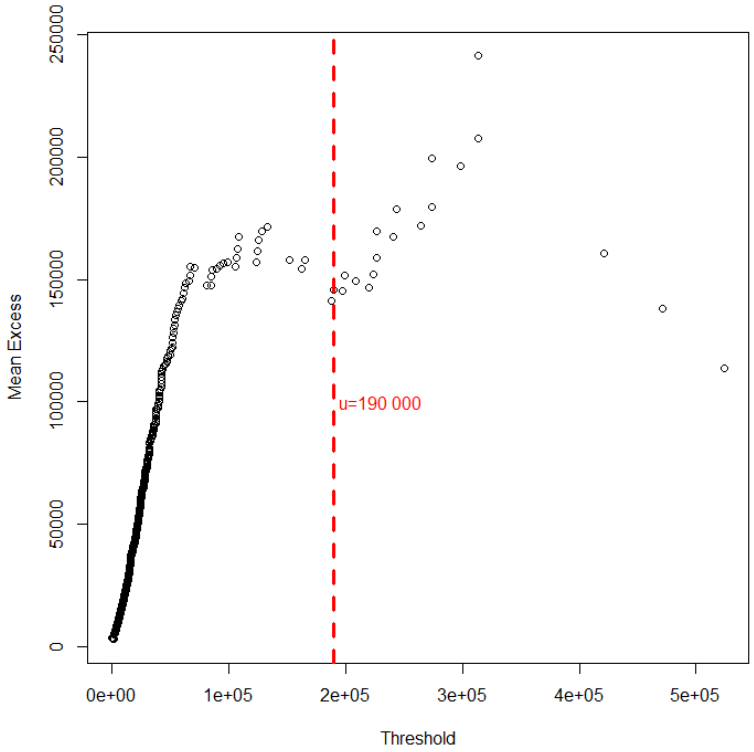

FIGURE 22 – Fonction moyenne des excès de notre distribution

On regarde les tendances linéaires du graphique pour en déduire le seuil empirique û. On constate que vers 190 000 la courbe se stabilise et devient linéaire puis vers 320 000 on retrouve une deuxième tendance linéaire. Cependant on rejette cette deuxième tendance car le nombre de données y participant est très faible. On retient donc le seuil empirique û de 190 000.

La 1^{ere} méthode ayant donné $\hat{u}$=270 000 et la deuxième méthode $\hat{u}$=190 000, on moyenne les deux pour arriver à un seuil grave u_{final}=230 000€

On écrête donc tous les sinistres de la base qui dépassent 230 000€ et on répartit l'excédent sur la prime pure de tous les contrats de la base. Pour cela, on sépare la prime pure en deux :

$$PP_{finale} = PP_{attritionnelle} + PP_{grave}$$

avec

— $PP_{attritionnelle}$ la prime pure couvrant les sinistres attritonnels

— PP_{grave} la prime pure couvrant l'excédent de l'écrêtage des sinistres graves

Le seuil grave ayant été trouvé, PP_{grave} peut être calculé trivialement :

$$PP_{grave} = \frac{M_{graves}}{N_{contrats}} = 7,3€$$

On majorera donc la prime pure calculée pour les sinistres attritionnels de 7,3€ pour obtenir la primue pure finale.
Par la suite, tous les calculs effectués concerneront les sinstres attritionnels et écrêtés à 230 000€.

4 Modèles linéaires généralisés

4.1 Théorie

Le modèle linéaire gaussien a été pendant longtemps l'outil principal des actuaires pour quantifier l'influence de variables explicatives sur une variable réponse comme une fréquence ou un coût sinistre.
Cependant à cause de la compléxité grandissante des problèmes statistiques ces modèles sont devenus obsolètes de par leurs hypothèses fortes comme une obligation de densité de probabilité gaussienne, une linéarité du score ou encore une homoscédasticité. Les modèles linéaires généralisés ont donc été introduits fin 20^{ime} siècle pour pallier à ces problèmes notamment en s'affranchissant de l'hypothèse de normalité.
Ils sont maintenant la référence en tarification non-vie ainsi que sur certaines problèmatiques en assurance-vie comme l'établissement des tables de mortalité ou encore l'estimation des indicateurs démographiques.

Modèle linéaire	Modèle linéaire généralisé
Variables aléatoires $Y_1, \ldots, Y_n$	Variables aléatoires $Y_1, \ldots, Y_n$
$Y_i \rightarrow Normale$	Y_i n'est pas obligatoirement normale
	mais doit l'être dans la famille exponentielle
$E[Y_i] = \nu_i = x_i^{'}\beta$	$g(E[Y_i]) = g(\mu_i) = x_i^{'}\beta$
et $V[Y_i] = \sigma^2$	ou $\mu_i = g^{-1}(x_i^{'}\beta)$

TABLE 8 – Différences entre le modèle linéaire et le modèle linéaire généralisée

Soit Y une variable réponse aléatoire, on définit dans un modèle GLM la relation suivante :

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

avec g la fonction de lien du modèle, x_i les variables explicatives.
Cette formule permet de faire le lien entre l'espérance de la variable expliquée et le prédicteur linéaire.

La loi de cette variable réponse Y doit faire partie de la famille exponentielle, de densité :

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

avec b définie sur R et deux fois dérivable et c définie sur R^2.

Exemple classique : Loi Gamma

densité	$f(y) = exp(\frac{-(\mu/\nu^2)y + ln(\mu/\nu^2)}{1/\nu} + c(y,\nu)$
θ	$-(\mu/\nu^2)$
ϕ	$1/\nu$
$b(\theta)$	$ln(-1/\theta)$
$E[Y]$	ν
Fonction variance	$v(\mu) = \mu^2$

TABLE 9 – Différents paramètres de la loi de Gamma

La deuxième partie de la formule (1) qui est similaire au modèle linéaire concerne la partie déterministe du modèle :

$$\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

L'intérêt de la modélisation se trouve dans le choix des variables X_i : sélectionner un trop petit nombre de variables explicatives peut baisser le pouvoir de généralisation du modèle et en séléctionner un trop grand nombre peut le rendre inutilisable.
Il faut donc trouver les variables qui expliquent le mieux le modèle et éliminer celles qui l'expliquent trop peu ou pas du tout.
La méthode de séléction de ces variables se fait généralement par maximum de vraisemblance.

Nous nous intéressons enfin à la fonction g de la formule (1) aussi appelée la fonction de lien. Cette fonction assure la liaison entre l'espérance de la variable réponse et la relation déterministe entre les variables explicatives. Les fonctions usuelles sont :

- la fonction logit : $g(x) = ln(\frac{x}{1-x})$
- la fonction identité : $g(x) = x$
- la fonction cloglog : $g(x) = ln(-ln(1-x))$
- la fonction ln : $g(x) = ln(x)$

Cette dernière fonction est très pratique car cela permet d'avoir un modèle multiplicatif :

$$g(E[Y]) = ln(E[Y])$$

et donc on a :

$$E[Y] = exp(\sum_{k=1}^{p} \beta_k x_k) = exp(\beta_0) \times exp(\beta_1 X_1) \times \cdots \times exp(\beta_p X_p)$$

Cette propriété de multiplicativité est essentielle quand on travaille sur un modèle Coût X Fréquence.

4.1.1 Estimation des paramètres d'une GLM

Les (β_i) sont estimés par le maximum de vraisemblance et plus particulièrement en maximisant la log-vraisemblance.
La log-vraisemblance s'écrit :

$$ln(L(y,\theta,\phi)) = \sum_{i=1}^{n} ln(f_{\theta,\phi}(y_i))$$

On cherche à maximiser cette relation et donc en dérivant la formule, on cherche à rendre nulle la quantité m_j $\forall j \in [1,\ldots,p]$ définie par :

$$m_j = \sum_{i=1}^{n} \frac{\partial ln(f_{\theta,\phi}(y_i))}{\partial \beta_j}$$

Après quelques calculs on arrive au résultat suivant :

$$m_j = \sum_{i=1}^{n} \frac{(y_i-\mu_i)x_i^j \partial \mu_i}{Var(Y_i)\partial \eta_i}$$

avec $\mu_i = b^{'}(\theta_i)$ et $\eta_i = \theta_i$
Pour résoudre ces équations il existe plusieurs méthodes itératives que nous ne décrirons pas ici. L'ensemble des paramètres (β_i) est calculé sur le logiciel R.

4.1.2 Validation du modèle

Déviance
La déviance est une mesure courante du bon ajustement d'une GLM. Elle s'exprime comme l'écart entre la log-vraisemblance obtenue en β et celle obtenue avec un modèle saturé. Elle est définie par :

$$D = 2 \times (lnL(Y|Y) - lnL(\hat{\mu}|Y))$$

avec $L(Y|Y)$ la vraisemblance du modèle saturé et $L(\hat{\mu}|Y)$ la vraisemblance maximale du modèle calculé.
Plus D est petit, meilleur sera l'ajustement du modèle.

Analyse des résidus
On couple généralement l'analyse de la déviance avec l'analyse des résidus d'une GLM. Un résidu permet de mesurer l'écart entre les valeurs cibles observées et celles qui sont calculées.
Les résidus les plus courants sont :

- Résidus de déviance $\quad r_i^p = \epsilon(y_i - \mu_i)\sqrt{d_i}$
- Résidus de Pearson $\quad r_i^p = \sqrt{\omega_i}\frac{y_i-\mu_i}{\sqrt{Var(\mu_i)}}$

avec ϵ la fonction *signe*, d_i et ω_i respectivement la déviance et la statistique de Pearson de l'observation i, μ_i la variable réponse empirique et y_i la variable réponse théorique.
Des résidus à structure aléatoire et centrés en 0 sont un signe d'une faible erreur de modélisation.

4.1.3 Choix du modèle

Critère d'information d'Akaike (AIC)

Le critères AIC est utilisé lorsqu'il est nécessaire de comparer deux modèles entre eux.Il pénalise les modèles avec trop de paramètres qui peuvent présenter un risque d'overfitting.
L'AIC est défini par :

$$AIC = 2 \times (p - lnL(\hat{\mu}|Y))$$

avec $lnL(\hat{\mu}|Y)$ et p respectivement la log-vraisemblance maximale et le nombre de variables explicatives du modèle.
Le modèle avec le critère AIC le plus faible est le plus optimal.

4.2 Etablissement des modèles

Intéressons-nous maintenant à la modélisation de la prime pure. Dans l'approche classique, la modélisation Coût-Fréquence est utilisée, c'est-à-dire qu'on modélise deux GLM séparément respectivement pour le coût et pour la fréquence sinistre, puis qu'on multiplie les deux obtenues pour trouver la prime pure finale comme suivant :

$$PP_{finale} = GLM_{frquence} \times GLM_{cout}$$

Le problème dans notre cas vient de l'application de la convention IRSA qui vient fausser la distribution des coûts sinistres. En effet il n'existe pas de loi pour la variable Y qui va correctement approximer la distribution décrite en figure 23.

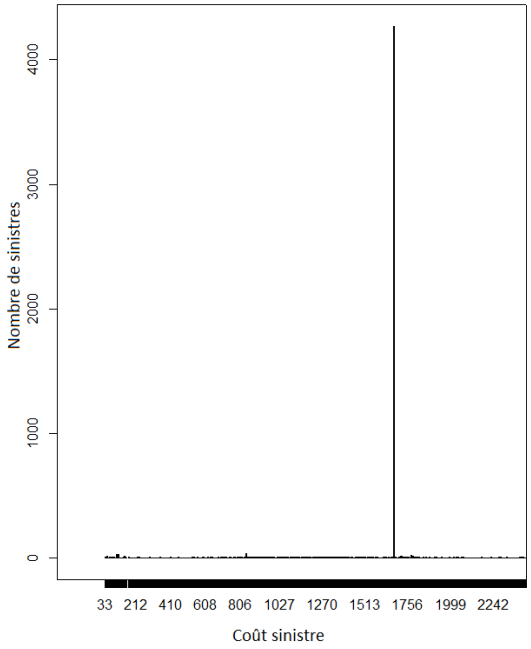

FIGURE 23 – Nombre de sinistres en fonction du coût sinistre

On propose donc un nouveau modèle qui va traiter les sinistres IRSA à part :

$$PP_{finale} = PP_{IRSA} + PP_{NONIRSA}$$

avec PP_{IRSA} la prime pure couvrant les sinistres IRSA et $PP_{NONIRSA}$ la prime pure couvrant les sinistres non IRSA.
On a donc en termes de GLM si on considère un modèle Coût/Fréquence :

$$PP_{finale} = GLM^{IRSA}_{frquence} \times GLM^{IRSA}_{cout} + GLM^{NONIRSA}_{frquence} \times GLM^{NONIRSA}_{cout}$$

avec :

— $GLM^{IRSA}_{frquence}$ et GLM^{IRSA}_{cout} des GLM qui ont été entrainées sur la base de départ avec les sinistres uniquement IRSA ;

— $GLM^{NONIRSA}_{cout}$ et $GLM^{NONIRSA}_{frquence}$ des GLM qui ont été entrainées sur la base de départ sans les sinistres IRSA.

On a de plus GLM^{IRSA}_{cout} qui n'existe pas car le coût moyen n'a pas besoin d'être modélisé, étant toujours égal à 1678€. Il suffit donc dans le cas des sinistres IRSA d'uniquement prédire la fréquence sinistre puis de multiplier par 1678€ pour avoir la prime pure IRSA.
On se retrouve au bilan avec deux GLM fréquence et une GLM coût à modéliser.

Séparation de la base
Pour la suite du mémoire, nous séparons la base en deux de façon aléatoire, en considérant 80% de la base comme base d'entrainement et le reste comme base de test.

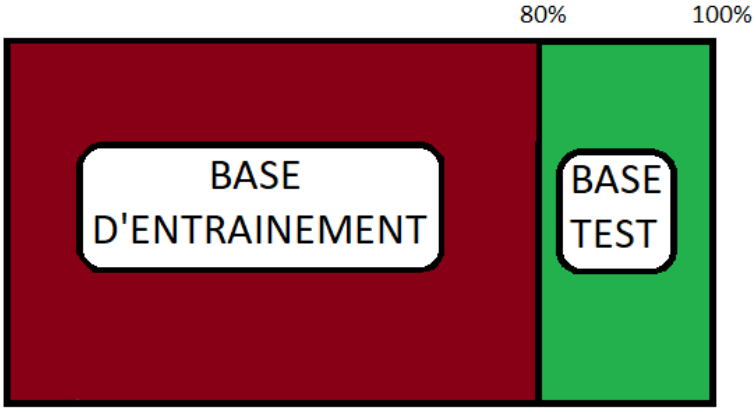

FIGURE 24 – Séparation de la base de départ

4.3 Modèles fréquences

La fréquence sinistre qu'on cherche à modéliser se caractérise par :

$$F = \frac{Nombresinistres}{Exposition}$$

L'exposition définie au-dessus est prise sur une année. On ne prédit pas la fréquence à proprement parler mais plutôt le nombre de sinistres sur une année. L'exposition est une variable qui sera mise en offset dans la GLM fréquence,ie, on a :

$$log(E[Frequence]) = log(E[\frac{Nombresinistres}{Exposition}]) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

donne :

$$log(E[Nombresinistres]) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + ln(Exposition)$$

et donc un final :

$$E[Nombresinistres] = Exposition \times exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$$

Comme décrit plus au-dessus, nous aurons deux modèles fréquence dans notre étude : un 1^{er} modèle qui prédira la fréquence des sinistre non-IRSA et la deuxième qui prédira la fréquence des sinistre IRSA.

4.3.1 Modèle fréquence NON IRSA

Nous choisissons tout d'abord comme fonction de lien le logarithme néperien pour conserver le caractère multiplicatif du modèle coût-fréquence.
Deux lois peuvent être utilisées pour modéliser le nombre de sinistres, à savoir la loi de Poisson et la loi binomiale négative.
La loi de Poisson est celle qu'on utilise traditionnellement mais lorsqu'on est dans un cas de surdispersion de nos données on utilise la loi binomiale négative.
Plusieurs critères existent pour choisir la bonne loi :

Egalité Espérance-Variance
Une des propriétés de la loi de Poisson est l'égalité de sa variance et de son espérance. On regarde donc l'espérance et la variance de la variable nombre de sinistres de notre base :

Espérance	Variance
0.042	0.044

TABLE 10 – Espérance et variance du nombre de sinistres

On constate que la variance de l'échantillon est supérieure à son espérance, cela fait plutôt pencher notre choix vers la loi binomiale négative.

Densité empirique et théorique
Un deuxième critère de jugmement est la superposition de la densité empirique de notre nombre de sinistres avec la densité théorique des lois cibles. On précise que nos densités concernent des lois discrètes.

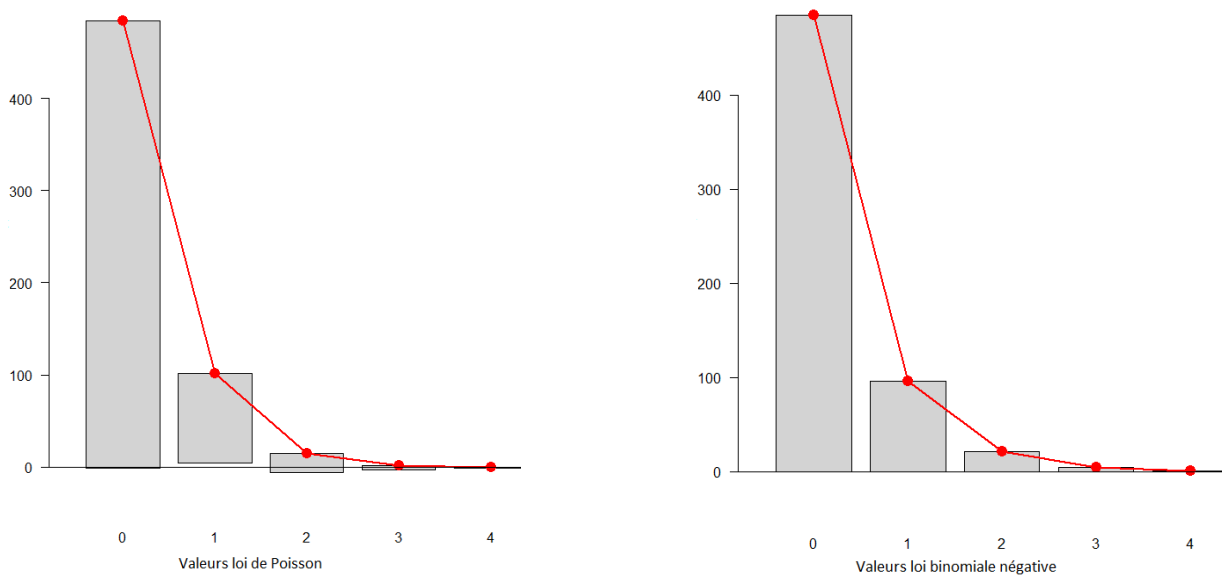

FIGURE 25 – Comparaison des densités théorique et empiriques

On constate d'après la figure 30 que la loi binomiale négative est mieux ajustée par notre échantillon que la loi de Poisson.

Critère AIC

Bien que le critère AIC serve habituellement à filtrer les variables explicatives, on peut aussi l'utiliser pour choisir la loi adaptée en gardant l'AIC le plus faible parmi deux modèles identiques à l'exception de leur loi. On obtient dans notre cas en considérant un modèle avec nos 17 variables explicatives :

Loi	AIC
Poisson	82 814
Binomiale négative	82 696

TABLE 11 – AIC pour chacune de nos lois

La loi binomiale négative a un AIC plus faible que celui de Poisson.

Les trois critères cités ci-dessus confirment tous la pertinence du choix de la loi binomiale négative à la place de la loi de Poisson. C'est donc cette loi qu'on choisit pour modéliser notre fréquence sinistre.

Premier modèle

On crée un premier modèle qui contient toutes les variables explicatives :

```
Call:
glm.nb(formula = Nbre_sin ~ Ant_assurance + SOR_FORMULE + POL_FRACTION +
    VEH_TARIF_ZONE + Age_cond + Age_vehicule + Anc_permis + Resil_n_paiement +
    Resil_pmq + Resil_freq_sin + Usage + Classe + Groupe + CRM +
    Nbre_sin_declar + dernier_sinistre + Duree_assur + offset(log(Exposition)),
    data = base_train_freq_nida, link = log, init.theta = 1.436086492)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7605  -0.3381  -0.2460  -0.1571   5.2972

(Dispersion parameter for Negative Binomial(1.4361) family taken to be 1)

    Null deviance: 59447  on 244524  degrees of freedom
Residual deviance: 56950  on 244469  degrees of freedom
AIC: 82696

Number of Fisher Scoring iterations: 1
```

FIGURE 26 – Sortie R du modèle Fréquence non IRSA complet

Choix des variables

L'étape de sélection des variables est cruciale pour obtenir une GLM de qualité avec des résidus faibles et donc avec une bonne justesse de prédiction.
Comme expliqué lors de la définition de l'AIC, trop de variables explicatives empêcherait le pouvoir de généralisation du modèle et trop peu de variables donneraient des prédictions à fiabilité faible.

Bien que l'expertise d'un professionnel de l'assurance est généralement la meilleur solution pour trouver les variables à garder dans le modèle, plusieurs méthodes ont été mises en place :

- Procédé **Forward** :A partir d'un modèle constant qui ne contient aucune variable explicative, on ajoute la variable qui fait baisser le plus l'AIC par rapport au modèle précédent.On répète le processus jusqu'au moment où l'AIC ne baisse plus en ajoutant des variables.
- Procédé **Backward** : A partir d'un modèle complet, on enlève la variable qui fait baisser le plus l'AIC. On répète le processus jusqu'à ce que l'AIC ne baisse plus en enlevant une variable.
- Procédé **Bidirectionnel** : Ce procédé est un mélange des deux précédents : à chaque itération, l'algorithme choisit d'ajouter ou d'enlever une variable en fonction de l'AIC.

Nous utilisons la méthode Backward dans notre séléction de variables.

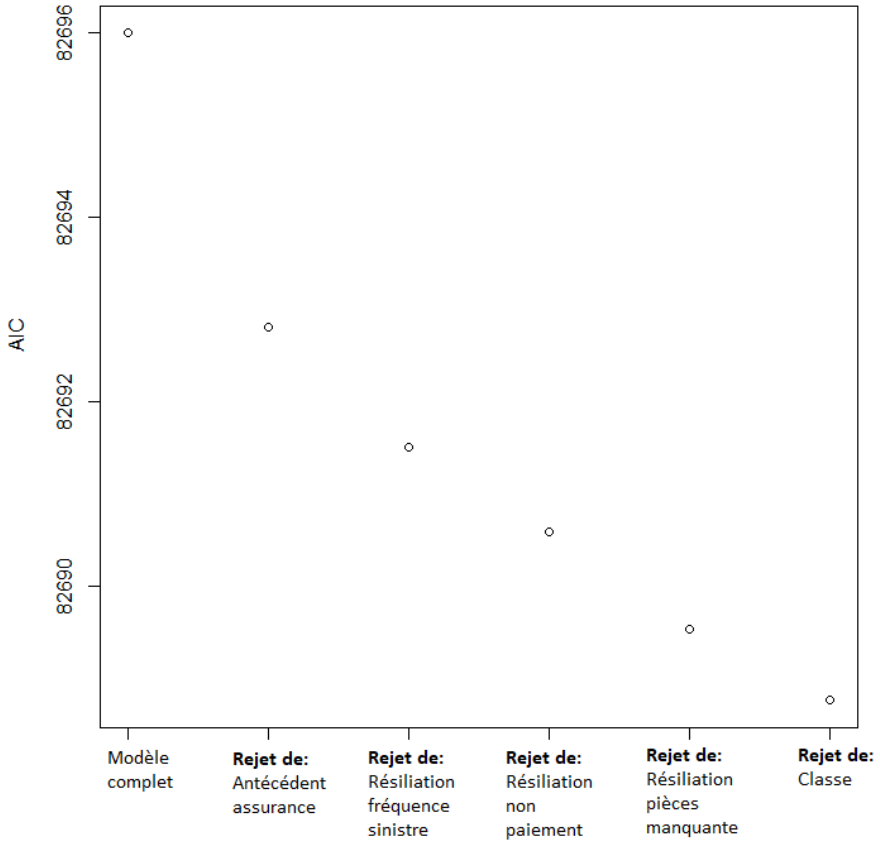

FIGURE 27 – Procédure Backward sur le modèle

On obtient donc à notre modèle de fréquence optimal, avec un AIC diminué de 7 par rapport au modèle complet :

```
Call:
glm.nb(formula = Nbre_sin ~ SOR_FORMULE + POL_FRACTION + VEH_TARIF_ZONE +
    Age_cond + Nbre_sin_declar + Age_vehicule + Anc_permis +
    Usage + Groupe + CRM + dernier_sinistre + Duree_assur + offset(log(Exposition)),
    data = base_train_freq_nida, link = log, init.theta = 1.431804884)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7511  -0.3380  -0.2460  -0.1571   5.3074

(Dispersion parameter for Negative Binomial(1.4318) family taken to be 1)

    Null deviance: 59433  on 244524  degrees of freedom
Residual deviance: 56948  on 244478  degrees of freedom
AIC: 82689

Number of Fisher Scoring iterations: 1
```

FIGURE 28 – Sortie R du modèle Fréquence non IRSA optimal

5 des 17 variables explicatives ont été rejetées du modèle final :

Variables conservées	variables enlevées
Age véhicule, Fractionnement Formule,Ancienneté permis Groupe, CRM,Nombre sinistres déclarés Age conducteur,Usage Dernier sinistre,Interruption Assurance Zone	Antécédent assurance Résiliation non paiement Résiliation pièces manquantes, Classe Résiliation fréquence sinistre

TABLE 12 – Variables en jeu dans le modèle fréquence non IRSA

4.3.2 Modèle fréquence IRSA

On reprend les mêmes étapes pour le modèle fréquence IRSA que pour le modèle fréquence non IRSA, la seule différence étant la base sinistre qui ne contient que des sinistres IRSA à 1678€.

Choix de la loi

Comme dans la sous-section précédente, on étudie les critères **Espérance/variance**,**densité empirique/théorique** et **AIC** sur les lois de Poisson et Binomiale négative.

Espérance	Variance
0.017	0.018

TABLE 13 – Critère Espérance/Variance

Loi	AIC
Poisson	42 172
Binomiale négative	42 088

TABLE 14 – Critère AIC

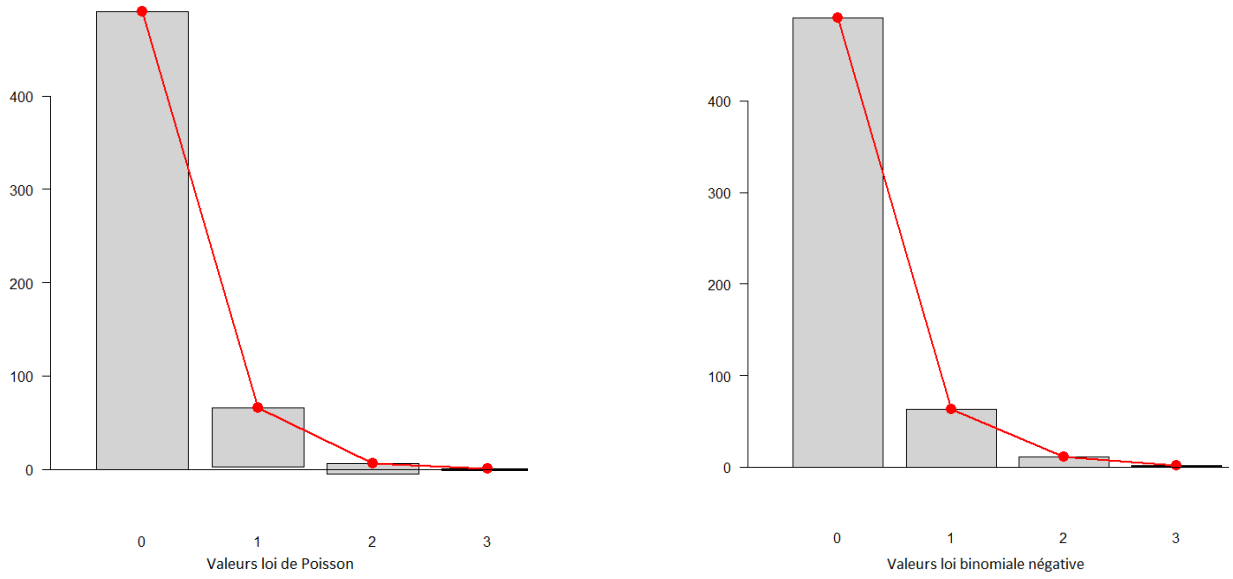

FIGURE 29 – Comparaison des densités théorique et empiriques

Au vu de ces considérations, on choisit encore une fois la loi binomiale négative.

Choix des variables

Le modèle complet étant le même que le modèle fréquence non IRSA, on passe directement au choix des variables en appliquant la méthode Backward. On retient à l'issue du processus 14 variables explicatives et on en rejette 3 :

Variables conservées	variables enlevées
Age véhicule, Fractionnement Formule, Ancienneté permis, Résiliation non paiement CRM, Nombre sinistres déclarés Age conducteur, Usage, Classe Dernier sinistre, Interruption Assurance Zone, Résiliation fréquence sinistre	Groupe Antécédent assurance Résiliation pièces manquantes

TABLE 15 – Variables en jeu dans le modèle fréquence IRSA

On recrée un modèle fréquence final avec ces 14 variables explicatives séléctionnées ce qui a pour conséquence de diminuer l'AIC de 5 par rapport au modèle complet, on obtient donc finalement :

```
Call:
glm.nb(formula = Nbre_sin ~ SOR_FORMULE + POL_FRACTION + VEH_TARIF_ZONE +
    Age_cond + Nbre_sin_declar + Age_vehicule + Anc_permis +
    Usage + Resil_n_paiement + Resil_freq_sin + Groupe + CRM +
    dernier_sinistre + Duree_assur + offset(log(Exposition)),
    data = base_train_freq_ida, link = log, init.theta = 0.6143997464)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5200  -0.2176  -0.1667  -0.1169   4.4587

(Dispersion parameter for Negative Binomial(0.6144) family taken to be 1)

    Null deviance: 29882  on 244524  degrees of freedom
Residual deviance: 28913  on 244476  degrees of freedom
AIC: 42083

Number of Fisher Scoring iterations: 1
```

FIGURE 30 – Sortie R du modèle Fréquence IRSA optimal

4.4 Modèle coût

Intéressons-nous maintenant à la modélisation des coûts sinistres **hors sinistres IRSA**. Dans la littérature actuarielle il est très souvent préconisé d'utiliser la loi Gamma, bien que la loi log-normale ou la loi de Weibull soient aussi une option.

4.4.1 Choix préliminaires

Pour justifier l'utilisation de la loi Gamma nous pouvons rapidement regarder quelques indicateurs qui peuvent nous aider dans notre choix.

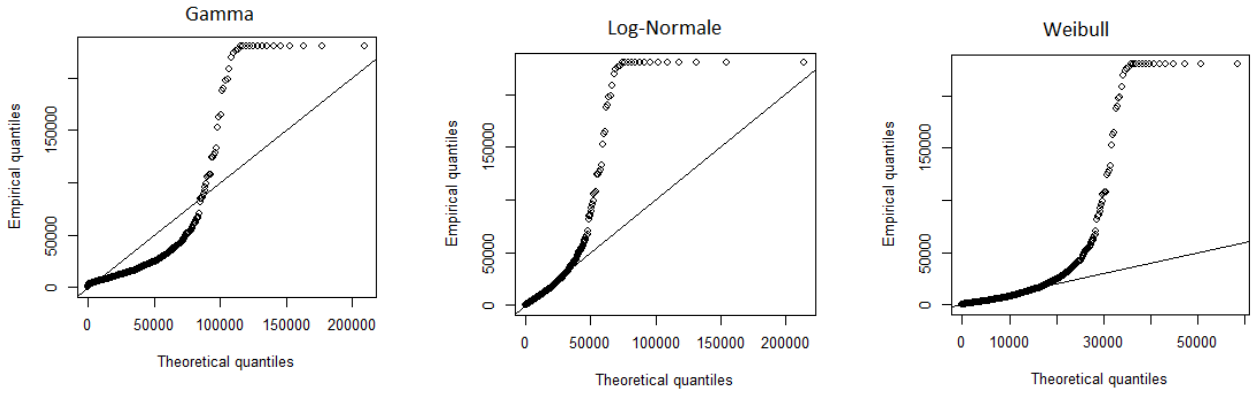

FIGURE 31 – QQ-plot des différentes lois

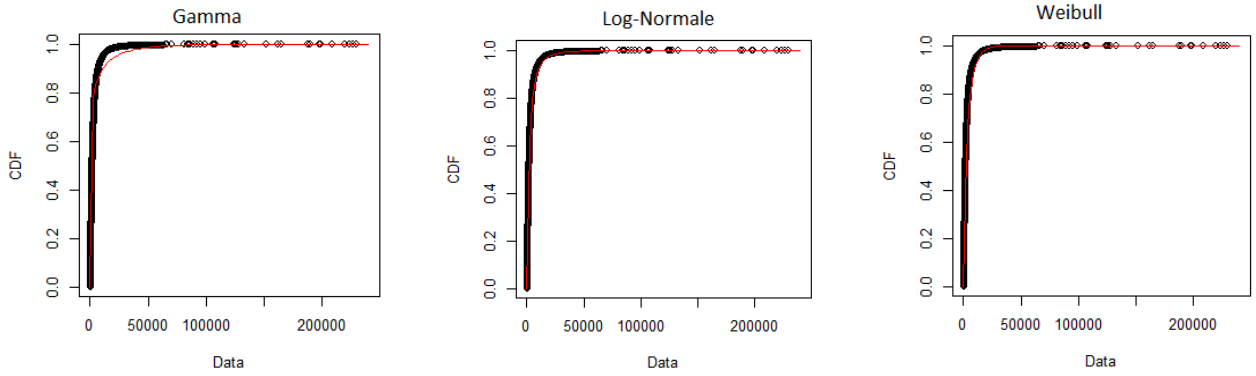

FIGURE 32 – Comparaison des fonctions de répartitions théorique et empiriques

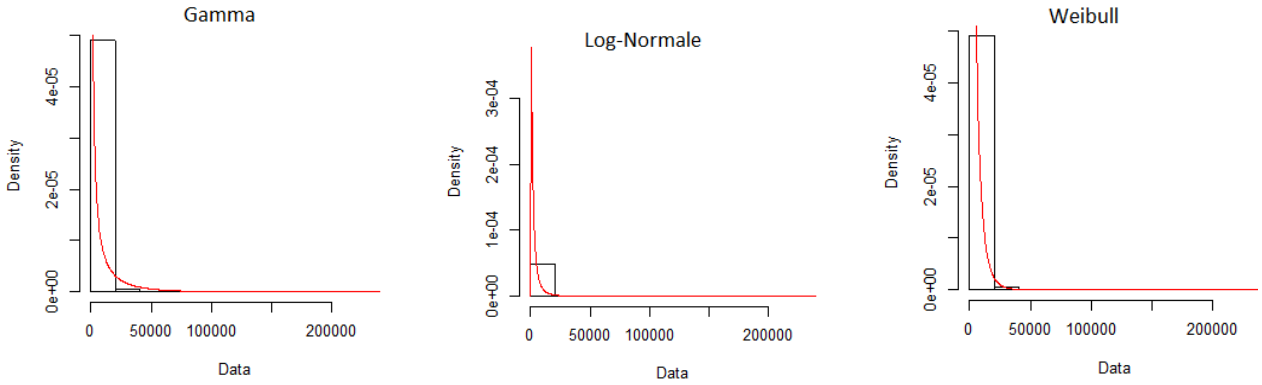

FIGURE 33 – Comparaison des densités théorique et empiriques

Le QQ-plot ainsi que la comparaison des densités montrent que la loi Gamma est la meilleur, bien que le graphique des fonctions de répartition est plus en faveur des lois Log-Normale et de Weibull.
Nous utiliserons donc par la suite la loi Gamma.

La fonction de lien généralement associée à la loi Gamma est la fonction inverse, mais

de par le besoin d'un modèle multiplicatif et de valeurs positives en sortie, nous choisissons le logarithme néperien comme fonction de lien.

4.4.2 Modélisation

On crée un premier modèle "complet" qui comporte toutes les variables explicatives :

```
Call:
glm(formula = cout_tot ~ Ant_assurance + SOR_FORMULE + POL_FRACTION +
    VEH_TARIF_ZONE + Age_cond + Age_vehicule + Anc_permis + Resil_n_paiement +
    Resil_pmq + Resil_freq_sin + Usage + Classe + Groupe + CRM +
    Nbre_sin_declar + dernier_sinistre + Duree_assur, family = Gamma(link = "log"),
    data = base_train_cout_nida)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8350  -1.0813  -0.5894   0.0135  10.6683

(Dispersion parameter for Gamma family taken to be 6.169244)

    Null deviance: 15514  on 10263  degrees of freedom
Residual deviance: 14711  on 10208  degrees of freedom
AIC: 189268
```

FIGURE 34 – Sortie R du modèle coût complet

Tout comme les modèles fréquences, on procède à la sélection des variables par la méthode Backward. Pour plus de précisions théoriques voir la section 4.3. On trouve finalement notre modèle optmial avec une perte d'AIC de 2 :

```
Call:
glm(formula = cout_tot ~ Ant_assurance + SOR_FORMULE + VEH_TARIF_ZONE +
    Age_cond + Age_vehicule + Anc_permis + Resil_freq_sin + Classe +
    Groupe + CRM + Nbre_sin_declar + dernier_sinistre + Duree_assur,
    family = Gamma(link = "log"), data = base_train_cout_nida)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8414  -1.0836  -0.5944   0.0116  10.6943

(Dispersion parameter for Gamma family taken to be 6.225266)

    Null deviance: 15514  on 10263  degrees of freedom
Residual deviance: 14728  on 10216  degrees of freedom
AIC: 189266

Number of Fisher Scoring iterations: 9
```

FIGURE 35 – Sortie R du modèle optimal

Sur nos 17 variables explicatives de départ, 13 ont été retenues et 4 ont été écartées du modèle final :

Variables conservées	variables enlevées
Antécédent assurance,Age véhicule, Formule,Ancienneté permis,Classe, Groupe, CRM,Nombre sinistres déclarés Age conducteur,Résiliation fréquence sinistre Dernier sinistre,Interruption Assurance Zone	Fractionnement Résiliation non paiement Résiliation pièces manquantes Usage

TABLE 16 – Variables en jeu dans le modèle coût

4.5 Limites

Plusieurs facteurs dans les modèles linéaires généralisés font qu'ils peuvent souvent s'éloigner de la réalité :

- le modèle Coût/Fréquence de la prédiction de la prime pure nécessite des hypothèses fortes comme l'indépendance la fréquence sinistre F (et donc par extension le nombre de sinistres de l'assuré) avec les coûts sinistres (M_k), ce qui n'est pas toujours vérifié ;
- les GLM sont paramétriques, c'est-à-dire que l'on choisit une loi au regard de différents indicateurs pour la variable cible, loi qui ne colle pas toujours avec les phénomènes réels ;
- comme son nom l'indique la GLM est linéaire et donc l'influence des variables explicatives aussi alors qu'en pratique le réel est assez rarement linéaire ;
- la discrétisation obligatoire des variables continues fait perdre de l'information ;
- les valeurs extrêmes doivent être traitées à part et le seuil grave à fixer est une autre source d'incertitude ;
- le choix des variables à modéliser ainsi que les interactions entre elles requièrent généralement l'avis d'un expert. Bien que la méthode de séléction Backward s'appuyant sur l'AIC est performante, elle peut rejeter des variables qui auraient été pertinentes combinées à d'autres.

5 Méthode Tweedie

La méthode Tweedie est une autre approche statistique utilisée pour estimer la prime pure en assurance automobile. Elle repose sur la distribution de Tweedie, qui est une distribution composée appropriée pour modéliser les données de sinistres. La distribution de Tweedie est définie par un paramètre d'index p, qui détermine la relation entre la moyenne et la variance. Les modèles de régression basés sur la distribution de Tweedie sont appelés Modèles de Régression de Tweedie (MRT).

5.1 Distribution de Tweedie

La distribution de Tweedie est une distribution continue qui appartient à la famille exponentielle. Elle est particulièrement adaptée pour modéliser des données avec une proportion importante de zéros et une dispersion élevée. La fonction de masse de probabilité (fmp) de la distribution de Tweedie est définie comme suit :

$$f(y;\mu,p,\phi) = C(y,p,\phi)\exp\left(\frac{y\mu^{1-p}-\mu^{2-p}}{(1-p)(2-p)\phi}\right) \tag{5}$$

où y est la variable de réponse, μ est la moyenne, p est le paramètre d'index et ϕ est le paramètre de dispersion. Le paramètre d'index p détermine la relation entre la moyenne et la variance :

$$\mathrm{Var}(Y) = \phi\mu^p \tag{6}$$

Les valeurs courantes de p pour les modèles de sinistres en assurance automobile sont $p = 1.5$ pour la fréquence des sinistres et $p = 2$ pour la sévérité des sinistres.

5.2 Modèles de régression de Tweedie

Les modèles de régression de Tweedie (MRT) sont une extension des modèles linéaires généralisés (GLM) qui permettent d'utiliser la distribution de Tweedie comme distribution de réponse. Les MRT sont définis par une fonction de lien $g(\cdot)$ qui relie la variable de réponse Y aux variables explicatives $\boldsymbol{X}$:

$$g(\mu) = \boldsymbol{X}\boldsymbol{\beta} \tag{7}$$

où μ est la moyenne de la distribution de Tweedie, $\boldsymbol{\beta}$ est un vecteur de coefficients de régression et $\boldsymbol{X}$ est une matrice de variables explicatives. La fonction de lien la plus couramment utilisée pour les MRT est la fonction de lien log :

$$g(\mu) = \log(\mu) \tag{8}$$

Ainsi, le modèle de régression de Tweedie s'exprime comme suit :

$$\log(\mu) = \boldsymbol{X}\boldsymbol{\beta} \tag{9}$$

5.3 Estimation des paramètres

L'estimation des paramètres du modèle de régression de Tweedie, $\boldsymbol{\beta}$ et ϕ, peut être réalisée par la méthode du maximum de vraisemblance. La log-vraisemblance pour un échantillon de données $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ est donnée par :

$$\ell(\boldsymbol{\beta}, \phi) = \sum_{i=1}^{n} \log f(y_i; \mu_i, p, \phi) \tag{10}$$

où $\mu_i = g^{-1}(\boldsymbol{x}_i\boldsymbol{\beta})$. L'estimation des paramètres est réalisée en maximisant la log-vraisemblance :

$$\hat{\boldsymbol{\beta}}, \hat{\phi} = \arg\max_{\boldsymbol{\beta},\phi} \ell(\boldsymbol{\beta}, \phi) \tag{11}$$

Des algorithmes tels que l'algorithme de Newton-Raphson, l'algorithme de Fisher Scoring ou la méthode de quasi-Newton Broyden-Fletcher-Goldfarb-Shanno (BFGS) peuvent être utilisés pour résoudre ce problème d'optimisation.

5.4 Calcul de la prime pure

Une fois les paramètres $\boldsymbol{\beta}$ et ϕ estimés, la prime pure pour un assuré avec des caractéristiques $\boldsymbol{x}_{\text{new}}$ peut être calculée en utilisant la formule suivante :

$$\text{Prime Pure}(\boldsymbol{x}_{\text{new}}) = \hat{\mu}_{\text{new}} = g^{-1}(\boldsymbol{x}_{\text{new}}\hat{\boldsymbol{\beta}}) \tag{12}$$

Dans le cas d'une fonction de lien log, la prime pure s'exprime comme :

$$\text{Prime Pure}(\boldsymbol{x}_{\text{new}}) = \exp(\boldsymbol{x}_{\text{new}}\hat{\boldsymbol{\beta}}) \tag{13}$$

La prime pure représente le coût moyen des sinistres pour un assuré ayant les caractéristiques $\boldsymbol{x}_{\text{new}}$.
Par la suite, nous appliquerons la méthode Tweedie pour calculer la prime pure non IRSA, le but étant de comparer l'utilisation d'un modèle coût fréquence VS un modèle Tweedie, ce qui n'est pas le cas de la prime pure IRSA, le modèle coût de cette dernière étant constant.
Nous additionnerons donc la prime pure IRSA et la prime pure grave mentionnée précédemment au modèle Tweedie pour donner la prime pure Tweedie.

5.5 Modélisation

En utilisant la méthode Backward avec comme indicateur l'AIC pour séléctionner nos variables, nous obtenons :

Variables conservées	variables enlevées
Antécédent assurance,Age véhicule,	Fractionnement
Formule,Ancienneté permis,Classe,	Résiliation non paiement
Groupe, CRM,Nombre sinistres déclarés	Résiliation pièces manquantes
Age conducteur,Résiliation fréquence sinistre	Usage
Interruption Assurance	Dernier sinistre
Zone	

TABLE 17 – Variables en jeu dans le modèle Tweedie

6 Réseaux de neurones

Les algorithmes de machines learning faisant appels à des réseaux de neurones font aujourd'hui face à une popularité grandissante et sont le carburant de plusieurs avancées techonologiques, telles que les voitures autonomes, les reconnaissances faciales, les deep fakes ou encore la conception d'IA performantes comme qu'AlphaGo.
Ils peuvent aussi être utilisés pour faire de la régression comme nous allons procéder dans la suite de ce mémoire.

Rapide historique du réseau de neurones
Les réseaux de neurones ont été théorisés à partir du modèle du neurone biologique. Le premier neurone artificiel, le neurone formel, a été inventé par Warren McCulloch et Walter Pitts fin années 50. Puis Frank Rosenblatt en 1957 inventa le perceptron, qui est un neurone formel muni d'une règle d'apprentissage capable de faire de la classification binaire.
En 1986 le perceptron multicouche est apparu, réseau de neurones qui pour la première fois était capable de traiter des problèmes non linéaires. Plusieurs méthodes ont été développées pour optimiser ce perceptron multicouches comme la rétropropagation du gradient de l'erreur que nous verrons par la suite.
Par la suite les réseaux de neurones ont connu un essor de plus en plus important pour arriver à la popularité que nous connaissons aujourd'hui.

Présentation du neurone artificel
Le neurone artificel est l'élement de base de tout réseau de neurones. On le considère formellement comme une fonction g_i qui prend en entrée le vecteur $x = (x_1, \ldots, x_n)$ lié à des poids $\beta_i = (\beta_{i,1}, \ldots, \beta_{i,n})$. Cette fonction g fait le produit scalaire de x et β et y ajoute un biais m_i et combine le tout avec une fonction d'activation ψ :

$$g_i(x) = \psi(< \beta_i, x > + m_i)$$

Il existe beaucoup de fonction d'activation dans le domaine du machine learning, néanmoins les plus populaires sont :

— La fonction sigmoide

$$\psi(x) = \frac{1}{1+exp(-x)}$$

— La fonction tangente hyperbolique

$$\psi(x) = \frac{exp(2x-1}{exp(2x)+1}$$

— La fonction identité

$$\psi(x) = x$$

— La fonction ReLU (Rectified Linear Unit)

$$\psi(x) = max(x, 0)$$

La fonction d'activation doit généralement respecter certaines caractéristiques pour être efficace dans un réseau de neurones :

- différentiable sur son intervalle de définition : permet d'appliquer les algorithmes d'optimisation par le gradient que nous verrons par la suite ;
- non-linéaire : permet d'approximer toutes les fonctions (théorème à voir par la suite) ;
- $f(x) \approx 0$ quand $x \approx 0$: permet un apprentissage plus rapide ;
- dérivée monotone : permet au neurone artificiel de mieux généraliser ;

La fonction d'activation la plus souvent utilisée est généralement la sigmoide car elle est differentiable et elle regroupe les valeurs obtenues en entrée dans l'intervalle $[0, 1]$. Elle peut cependant présenter certains inconvénients car son gradient est très proche de 0 pour des valeurs de x éloignées de 0. Cela entraine notamment des lenteurs dans l'apprentissage.

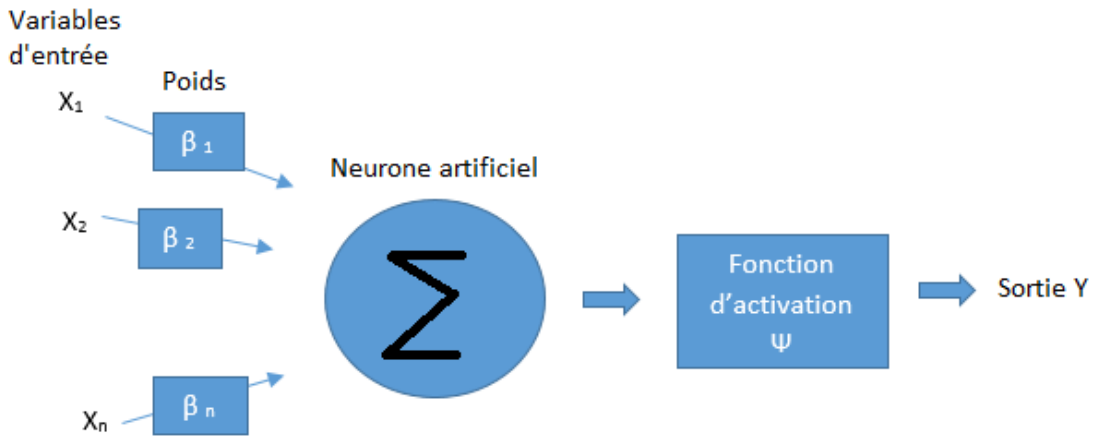

FIGURE 36 – Neurone artificiel

6.1 Différents réseaux de neurones

La recherche scientifique sur les réseaux de neurones étant prolifique, une mutlitude de réseaux de neurones différents a été créee. Nous présenterons les plus classiques, avec un appui sur le perceptron multicouches qu'on utilisera pour prédire notre prime pure.

Le perceptron multicouches

Un perceptron multicouches est une structure composée de plusieurs couches de neurones et où la sortie d'un neurone d'une certaine couche devient l'entrée d'un neurone de la couche suivante.

La dernière couche est appelée la couche de sortie. Elle possède une fonction d'activation différente des autres couches en fonction du type de problème traité (classification ou régression).

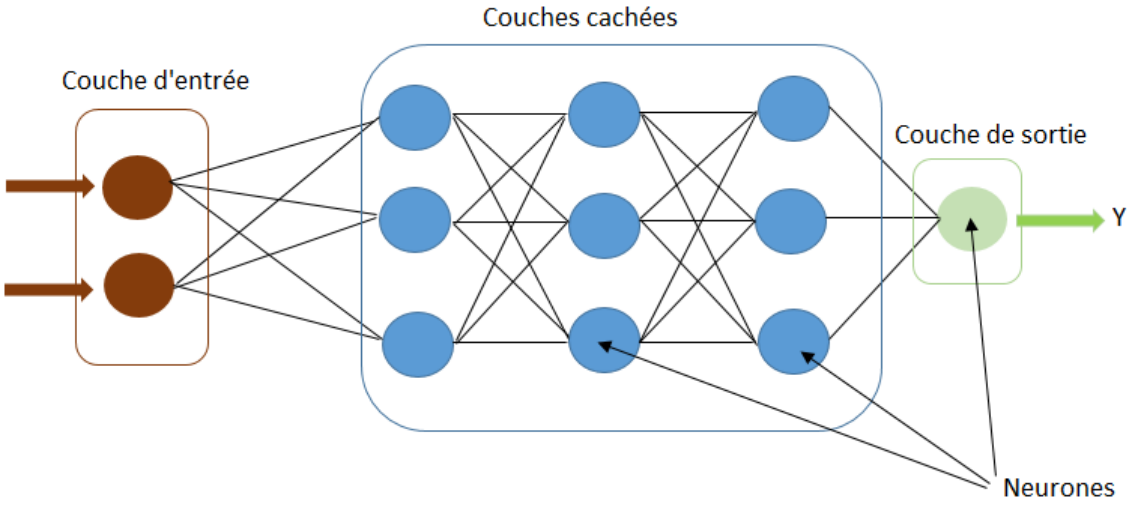

FIGURE 37 – Représentation du perceptron multicouches

En posant N_j le nombre de neurones de la couche j, on a :

1. Couche d'entrée 0 : $g^0(x) = x$ avec $x \in R^p$
2. Couches cachées $j = 1 \dots J$:
   — N_j neurones sont crées :
   $\forall i = 1 \dots N_j$, on a :

$$u_i^{(j)}(x) = m_i^{(j)} + \sum_{k=1}^{N_{j-1}} \beta_{i,k}^{(j)} g_k^{(j-1)}(x)$$

   et donc :

$$u_i^{(j)}(x) = m_i^{(j)} + < \beta_i^{(j)}, g^{(j-1)}(x) >$$

   avec :

$$g_i^{(j)}(x) = \psi(u_i^j(x))$$

3. Enfin la couche de sortie : $j = J + 1$:

$$u^{(J+1)}(x) = m^{(J+1)} + \beta^{(J+1)} g^J(x)$$

   avec :

$$g^{(J+1)}(x) = \psi_{finale}(u^{(J+1)}(x))$$

avec :

— $\beta^(j)$ la matrice des poids de la j^{ieme} couche cachée comportant N_j lignes et N_{j-1} colonnes ;

— $\beta^(J+1)$ la matrice des poids de la couche de sortie comportant 1 ligne et N_{j-1} colonnes ;

— $u_i^{(j)}$ la valeur renvoyée par le i_{eme} neurone de la couche j avant l'application de sa fonction d'activation.

Dans notre cas où on cherche à prédire une prime pure, ie un problème de régression, la fonction d'activation de la couche de sortie $\psi^{(J+1)}$ est l'identité.

6.2 Théorie

Revenons sur le perceptron multicouches présenté ci-dessus. Un des théorèmes fondateurs du succès des réseaux de neurones est le suivant :

Théorème d'approximation universelle

Soit $g : [0,1]^n \to [0,1]^m$, $\forall \mu > 0, \exists M$ tel que :

$$\|M - g\| < \mu$$

avec M un réseau de neurones à une couche cachée

Conséquences du théorème
Ce théorème implique que toute fonction bornée peut être approchée par un réseau de neurones. Il existe donc un réseau de neurones qui peut potentiellement résoudre n'importe quel problème impliquant une fonction bornée.
En revanche, on ne connait pas le nombre de neurones de la couche cachée et il doit être cherché soit par tâtonnement, soit par des techniques d'optimisation que nous verrons par la suite.

6.2.1 Estimation des paramètres

Une fois que notre réseau de neurones a été mis en place (choix du nombre de couches cachées, du nombre de neurones par couches etc), les paramètres internes du réseau doivent être estimés, à savoir la matrice des poids $\beta^{(j)}$ et les biais (m^i).

On maximise généralement la log-vraisemblance ou on minimise la fonction de perte donnée par la relation $F_{perte} = -logvraisemblance$.
Dans notre cas de régression , on a :

$$Y \hookrightarrow N(f(x, \beta^{(j)}), I)$$

Comme le modèle est gaussien, maximiser la logvraisemblance revient à minimiser la perte quadratique :

$$l(f(x, \beta^{(j)}), Y) = \|Y - f(x, \beta^{(j)})\|^2$$

Il y a par la suite plusieurs étapes à respecter pour estimer et optimiser les paramètres du réseau :

Etape 1 : Initialisation
On initialise les poids (β_i) et les biais (m_i) (souvent de façon aléatoire) et on définit un taux d'apprentissage ϵ.

Etape 2 : Propagation forward
On part de la première couche vers la dernière couche et on calcule les $u_j^{(j)}(x_k)$ et les $g_i^{(j)}(x_k)$ de façon itérative.

Etape 3 : Propagation backward
On part de la dernière couche vers la première couche et on calcule les dérivés partielles par descente du gradient. On met grâce à cela à jour par récurrence les poids et les biais de chaque couche.
Pour cela, on cherche à minimiser la fonction de perte F, on utilise la descente du gradient.
En posant $\theta = (\beta, m)$ les paramètres du réseau, on itère sur :

$$\theta := \theta - \epsilon \nabla F(\theta)$$

avec θ le taux d'apprentissage qui permet de contrôler la convergence dans la descente du gradient.

On cherche donc à calculer pour chaque donnée k de l'échantillon d'apprentissage :

$$\frac{\partial F_k}{\partial m_i^{(j)}} \text{ et } \frac{\partial F_k}{\partial \beta_i^{(j)}}$$

avec $\beta_i^{(j)}$ le poids du neurone i de la couche j et $m_i^{(j)}$ le biais du neurone i de la couche j.
Cas de la couche finale J :

$$\frac{\partial F_k}{\partial \beta_i^{(J)}} = \frac{\partial F_k}{\partial \widehat{Y_{k_i}}} \frac{\partial \widehat{Y_{k_i}}}{\partial \beta_i^{(J)}} \text{ et } \frac{\partial F_k}{\partial m_i^{(J)}} = \frac{\partial F_k}{\partial \widehat{Y_{k_i}}} \frac{\partial \widehat{Y_{k_i}}}{\partial m_i^{(J)}}$$

avec $\widehat{Y_{k_i}} = \psi(u_i^{(J)} = \beta_i^{(J)} g^{(J-1)} + m_i^J)$ en reprenant les mêmes notations que dans la section 5.1. et :

$$\frac{\partial \widehat{Y_{k_i}}}{\partial \beta_i^{(J)}} = \frac{\partial \widehat{Y_{k_i}}}{\partial g_i^{(J)}} \frac{\partial g_i^{(J)}}{\partial \beta_i^{(J)}} = \psi'(u_i^{(J)}) g^{(J-1)} \text{ et } \frac{\partial \widehat{Y_{k_i}}}{\partial m_i^{(J)}} = \frac{\partial \widehat{Y_{k_i}}}{\partial g_i^{(J)}} \frac{\partial g_i^{(J)}}{\partial m_i^{(J)}} = \psi'(u_i^{(J)})$$

donc au final en remplaçant :

$$\frac{\partial F_k}{\partial \beta_i^{(J)}} = \frac{\partial F_k}{\partial \widehat{Y_{k_i}}} \psi'(u_i^{(J)}) g^{(J-1)} \text{ et } \frac{\partial F_k}{\partial m_i^{(J)}} = \frac{\partial F_k}{\partial \widehat{Y_{k_i}}} \psi'(u_i^{(J)})$$

On s'ocupe ensuite des autres couches j :

$$\frac{\partial F_k}{\partial \beta_i^{(j)}} = \frac{\partial F_k}{\partial g_i^{(j)}} \frac{\partial g_i^{(j)}}{\partial \beta_i^{(j)}} = \frac{\partial F_k}{\partial g_i^{(j)}} \psi'(u_i^{(j)}) g^{(j-1)} \text{ et } \frac{\partial F_k}{\partial m_i^{(j)}} = \frac{\partial F_k}{\partial g_i^{(j)}} \frac{\partial g_i^{(j)}}{\partial m_i^{(j)}} = \frac{\partial F_k}{\partial g_i^{(j)}} \psi'(u_i^{(j)})$$

avec :

$$\frac{\partial F_k}{\partial g_i^{(j)}} = \sum_l \frac{\partial F_k}{\partial g_l^{(j+1)}} \frac{\partial g_l^{(j+1)}}{\partial g_i^{(j)}}$$

et :

$$g_l^{(j+1)} = \psi(u_l^{(j+1)} = \sum_{l'} (\beta_l^{(j+1)})_{l'} g_{l'}^{(j)} + m_l^{(j+1)})$$

donc :

$$\frac{\partial g_l^{(j+1)}}{\partial g_i^{(j)}} = (\beta_l^{(j+1)})_i \psi'(u_l^{j+1})$$

Grâce à ces calculs on a les valeurs de $\frac{\partial F_k}{\partial \beta_i^{(j)}}$ et $\frac{\partial F_k}{\partial m_i^{(j)}}$. On peut donc mettre à jour les poids et les biais de chaque neurone de chaque couche :

$$\beta_i^j := \beta_i^j - \epsilon \frac{1}{N_{echantillon}} \sum_k \frac{\partial F_k}{\partial \beta_i^j}$$

et :

$$m_i^j := m_i^j - \epsilon \frac{1}{N_{echantillon}} \sum_k \frac{\partial F_k}{\partial m_i^j}$$

— si ϵ est trop petit, la convergence peut être lente et le minimum trouvé peut être local ;

— si ϵ est trop grand, les valeurs peuvent osciller autour d'un point optimal sans se stabiliser.

Arrêt de l'entrainement

Pour entrainer un réseau de neurones, il est nécessaire de séparer la base de départ en trois parties :

— la base d'entrainement qui va nous servir à estimer les poids et biais notre réseau de neurones ;

— la base de validation : le réseau de neurones de la base d'entrainement va être testé sur la base de validation pour détecter tout suraprentissage ;

— la base de test qui nous permet de mesurer la performance finale de notre réseau.

Le suraprentissage désigne le fait pour un réseau de neurones de "trop" apprendre la base d'entrainement au point où ce dernier n'est plus capable de généraliser ses connaissances sur d'autres données.
On détecte le suraprentissage en regardant la fonction de perte de la base d'entrainement et de la base de validation. Le changement de croissance de la fonction de perte de la base de validation (décroissante vers croissante) nous indique que le réseau "surapprend" et qu'il faut donc arrêter le nombre d'itérations au point de croissance de la courbe.
On a donc un critère d'arrêt de l'entrainement de notre réseau de neurones.

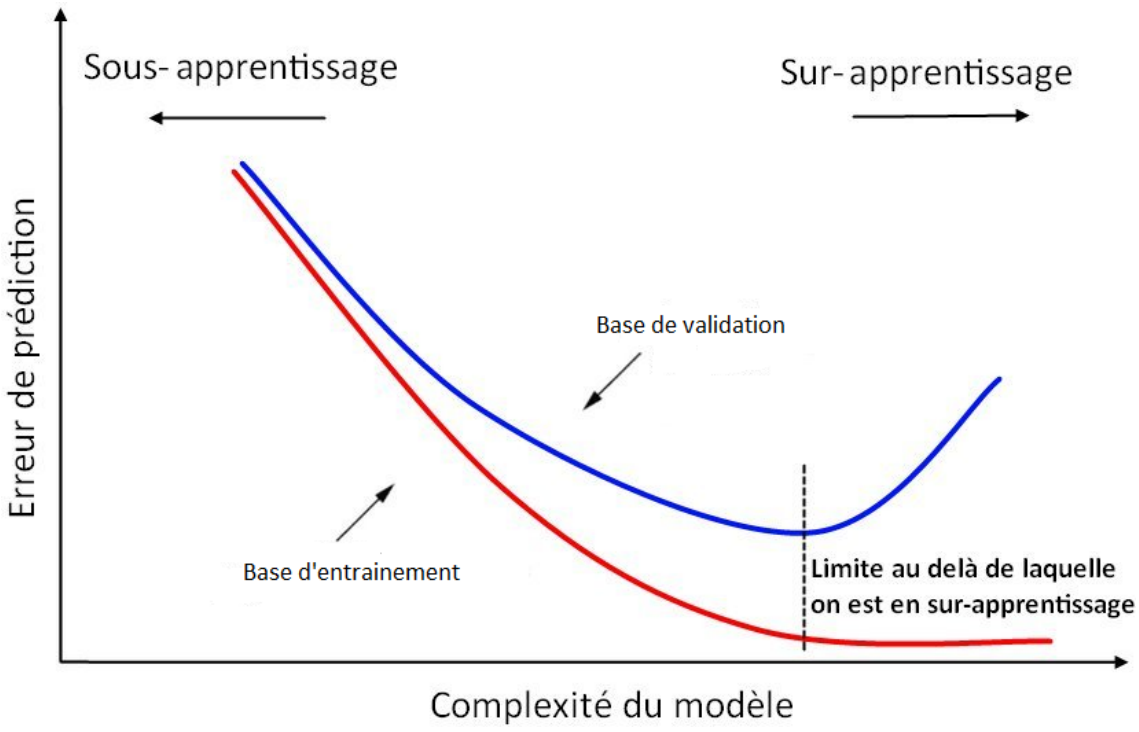

FIGURE 38 – Représentation du surapprentissage

6.3 Prétraitement de la base

Avant d'entrainer notre réseau, il est nécessaire de traiter notre base de données. Un réseau de neurones ne prenant que des variables continues en entrée, on doit d'abord transformer toutes nos données catégorielles. Pour cela il existe deux méthodes généralement utilisées :

Encodage one-hot
Pour toutes les variables catégorielles de notre base de données, on transforme chacune deleurs modalités en variables binaires.
L'avantage de cette méthode est qu'il n'y a pas de perte d'information contrairement à une transformation quantitative vers catégorielle.
L'inconvénient est que le nombre de variables peut exploser et de ce fait ralentir considérablement la vitesse de calcul. Dans notre base de données nous avons 13 variables catégorielles avec au total 57 modalités. Il y aurait donc 57 nouvelles variables dans notre modèle.

Formule
Formule 1
Formule 2
Formule 2
Formule 3
Formule 1

Formule 1	Formule 2	Formule 3
1	0	0
0	1	0
0	1	0
0	0	0
1	0	0

FIGURE 39 – Principe de l'encodage one-hot

Encodage variable cible

Une autre méthode consiste à remplacer les modalités par la moyenne de la variable cible sur chacune de ces modalités. Cette méthode présente l'avantage de ne pas ajouter de variables supplémentaires dans l'analyse.
Elle permet de plus de respecter le poids que chacune des modalités a sur la variable cible qui est le coût sinistre ici.

Formule	Coût sinistre
Formule 1	0
Formule 2	200
Formule 2	1000
Formule 3	100
Formule 1	500

Formule
250
600
600
100
250

FIGURE 40 – Principe de l'encodage variable cible

On choisira donc cette dernière méthode dans la suite du mémoire.

Normalisation

L'étape de normalisation des variables explicatives est essentielle dans un réseau de neurones car elle permet d'approcher plus rapidement les minima globaux dans l'estimation des poids et des biais de chacun des neurone en allégeant les calculs lors de la propagation forward.
Une autre raison est de redimensionner les valeurs extrêmes qui peuvent avoir un poids trop important dans les prédictions.

A chacune de nos variables, on applique la transformation :

$$X_i^{normalise} = \frac{X_i - m_i}{\sigma_i}$$

avec m_i la moyenne de la variable X_i et σ_i son écart-type

6.4 Implémentation du modèle

L'implémentation de notre réseau de neurones se fera sur python avec la bibliotheque TensorFlow et l'API Keras.

Structure du réseau
Notre réseau se compose de :

- 4 couches cachéees ;
- 150 neurones par couche ;
- fonction d'activation des neurones des couches cachées ReLU ;
- fonction d'activation du neurone de sortie linéaire
- optimisateur Adam avec un learning rate de 0.01 ;
- dropout de 0.2 sur chacune de nos couches cachées ;

Adam est un des optimisateur les plus populaires et est une méthode améliorée de mise à jour des poids et biais par rapport à la descente de gradient classique.
Le dropout de 0.2 implémenté est un procédé qui donne une probabilité de 0.2 à un neurone d'être ignoré lors de l'apprentissage du réseau. Cela permet d'éviter un surapprentissage rapide et d'augmenter la capacité de généralisation du réseau.

6.5 Optimisation

Dans le réseau de neurones crée ci-dessus nous avons choisi arbitrairement nos hyperparamètres comme le nombre de neurones et de couches cachées. Il existe plusieurs méthodes d'optimisation pour trouver la bonne combinaison d'hyperparamètres de manière à augmenter le pouvoir de prédiction de notre réseau. Les méthodes que nous verrons chercheront à trouver les bons hyperparamètres parmi :

- le dropout
- le nombre de couches cachées
- le nombre de neurones par couche
- la fonction d'activation des couches cachées
- le learning rate de la méthode Adam

Gridsearch
La méthode Gridsearch consiste à tester toutes les combinaisons possibles d'hyperparamètres à notre disposition.
L'avantage de ce procédé est qu'on trouve forcément à la fin le meilleur réseau de neurones possible. L'inconvénient majeur est que sur un nombre d'hyperparamètres important à optimiser, le temps de calcul explose et est en pratique inatteignable. Il est donc à privilégier pour un faible nombre d'hyperparamètres à gérer.

Recherche aléatoire

La recherche aléatoire est comme son nom l'indique une combinaison aléatoire de nos hyperparamètres. Elle donne de bons résultats en pratique mais l'inconvénient est qu'on ignore l'information donnée par les combinaisons d'hyperparamètres précédentes. Elle peut aussi avoir des résultats qui fluctuent beaucoup d'un test à un autre. La méthode est à privilégier sur un problème d'optimisation avec beaucoup d'hyperparamètres.

Optimisation bayésienne

L'optimisation bayésienne se base sur l'inférence bayésienne, c'est-à-dire qu'elle se sert de l'information connue pour déduire des résultats. Cette méthode est très utilisée en machine learning car avec peu de donnéees, elle donne des très bon résultats tout en étant peu couteuse en calculs.
A partir de points déjà connus, la méthode génère une distribution de probabilité de moyenne μ et d'écart-type σ en utilisant les processus gaussiens.

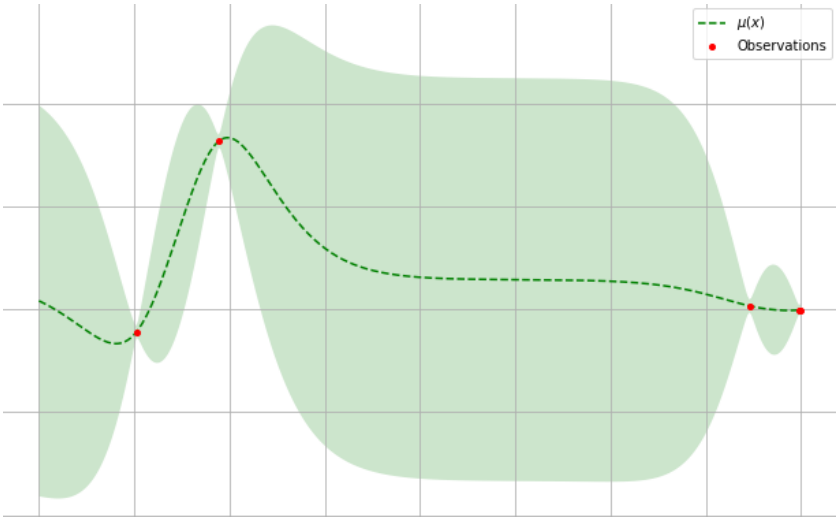

FIGURE 41 – Distribution de probabilité pour chaque point calculé

L'objectif final est de minimiser la fonction de perte, c'est-à-dire de chercher un point qui une fois évalué par la fonction de perte nous donne le minimum de cette fonction. Pour cela, on évalue une série de points en respectant à chaque fois deux principes :

- l'exploration : on veut chercher le prochain point à évaluer dans des endroits encore peu explorés, c'eest-à-dire avec un écart-type σ grand ;
- l'exploitation : on veut trouver concrètement le point qui minimise la fonction, c'est-à-dire chercher dans des endroits où l'espérance μ est petite.

On utilise pour cela la fonction d'acquisition :

$$A(x) = -\mu(x) + \lambda \times \sigma(x)$$

avec λ le poids donné à l'écart-type

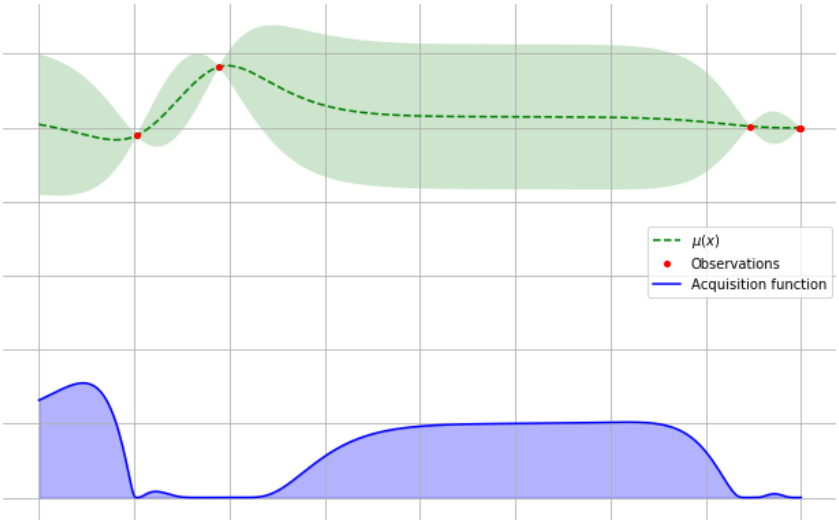

FIGURE 42 – Fonction d'acquisition calculée

A chaque étape de l'optimisation, le point choisi est celui qui maximise notre fonction d'acquisition. On doit donc calculer la fonction d'acquisition sur l'espace de recherche entier pour y trouver le maximum.
Comme le calcul de la fonction d'acquisition est beaucoup moins coûteux que l'évaluation de la fonction observée. On peut l'évaluer en un grand nombre de points dans le but de trouver le maximum. Ce maximum sera le prochain point à tester et une fois l'observation réalisée, il suffit de recalculer les distributions de probabilités en y ajoutant le nouveau point puis réitérer la procédure jusqu'à convergence.

Population based training
La méthode "Population Based Training" (PBT) est une approche d'optimisation des hyperparamètres pour les réseaux de neurones, conçue pour améliorer l'efficacité et la rapidité de la recherche d'hyperparamètres optimaux par rapport aux méthodes traditionnelles.
La méthode PBT repose sur le concept d'évolution des populations. Considérons un ensemble d'individus (modèles de réseaux de neurones) M_i avec des configurations d'hyperparamètres distinctes $\boldsymbol{\theta}_i$. Les modèles sont formés en parallèle sur un certain nombre d'époques.
À intervalles réguliers, les performances de chaque modèle sont évaluées. Les modèles les moins performants sont remplacés par les meilleurs modèles de la population. Les hyperparamètres des nouveaux modèles sont légèrement modifiés pour explorer de nouvelles combinaisons d'hyperparamètres. Considérons une population de modèles : $M_i, i = 1, ..., N$ avec des hyperparamètres correspondants : $\boldsymbol{\theta}_i$
et des performances évaluées par une fonction de coût : $L(M_i, \boldsymbol{\theta}_i)$

À chaque étape de sélection, les modèles les moins performants sont remplacés par les meilleurs modèles de la population :

Si $L(M_i, \boldsymbol{\theta}_i) > L(M_j, \boldsymbol{\theta}_j)$ alors :

$M_j \leftarrow M_i$
$\boldsymbol{\theta}_j \leftarrow \boldsymbol{\theta}_i$
Les hyperparamètres des nouveaux modèles sont légèrement modifiés : $\boldsymbol{\theta}_j \leftarrow \boldsymbol{\theta}_j + \epsilon$ où $\epsilon \sim \mathcal{N}(0, \sigma^2)$

Algorithm 1 Population Based Training (PBT)

procedure TRAIN($\mathcal{P}$) ▷ initial population $\mathcal{P}$
  for $(\theta, h, p, t) \in \mathcal{P}$ (asynchronously in parallel) **do**
    while not end of training **do**
      $\theta \leftarrow \texttt{step}(\theta|h)$ ▷ one step of optimisation using hyperparameters h
      $p \leftarrow \texttt{eval}(\theta)$ ▷ current model evaluation
      if $\texttt{ready}(p, t, \mathcal{P})$ **then**
        $h', \theta' \leftarrow \texttt{exploit}(h, \theta, p, \mathcal{P})$ ▷ use the rest of population to find better solution
        if $\theta \neq \theta'$ **then**
          $h, \theta \leftarrow \texttt{explore}(h', \theta', \mathcal{P})$ ▷ produce new hyperparameters h
          $p \leftarrow \texttt{eval}(\theta)$ ▷ new model evaluation
        end if
      end if
      update $\mathcal{P}$ with new $(\theta, h, p, t+1)$ ▷ update population
    end while
  end for
  return θ with the highest p in $\mathcal{P}$
end procedure

FIGURE 43 – Principe de l'algorithme PBT

Dans le contexte de l'optimisation des hyperparamètres d'un réseau de neurones, la méthode PBT permet d'ajuster efficacement les hyperparamètres tels que le taux d'apprentissage α, la régularisation λ, la taille des lots et d'autres paramètres spécifiques à l'architecture du réseau, afin d'obtenir la meilleure performance possible.

Comparaison des optimisations usuelles VS PBT

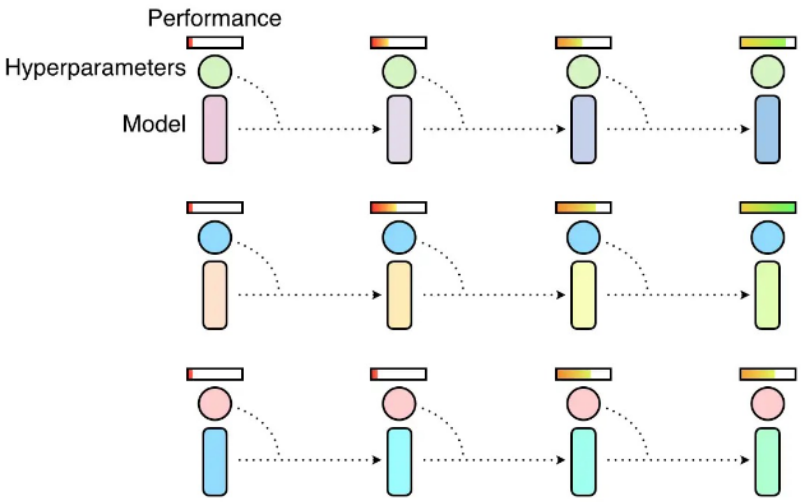

FIGURE 44 – Recherche aléatoire des hyperparamètres, où de nombreux hyperparamètres sont testés en parallèle et indépendamment. Certains hyperparamètres mèneront à des modèles avec de bonnes performances, tandis que d'autres ne le feront pas.

FIGURE 45 – Des méthodes telles que l'optimisation "manuelle" et l'optimisation bayésienne apportent des modifications aux hyperparamètres en observant de nombreuses exécutions d'entraînement de manière séquentielle, rendant ces méthodes lentes.

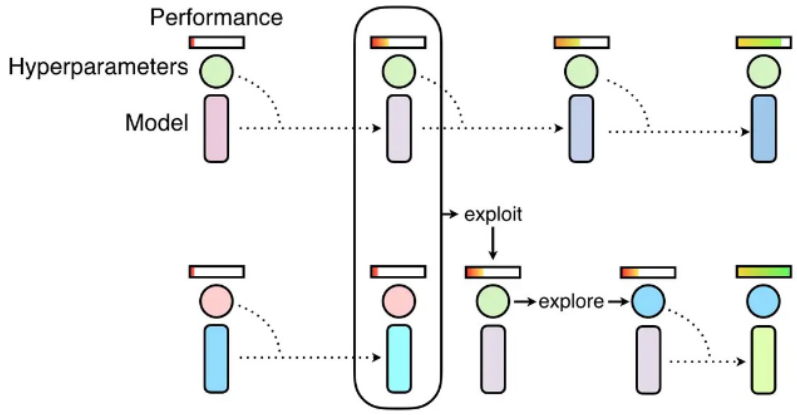

FIGURE 46 – L'entraînement de réseaux de neurones basé sur PBT commence comme une recherche aléatoire, mais permet aux différents MLP d'exploiter les résultats partiels des autres MLP et d'explorer de nouveaux hyperparamètres au fur et à mesure de la progression de l'entraînement.

A l'issue de cette étape d'optimisation, nous obtenons deux modèles, issus respectivement de l'optimisation PBT et bayésienne. Les hyperparamètres finaux obtenus par les deux méthodes sont les suivants :

Optimisation bayésienne

— le nombre de couches cachées : 19

— le nombre de neurones par couche : 113

— la fonction d'activation des couches cachées : *sigmoide*

— le learning rate de la méthode Adam : 0.0021

— le dropout : 0.05

Optimisation PBT

— le nombre de couches cachées : 7

— le nombre de neurones par couche : 318

— la fonction d'activation des couches cachées : *sigmoide*

— le learning rate de la méthode Adam : 0.0004

— le dropout : 0.1

Les paramètres à optimiser ont été choisi de manière arbitraire mais il est possible de rechercher l'optimalité sur plusieurs autres facteurs tels que le type d'optimiseur (Adam, SGD, RMSprop, etc.), la taille des batchs d'entrainement ou bien encore le nombre d'époques d'entrainement.

7 Analyse des résultats

Nous allons comparer maintenant les performances de prédiction de notre réseau de neurones avec celles du modèle GLM que nous avons mis en place dans la partie 4.

Nous nous concentrons principalement sur l'indicateur d'erreur quadratique moyenne c'est-à-dire :

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$$

Dans notre contexte, notre MSE est de la forme :

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

avec n le nombre de contrats de notre base de test, y_i les valeurs observées de notre base de test et $\hat{y}_i$ les valeurs prédites par nos modèles.

7.0.1 Sélection préliminaire GLM C-F vs Tweedie

Nous sélectionnons tout d'abord la méthode la plus performante au regarde de l'écart quadritique moyen entre le modèle Coût-Fréquence et le modèle Tweedie :

	Prime pure observée	Prime pure GLM (C-F)	Prime pure Tweedie
Moyenne	193€	205€	213€
Prime pure maximale	/	2 423€	4317€
Prime pure minimale	/	7€	15€
Ecart quadratique moyen	/	91 550	91 622

TABLE 18 – Comparaison entre GLM C-F et Tweedie

On remarque que globalement le modèle cout-fréquence a une erreur quadratique moyenne plus faible que le modèle Tweedie. Ses primes pures maximales et minimales sont de plus plus resserrées au tour de la moyenne. Il semblerait donc que l'hypothèse d'indépendance du coût et de la fréquence sinistre n'ait un impact significatif sur la qualité des prédictions.
Nous choisissons donc lle modèle Coût-Fréquence pour la suite.

7.0.2 Sélection préliminaire du meilleur réseau de neurones

Nous sélectionnons tout d'abord le réseau de neurones le plus abouti parmi celui entrainé avec la méthode bayésienne (BAY) et celui entrainé avevc la méthode "population based training" (PBT).

	Prime pure observée	PP RN (PBT)	PP RN (BAY)
Moyenne	193€	189€	200€
Prime pure maximale	/	949€	725€
Prime pure minimale	/	12€	2€
Ecart quadratique moyen	/	91 539	91 472

TABLE 19 – Comparaison entre les deux réseaux de neurones

En examinant l'écart quadratique moyen, on observe que l'optimisation bayésienne surpasse l'optimisation PBT. À première vue, cela peut sembler inattendu, étant donné que la littérature scientifique présente souvent la méthode PBT comme étant la meilleure. Cependant, il est important de tenir compte du volume de données, des conditions initiales de l'entraînement du réseau et de la variabilité des résultats obtenus avec chaque méthode.
Nous choisissons donc le réseau de neurone entrainé grâce à l'optimisation bayésienne pour la suite.

7.0.3 Analyse globale

	Prime pure observée	Prime pure GLM	Prime pure RN
Moyenne	193€	205€	200€
Prime pure maximale	/	2 423€	725€
Prime pure minimale	/	7€	2€
Ecart quadratique moyen	/	91 550	91 472

TABLE 20 – Comparaison globale GLM et réseaux de neurones

On remarque que l'erreur quadratique moyenne est plus faible dans le modèle réseaux de neurones que dans le modèle GLM.
La prime pure maximale calculée est plus faible pour les réseaux de neurones, ce qui est plus en raccord avec la réalité.
Enfin la prime moyenne est légèrement plus proche de celle observé pour notre modèle neuronal.

7.0.4 Analyse par profil de risque

Nous examinons les divers classes de risque contenus dans notre portefeuille et établissons un classement des 10 risques les plus fréquents. Ensuite, nous évaluons la précision des prédictions de nos deux modèles sur ces profils spécifiques.

Risques	PP observée	PP GLM	PP RN	Ecart GLM	Ecart RN
Risques 1	187€	387€	130€	106.95%	-30.48%
Risques 2	203€	391€	296€	92.61%	45.81%
Risques 3	196€	169€	156€	-13.78%	-20.41%
Risques 4	182€	348€	296€	91.21%	62.64%
Risques 5	198€	302€	294€	52.53%	48.48%
Risques 6	190€	385€	320€	102.63%	68.42%
Risques 7	210€	296€	216€	40.95%	2.86%
Risques 8	178€	305€	126€	71.35%	-29.21%
Risques 9	203€	321€	192€	58.13%	-5.42%
Risques 10	188€	99€	167€	-47.34%	-11.17%

TABLE 21 – Profils de risques

On remarque que l'écart est quasiment toujours plus faible pour les réseaux de neurones. Si l'on regarde au gobal, l'écart de primes pure est de 3,6% pour le cas du réseau de neurones versus 6,2% pour le cas de la GLM.

7.0.5 Analyse par variable explicative

On regarde les primes pures par modalités de quelques variables explicatives classiques en tarification automobile

Formule	PP observée	PP GLM	PP RN
Formule 1	132€	141€	143€
Formule 2	165€	184€	181€
Formule 3	304€	315€	297€

TABLE 22 – Comparaison GLM et RN variable Formule

Usage	PP observée	PP GLM	PP RN
Affaires	249€	141€	230€
Privés/retraite	192€	199€	181€
Promenade trajet	193€	160€	297€
Tous déplacements	123€	80€	297€

TABLE 23 – Comparaison GLM et RN variable Usage

Zone	PP observée	PP GLM	PP RN
2	173€	159€	181€
3	180€	198€	183€
4	227€	220€	210€
5	240€	292€	267€
6	317€	441€	358€

TABLE 24 – Comparaison GLM et RN variable Zone

CRM	PP observée	PP GLM	PP RN
0.50	171€	158€	160€
0.51-0.60	162€	174€	164€
0.61-0.70	215€	248€	232€
0.71-0.80	206€	198€	218€
0.81-0.90	203€	214€	221€
0.91-0.99	326€	228€	287€
1	317€	158€	283€
plus de 1	292€	441€	301€

TABLE 25 – Comparaison GLM et RN variable CRM

Groupe	PP observée	PP GLM	PP RN
20-27	154€	119€	123€
28-29	182€	207€	242€
30-31	214€	208€	236€
32-33	181€	221€	121€
34-38	214€	243€	224€

TABLE 26 – Comparaison GLM et RN variable Groupe

On remarque que globalement les tendances sont respectées entre les modèles et la réalité, avec une précision plus élevée chez les réseaux de neurones

7.1 Limites de chaque méthode

Chaque méthode présente des avantages et des inconvénients qu'il est bon de considérer selon la situation.

Réseaux de neurones
Les réseaux de neurones ont prédit la prime pure dans notre étude avec une plus grande précision que les GLM. La base de données utilisée ne nécessite pas de traitement majeur des variables, tel que la discrétisation des variables continues ou l'écrêtage des sinistres graves, ce qui permet de conserver une grande quantité d'information.

Un nombre important de variables explicatives n'est pas un problème pour les réseaux de neurones et augmente même la qualité des prédictions. Ainsi, les réseaux de neurones sont à privilégier pour un portefeuille complexe avec une grande quantité de données.

Cependant, la méthode présente des inconvénients, notamment le manque de transparence du réseau et la faible flexibilité dans la modification des paramètres. Contrairement aux GLM, on ne connaît pas les coefficients affectés à chaque variable, et il est difficile d'ajouter un offset au modèle pour inclure une donnée supplémentaire. De plus, si l'on doit réentraîner le réseau à la suite d'une mise à jour des données, on obtiendra un réseau avec des biais et des poids complètement différents en raison de l'aléatoire impliqué.

Cette méthode n'est donc pas recommandée pour un assureur ayant un réseau de distribution composé d'agents ou de courtiers, qui pourraient être amenés à justifier un tarif précis à un client, que ce soit lors de la souscription d'un nouveau contrat ou lors d'un avenant.

GLM
Les GLM sont des méthodes classiques de prédiction, plus anciennes que les réseaux de neurones. De ce fait, il existe un volume conséquent d'articles et d'études sur le sujet, offrant un soutien important de la littérature scientifique pour l'utilisation de cette méthode comme première approche.

La méthode GLM présente également une grande transparence et flexibilité : les coefficients affectés à chaque variable sont connus, et il est possible de les modifier manuellement afin de donner plus d'influence à certaines modalités par rapport à d'autres. Les prédictions de la prime pure obtenues avec notre modèle GLM sont de bonne qualité.

Cependant, les GLM présentent plusieurs inconvénients, notamment des hypothèses très fortes concernant l'indépendance de la fréquence sinistre et des coûts sinistres. De plus, il y a une perte d'information lors du prétraitement des variables, et l'utilisation d'une loi connue pour modéliser la fréquence ou le coût sinistre constitue une approximation forte et restrictive.

Tweedie
La méthode Tweedie est une approche de modélisation qui présente plusieurs avantages et inconvénients. D'une part, elle simplifie le processus de modélisation en combinant la fréquence et le coût des sinistres en une seule étape.
Elle prend également en compte la corrélation entre la fréquence et le coût des sinistres, ce qui peut améliorer la précision des prédictions. De plus, la distribution de Tweedie est adaptée pour modéliser des données contenant des zéros en excès, c'est-à-dire des cas où il n'y a pas de sinistre.

D'autre part, la méthode Tweedie est plus complexe à mettre en œuvre et à interpréter que la méthode Coût-Fréquence. Elle nécessite l'estimation d'un paramètre supplémentaire, le paramètre de dispersion, pour ajuster la distribution de Tweedie aux données. Enfin, cette méthode peut être moins flexible pour incorporer des informations supplémentaires ou des ajustements spécifiques.

En conclusion, les GLM offrent une transparence et une flexibilité intéressantes pour les assureurs cherchant à justifier leurs tarifs, mais présentent des inconvénients liés aux hypothèses et approximations utilisées.

La méthode Tweedie permet d'éviter certaines limitations des GLM, telles que les hypothèses d'indépendance entre les coûts et la fréquence des sinistres. Cependant, elle est généralement moins flexible dans sa mise en œuvre.

Les réseaux de neurones, quant à eux, sont plus performants pour les portefeuilles complexes avec de nombreuses données, mais présentent des inconvénients en termes de transparence et de flexibilité.

7.2 Interprétabilité des réseaux de neurones

La question de l'explicabilité des modèles de réseaux de neurones est un sujet de recherche actif dans le domaine de l'apprentissage automatique. Les réseaux de neurones profonds sont souvent considérés comme des "boîtes noires" en raison de leur complexité.

Pourtant, plusieurs méthodes tentent de les rendre plus explicables. Nous présentons en suivant quelques moyens de rendre l'intérprétation de modèles neuronaux plus aisée.

1. **Méthodes post-hoc** :
   - **LIME (Local Interpretable Model-agnostic Explanations)** : Cette méthode vise à expliquer localement des résultats spécifiques d'un réseau en introduisant du bruit dans les données de départ et en utilisant un modèle tierce qui compare les différences obtenues entre les données d'entrée et de sortie.
   - **SHAP (SHapley Additive exPlanations)** : Fondée sur la théorie des jeux, elle attribue à chaque caractéristique une valeur d'importance pour une prédiction.
2. **Techniques de régularisation** : Des méthodes comme la régularisation par la norme L1 peuvent conduire à des modèles parcimonieux et donc potentiellement plus interprétables.
3. **Modèles surrogates** : L'idée est d'entraîner un modèle simple (comme un arbre de décision) pour imiter un réseau de neurones complexe, et d'utiliser ce modèle simple comme approximation interprétable.

Cependant, même avec ces méthodes, l'interprétabilité demeure un défi, en particulier avec les architectures de plus en plus sophistiquées. L'importance relative de la performance par rapport à l'explicabilité varie selon le domaine d'application.

8 Conclusion

Tout au long de ce mémoire, nous avons exploré, implémenté et comparé deux méthodes de tarification pour la garantie RC appliquée aux produits automobiles.

Nous avons débuté par l'adaptation de notre base de données au modèle GLM à travers une analyse approfondie et un prétraitement des variables explicatives. Des techniques telles que la discrétisation de Jenks et la théorie des valeurs extrêmes ont été utilisées pour gérer les sinistres graves et estimer la charge à répartir sur l'ensemble des contrats, aboutissant à une majoration de 7€ pour chaque prime pure calculée.

Ensuite, nous avons élaboré deux modèles coût x fréquence distincts pour traiter les sinistres relevant de la convention IRSA et les autres. La combinaison de ces deux modèles nous a permis d'obtenir une prime pure couvrant tous les sinistres de la garantie RC.

Nous avons de plus fait un modèle Tweedie pour modéliser la prime pure non IRSA.

Parallèlement, nous avons développé un réseau de neurones en Python afin de calculer directement la prime pure sans dépendre des hypothèses fortes comme l'indépendance entre la fréquence et les coûts sinistres. Ce réseau de neurones a été optimisé grâce à des techniques d'optimisation bayésienne et PBT, afin d'atteindre les meilleurs résultats possibles.

En comparant les performances des trois approches, nous avons constaté que le réseau de neurones offrait une meilleure prédiction en termes d'erreur quadratique moyenne par rapport au modèle GLM. Cependant, la nature de "boîte noire" du réseau de neurones et la difficulté à l'ajuster en fonction des besoins spécifiques rendent cette méthode moins pratique pour une utilisation professionnelle.

Il est important de souligner que d'autres algorithmes de machine learning auraient pu être explorés et auraient probablement offert des performances similaires aux réseaux de neurones. Des exemples notables incluent le populaire XGBoost, fréquemment utilisé dans les compétitions Kaggle, ou encore les forêts aléatoires.

Références

[1] DENUIT M.,CHARPENTIER A. (2004) *: Mathématiques de l'assurance non-vie : Tome 1, Principes fondamentaux de théorie du risque. Economica*

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

[3] RASCHKA S.,MIRJALILI V. (2019) *: Python Machine Learning : Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2, 3rd Edition. Packt Publishing*

[4] BODIN A., RECHER F. (2021) *: Deepmath - Mathématiques (simples) des réseaux de neurones (pas trop compliqués) : Algorithmes et mathématiques*

[5] BELLINA R. (2014) *: Méthodes d'apprentissage appliquées à la tarification non-vie. Mémoire d'actuariat*

[7] ATIA R. (2016) *: Mise en place de modèles de tarification alternatifs face à la suppression règlementaire d'une variable tarifaire en automobile. Mémoire d'actuariat*

[8] KOUO K. (2017) *: Tarification automobile : GLM vs Réseaux de neurones. Mémoire d'actuariat*

[9] KOEHRSEN W. (2018) *: A Conceptual Explanation of Bayesian Hyperparameter Optimization for Machine Learning. Towards data science*

[10] ELSHEIKH21 (2019) *: population-base-training-of-NNs. GitHub*

[11] KONIG D., LOSER F. (2020) *: GLM, Neural Nets and XGBoost for Insurance Pricing. Kaggle*

[12] LIANNA, JUSTIN (2020) *: Hyperparameter Tuning with Python : Keras Step-by-Step Guide. Just into Data*

[13] CHARPENTIER A. DUTANG C. (2012) *: L'Actuariat avec R. Version numérique*

[14] BENLAGHA N., GRUN-REHOMME M., VASECHKO O. (2009) *: Les sinistres graves en assurance automobile : Une nouvelle approche par la théorie des valeurs extrêmes. Revue MODULAD*

[15] PLANCHET F., MISERAY A. (2017) *: Tarification IARD. Introduction aux techniques avancées. Ressources-actuarielles.net*

Liste des tableaux

Table des figures