Mémoire présenté le :

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

Par : Tadeusz Deslandes

Titre : Construction d'une échelle Bonus-Malus à l'aide du score de conduite

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

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

Membre présents du jury de l'Institut des Actuaires
S. BAUMANN
L. LAURENT
N. GARRIGUE

Membre présents du jury de l'ISFA
Y. SALHI

Entreprise :
Nom : Société Générale Assurances
Signature : Bruno GERIN ROZE, DRH

Directeur de mémoire en entreprise :
Nom : Thibault Van Everbroeck
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

PO Elsa CHRETIEN

Signature du candidat

Remerciements

Je tiens à remercier en premier lieu Thibault VAN EVERBROECK, responsable de l'équipe Actuariat Dommages International pour son encadrement, sa disponibilité, son expérience et son soutien moral dans cette période exceptionnelle.

Je souhaite exprimer ma gratitude envers toute l'équipe Actuariat Dommages International pour leur accueil chaleureux, pour leur bienveillance et pour leur disponibilité à me transmettre leur savoir.

Je suis reconnaissant envers l'ensemble des personnes de Société Générale Assurances qui m'ont aidé à compléter et achever cette étude.

Je souhaite remercier l'ensemble du corps professoral de l'ISFA pour le savoir et les techniques qu'ils m'ont enseigné durant ces trois dernières années et particulièrement Diana DOROBANTU pour les précieux conseils qu'elle m'a prodigués tout au long de la rédaction de ce mémoire.

Enfin, je souhaite remercier plus généralement toutes les personnes qui m'ont soutenu dans la réalisation de ce mémoire, sans lesquelles rien n'aurait été possible.

Résumé

Mots-clés : échelle Bonus-Malus, flottes de véhicules, garantie responsabilité civile, modèles linéaires généralisés, tarification, théorie de la crédibilité, télématique.

Dans le cadre du développement de nouvelles offres assurantielles, Société Générale Assurance souhaite étudier un modèle de tarification intégrant un score de dangerosité de la conduite des véhicules estimé grâce à l'information collectée à l'aide de boîtiers télématiques installés dans certains d'entre eux.

Dans le cadre des contrats d'assurance de flottes automobiles, les informations relatives aux conducteurs ne sont généralement pas disponibles. C'est la raison pour laquelle la majorité des modèles de tarification sont peu segmentés. L'objectif de cette étude est de trouver la modélisation qui optimisera l'utilisation de la variable télématique.

Nous modéliserons, au moyen des modèles linéaires généralisés, les divers éléments de la sinistralité responsabilité civile. Pour ce faire, nous décomposerons fréquence et coût moyen d'une part, sinistres matériels et corporels d'autres part. Une fois nos modèles ainsi optimisés, nous y ajouterons l'information télématique afin de pouvoir en évaluer l'impact dans les diverses modélisations. Nous observerons que le score de conduite permet une nette amélioration des modèles, excepté pour le modèle de coût moyen corporel plus volatil.

Afin d'optimiser l'utilisation de la variable télématique, nous utiliserons la théorie de la crédibilité pour construire une modélisation Bonus-Malus basée sur le score de conduite. Cette modélisation semblant moins précise qu'une modélisation Bonus-Malus basée sur la fréquence de sinistres, nous avons décidé d'élaborer une modélisation Bonus-Malus "combinée" dépendante à la fois de la fréquence de sinistres et du score de conduite.

Cette dernière modélisation combinée apparaît la plus précise et la plus pertinente pour intégrer au mieux l'information télématique.

Abstract

Keywords : Bonus-Malus scale, fleets of vehicle, Motor third party liability, generalized linear model, pricing, credibility theory, telematics.

As part of the development of new insurance products, Société Generale Insurance wants to study a pricing model incorporating a driving score for the dangerousness of driving habits, estimated using information collected by telematics boxes installed in some vehicles.

In fleet insurance contracts, driver information is usually not available. For this reason, most pricing models are not very segmented. The objective of this study is to find the model that will optimise the use of the telematics variable.

We will model with generalized linear models, the various elements of the motor third party liability claims. To do this, we will split frequency and average cost on the one hand, and material and injury claims on the other. Once our models have been optimised in this way, we will add telematics information in order to assess its impact on the various models. We observe that the driving score allows a clear improvement of the models, except for the more volatile average injury cost model.

In order to optimise the use of the telematics variable, we will use credibility theory to construct a Bonus-Malus modelling based on the driving score. As this model appears to be less accurate than a Bonus-Malus model based on the frequency of claims, we have decided to develop a "combined" Bonus-Malus model dependent on both the frequency of claims and the driving score.

This last combined model appears to be the most accurate and relevant for integrating telematics information.

Table des matières

Introduction

En France, le marché de la location moyenne et longue durée (aussi appelée leasing) de flottes automobiles est un marché en expansion à contrario de celui des ventes automobiles destinés aux particuliers. L'assurance de ces flottes automobiles est un marché estimé à 2,2 milliards d'euros de chiffre d'affaires en 2018.

Bien qu'ayant une politique tarifaire concurrentielle, les compagnies d'assurances se doivent de proposer des produits rentables. Cette volonté se traduit par la mise en place de stratégies innovantes permises par le recours aux nouvelles technologies. Dans le cadre du marché de l'assurance automobile, la télématique et le développement de voitures autonomes sont principalement étudiés.

La télématique est l'ensemble des techniques et des services qui associent les télécommunications et l'informatique. Dans le cadre de ce mémoire, seul l'usage de la télématique appliqué à l'assurance des flottes automobiles sera étudié. L'utilisation de cette technologie permet la captation d'informations sur la conduite : accélération, vitesse, localisation... Ces informations permettent une analyse individualisée du comportement des conducteurs. Grâce à ce nouvel outil, les assureurs sont désormais capables d'associer des risques à des types de conduite.

A l'heure actuelle, la télématique est principalement utilisée pour les contrats destinés aux jeunes conducteurs. Elle permet, en effet, d'anticiper la probabilité de survenance d'un sinistre pour une population n'ayant aucun antécédent de sinistralité disponible. Cette utilisation peut s'étendre aux contrats d'assurance de flottes automobiles car l'identité et l'historique des conducteurs conduisant les véhicules de la flotte ne sont pas connus de l'assureur. La tarification se base uniquement sur les informations des véhicules et éventuellement des secteurs d'activité. Le manque d'informations liées au conducteur limite considérablement l'évaluation de la sinistralité : les produits d'assurances automobiles de particuliers utilisent une dizaine de variables de segmentation liées au conducteur qui sont inconnues en assurance de flottes automobiles. La télématique permettrait ainsi de répondre à ce besoin d'information.

Le but de ce mémoire est donc d'étudier l'impact de la connaissance des habitudes de conduite dans la tarification de produits assurantiels destinés aux flottes automobiles.

Dans un premier temps, nous expliquerons le contexte de ce mémoire ainsi que la variable score de conduite qui évaluera la dangerosité de la conduite pour un véhicule. Cette information sera obtenue grâce à l'analyse conjointe de quatre indicateurs : l'accélération, la décélération, le mouvement latéral et la vitesse angulaire.

Nous étudierons ensuite l'intérêt de la variable score de conduite dans un modèle linéaire généralisé [1] sur la sinistralité matérielle d'une part, et la sinistralité corporelle d'autre part. Chacune de ces modélisations sera abordée en termes de fréquence et de coût moyen. Un GLM utilisant les variables habituelles telles que la marque du véhicule ou la zone géographique sera calibré puis comparé à un GLM intégrant la variable score de conduite. Ce parallèle permettra d'évaluer l'apport de cette dernière.

Pour approfondir, nous créerons deux modèles Bonus-Malus : un indexé sur le score de conduite et l'autre sur la sinistralité. Il en découlera un modèle Bonus-Malus conjoint indexé simultanément sur la sinistralité et le score de conduite. Les différents modèles seront ensuite comparés sur différents échantillons.

1. GLM

Première partie

Cadre de l'étude

Comme énoncé dans l'introduction, nous commencerons par présenter le cadre de l'étude : l'histoire de l'assurance automobile, les projets liés à l'utilisation de la télématique au sein du groupe Société Générale Assurance et les bases de données utiles à cette étude.

L'objectif de cette partie est de présenter le contexte et l'enjeu de ce mémoire, ainsi que les bases de données utilisées afin de clarifier leur origine et leur fiabilité. Par exemple, les données issues de boîtiers télématiques sont plus fiables que celles obtenues via des smartphones.

Cette partie servira de base pour la suite de l'étude qui concernera les modèles linéaires généralisés.

Chapitre 1

Histoire de l'assurance automobile

Historiquement, les premières assurances pour un moyen de transport apparaissent dès les civilisations grecque et romaine sous la forme de « prêts à la grosse aventure ». Si les bateaux « assurés » faisaient naufrage, les marchands n'avaient rien à rembourser aux banques. Mais s'ils arrivaient à bon port, la banque se voyait rembourser un montant défini au préalable.

Le premier contrat d'assurance comparable aux contrats actuels apparaît au XIV^e siècle. C'est une assurance maritime pour le voyage du navire Santa Clara, de Gênes à Majorque. Un siècle plus tard, la première entreprise d'assurance est créée à Gênes. Suite à des abus dans la tarification, des blocages religieux empêcheront le développement de l'assurance maritime en France jusqu'en 1681, date à laquelle on constate l'apparition des premières lois régissant ce domaine.

Les assurances vie et assurances incendie seront encadrées par des lois au début du XIX^e siècle. Par la suite, l'assurance s'est ouverte à tous les types de risques.

Les automobiles apparaissent quant à elle à la fin du $XVIII^e$ siècle. En 1900, ce sont presque 10 000 véhicules qui sont produits. En 1908, Henry Ford lance son modèle T qui sera vendu à plusieurs millions d'exemplaires grâce à la mise en place du travail à la chaîne. L'industrialisation du secteur automobile se poursuit et s'intensifie au milieu du XX^e siècle.

C'est en 1951 que la première assurance automobile voit le jour en France. La réglementation inhérente à cette nouvelle assurance apparaît à la même période. Les conducteurs n'avaient jusqu'alors aucune couverture et devaient donc indemniser les dommages causés lors d'un accident responsable sur leurs fonds personnels.

En 1951, un fonds de garantie automobile [1] est créé pour dédommager les victimes des accidents de la route n'ayant pas pu être indemnisées par la partie adverse. Ce fonds de garantie est rapidement dépassé par l'ampleur des dédommagements à fournir car les véhicules non

1. FGA

assurés sont alors les plus dangereux. Le législateur impose donc l'assurance automobile à tous dès le 27 février 1958. Cette obligation persiste actuellement sous la forme de l'assurance responsabilité civile qui dédommage la partie adverse en cas d'accident responsable. Afin de responsabiliser les conducteurs dans leur sinistralité, le Bonus-Malus est créé en 1976. Il permet de pénaliser financièrement les conducteurs avec une sinistralité importante. Ce système permet d'optimiser et de fiabiliser les modélisations des contrats d'assurance automobile.

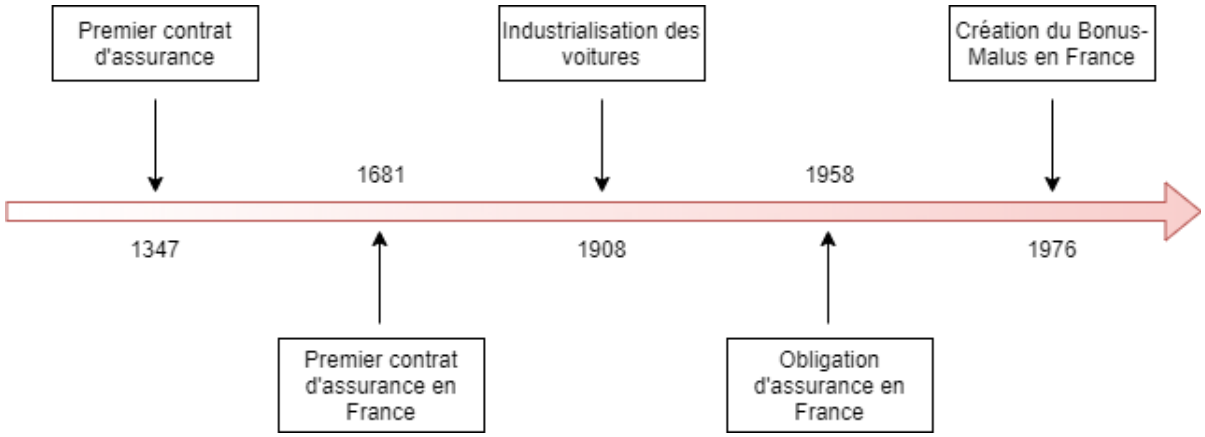

FIGURE 1.1 – Histoire de l'assurance

L'assurance automobile est désormais un secteur important et très compétitif. L'apparition de la télématique fournit aux assureurs de nouvelles possibilités de modélisation des contrats dans laquelle de nouvelles données sont à intégrer et à exploiter.

Chapitre 2

Contexte de l'étude

2.1 Introduction à la télématique

En France, la majorité des assurances automobiles sont dîtes « classiques ». Grâce à un historique de sinistres, il est possible de faire des classes de risques en fonction du profil de l'individu en utilisant notamment l'âge du conducteur, son lieu de résidence... On parle alors de segmentation. Le prix des contrats "classiques" est obtenu grâce aux caractéristiques relatives à l'individu et au bien à assurer qui, combinées, permettent de déterminer son profil de risque et la classe de tarification à laquelle il appartient.

Aujourd'hui, la méthode de segmentation la plus courante est le Modèle Linéaire Généralisé [1]. Elle s'appuie sur certaines variables caractéristiques telles que l'âge et le genre du conducteur, la puissance et l'ancienneté du véhicule.

Un GLM de fréquence est appliqué pour déterminer la probabilité de survenance d'un sinistre pour une classe d'individus. Puis, un GLM de coût est réalisé afin d'obtenir un coût théorique par sinistre. En combinant les deux résultats, le coût moyen (aussi appelé la charge sinistre) probable de chaque individu est estimé. Le montant de la prime commerciale est déterminé en ajoutant à cette charge sinistre les frais de gestion, frais de structure, le coût de la réassurance, d'éventuelles commissions ainsi que les taxes.

Cependant, dans la pratique, il existe au sein de chaque classe de conducteurs des écarts de sinistralité. Les assurés les plus prudents d'une classe d'individus voient leur prime surestimée du fait de la sinistralité plus importante d'une autre partie des individus de leur groupe. Afin de minimiser ces écarts de sinistralité, l'idée de l'hypersegmentation a vu le jour : cette approche s'apparente à une tarification individualisée. L'hypersegmentation permettrait également à l'assureur d'avoir une meilleure vision de ses assurés et donc de proposer davantage d'offres commerciales aux assurés présentant une faible sinistralité.

1. GLM

C'est dans ce contexte qu'est apparue l'idée de l'utilisation de la télématique en assurance. A l'origine, la télématique n'est pas destinée à l'assurance automobile mais à la sécurité routière. En 1988, les Etats-Unis projettent d'utiliser la télématique afin d'améliorer la sécurité routière et de proposer divers services tels que le suivi en temps réel des camions.

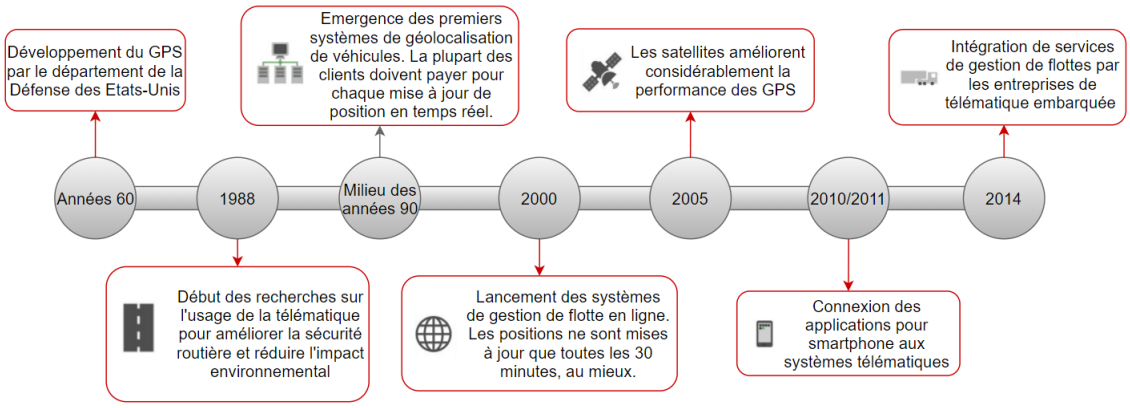

FIGURE 2.1 – Histoire de la télématique

Concrètement, la télématique étudiée dans ce mémoire se matérialise par un petit boîtier qui se branche sur le véhicule. Il contient une puce GPS, un gyroscope et une puce de communication (GSM). Les données du véhicule sont transmises en temps réel vers un serveur afin d'être traitées.

La première application utilisant ce système est le Pay As You Drive [2]. PAYD est une méthode de tarification qui est basée sur les distances parcourues en partant du principe que le risque augmente avec le nombre de kilomètres. La limite de ce système est qu'il n'intègre pas les conditions de conduite.

Afin d'intégrer cet élément, des variables caractérisant la conduite ont été prises en compte comme l'accélération et la vitesse angulaire. Ce type de tarification est Pay How You Drive [3]. C'est une tarification comportementale. Elle corrèle la manière de conduire au risque de sinistre.

Ces deux types de tarifications sont complémentaires et il est recommandé de combiner les deux pour obtenir la tarification la plus précise. PAYD permet une tarification équitable entre les assurés en ne faisant payer que les kilomètres effectués. PHYD a un objectif préventif puisqu'elle incite les assurés à bien conduire pour minimiser leur prime.

2. PAYD
3. PHYD

2.2 La télématique au sein de Société Générale Assurances

2.2.1 Présentation de Société Générale Assurances

Société Générale Assurances [4] est une société en expansion soucieuse de proposer de nouveaux produits techniques et innovants. Dans ce cadre, le recours à la télématique est étudié et ouvre de nouvelles possibilités.

La Société Générale est une firme internationale contenant de nombreuses sociétés, la principale étant la banque Société Générale. La SGA est une filiale du groupe qui gère l'ensemble de ses assurances avec deux sous-filiales : la Sogessur pour la partie non-vie et la Sogecap pour l'assurance vie.

FIGURE 2.2 – Les différentes filiales de SGA

SGA est présente en France et à l'international dans une dizaine de pays.

4. SGA

SOCIÉTÉS	PAYS
Sogécap	France
Succursale : Société Générale Insurance Allemagne	Allemagne
Succursale : Société Générale Insurance Italie	Italie
Succursale : Société Générale Insurance Pologne	Pologne
Antarius	France
Oradéa Vie	France
Sogessur	France
Succursale : Société Générale Insurance Allemagne	Allemagne
Succursale : Société Générale Insurance Italie	Italie
Succursale : Société Générale Insurance Pologne	Pologne
Succursale : Société Générale Insurance Roumanie (1)	Roumanie
Sogelife Bulgaria (2)	Bulgarie
Sogelife	Luxembourg
La Marocaine Vie	Maroc
Komerčni Pojištovna	Rep. tchèque
BRD Asigurari de Viata	Roumanie
Societe Generale Strakhovanie CSJC	Russie
Societe Generale Strakhovanie Zhizni LLC	Russie

(1) Succursale créée en 2018. Lancement de l'activité en 2019.
(2) Cession de l'entité Sogelife Bulgaria en janvier 2019 (cf. A.1.7).

FIGURE 2.3 – Sociétés composant SGA

Il existe deux principales manières pour une entreprise de s'étendre dans un pays, soit via une succursale soit via une filiale. Une succursale est une implantation de la société à l'étranger. Une filiale est une entreprise appartenant à une autre entreprise groupe ou mère. Une filiale a beaucoup plus d'autonomie juridique qu'une succursale. Toutes les décisions de la succursale doivent être approuvées par le Groupe.

Au Maroc, au Luxembourg, en République Tchèque et en Russie, les institutions présentes sont des filiales. C'est pourquoi les produits distribués dans ces pays donnent seulement lieu à une supervision technique.

En Roumanie, Allemagne, Pologne et Italie, les institutions présentes sont des succursales. Les travaux techniques et actuariels pour ces pays sont effectués au siège par la direction Technique Dommage et Prévoyance [5].

5. TDP

	Italie	Allemagne	Pologne	Roumanie
Moyens de paiement		Hanseatic Bank	Eurobank/Millenium	BRD, BRD Finance
Flottes automobiles (MTPL/ Protection du conducteur/ CASCO)	ALD Automotive, Alphabet, Car Server, Nissan, 24H Assistance, Poste, Valandro		ALD Automotive, Alphabet, ING, Hitachi, Business Lease	
MTPL Retail	24H Assistance			
CASCO Retail	Fiditalia, 24H Assistance			
Protection juridique	ALD Automotive, 24H Assistance			
Protection du conducteur	ALD Automotive, Car Server, 24H Assistance		ALD Automotive, Alphabet, ING, Hitachi, Business Lease	
Assurances voyages		Barclays	April	BRD, BRD Finance
Gap Insurance		BDK		
Pannes mécaniques	Fiditalia/Opteven			
Autres produits	Bike Insurance, Assurance Mobilité Assurance Chiens/Chats	Maxpool, Business Bike	UNOOPTIC	BRD

FIGURE 2.4 – Les différents produits non-vie commercialisés dans les succursales de SGA

2.2.2 Le projet télématique chez Société Générale Assurance

Star Drive est le premier essai de télématique au sein de SGA. Il s'agit d'une application mobile qui permet d'enregistrer le comportement de conduite des utilisateurs, de collecter des données et de promouvoir des conseils pour améliorer et sécuriser la conduite. Cette application et les données collectées ont permis de développer un produit d'assurance automobile connectée (Carapass) pour la banque Boursorama (filiale du Groupe Société Générale). Carapass est une offre PAYD. Le tarif se compose d'une partie fixe pour l'abonnement et d'une partie variable indexée sur le nombre de kilomètres parcourus.

Parallèlement, une initiative a vu le jour en Italie avec la société ALD, société de leasing automobile du groupe Société Générale, ayant pour vocation la location de véhicules à des entreprises. ALD propose à ses clients des services innovants : réduction de fraude, réduction des vols de véhicules et optimisation de l'utilisation de la flotte automobile. Ces services sont possibles grâce à des puces de localisation installées dans les véhicules et des algorithmes d'optimisation. L'intérêt pour l'entreprise locataire est de voir ses primes d'assurance baisser tout en

FIGURE 2.5 – Boîtier télématique embarqué

ayant le contrôle sur l'utilisation de son parc automobile par ses salariés et en minimisant, autant que faire se peut, les risques de dommages et d'accidents.

La flotte ALD Italie compte aujourd'hui plus de 150 000 véhicules dont 75% équipés avec un boîtier télématique. La mise en place de cet équipement a été favorisée par la segmentation tarifaire et la baisse de prix des contrats pour les véhicules équipés.

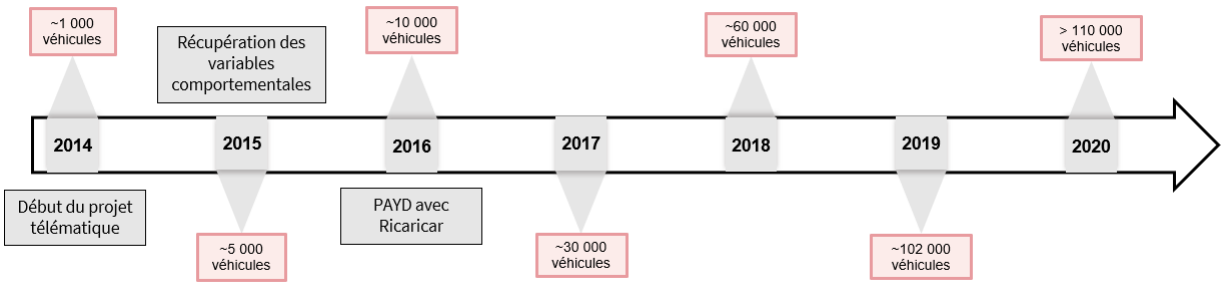

FIGURE 2.6 – Evolution du nombre de véhicules équipés au sein de la flotte ALD

Depuis 2014, ALD et SGA travaillent en étroite collaboration sur les données télématiques issues des véhicules équipés dans la flotte italienne.

Chapitre 3

Présentation des données

Cette étude utilisera les données issues de la flotte automobile ALD Italie. C'est une grande flotte automobile comprenant plus de 150 000 véhicules. Elle est assurée en totalité par SGA sur la garantie responsabilité civile [1] qui dédommage les tiers lors d'un accident responsable.

Les résultats présentés dans ce mémoire ont été modifiés par souci de confidentialité.

Cette étude porte sur les données recueillies de novembre 2017 à octobre 2020 et ne concerne que les véhicules ayant un score télématique représentant 75 % de la flotte.

3.1 Bases de données utilisées pour l'étude

Cette base de données est composée de trois groupes de données distincts : la base contrat, la base sinistre et la base télématique.

3.1.1 Base contrat

La base contrat est la base de données comportant toutes les informations des assurés et des véhicules assurés. Elle est composée de :

	Année Y6	Année Y7	Année Y8
Nombre de véhicules	167 791	197 084	177 018
Exposition	133 613	157 722	143 209

FIGURE 3.1 – Nombre de véhicules dans la base contrat

Cette base de données regroupe les variables explicatives qui seront utilisées dans les GLM :

1. RC

- **Age d'immatriculation** : nombre d'années écoulées depuis l'immatriculation.
- **Couverture maximale** : couverture maximale souscrite auprès de l'assureur.
- **Cylindrée du véhicule**.
- **Domaine d'activité** : code ATECO [2] à la maille deux chiffres permettant l'identification du domaine d'activité de l'entreprise.
- **Durée du contrat** : durée de location incluse dans le contrat de leasing.
- **Entité juridique de l'entreprise** : type d'identité morale de l'entreprise.
- **Evaluation client** : évaluation du client par le distributeur.
- **Exposition** : rapport entre le nombre de jours assurés et le nombre de jours total sur la période considérée.
- **Franchise**.
- **Lieu de location**.
- **Marque du véhicule**.
- **Nombre de chevaux fiscaux**.
- **Nombre de kilomètres au contrat** : nombre de kilomètres inclus dans le contrat de leasing.
- **Nombre de places dans le véhicule**.
- **Période d'activité**.
- **Masse du véhicule**.
- **Province**.
- **Puissance du véhicule**.
- **Puissance en kilowatt**.
- **Score de conduite** : variable comprise entre 0 et 100 représentant la dangerosité de la conduite du véhicule. Un score de 100 représente une conduite optimale alors qu'un score de 0 suppose une très mauvaise conduite.
- **Taille de la flotte** : somme des expositions des contrats composants la flotte de véhicules.
- **Type de carburant**.
- **Type de contrat de location** : anagramme en fonction du contrat de leasing souscrit par l'entreprise.
- **Zone aéroportuaire** : indique si le véhicule se trouve dans une zone aéroportuaire,

2. Code associant un nombre à des secteurs d'activités. Plus le nombre a de chiffres, plus la catégorie est affinée.

peuvant entraîner des montants de sinistres très élevés.

Ces variables permettent d'obtenir quelques statistiques sur la composition du portefeuille.

Les données récoltées permettent de connaître la répartition des véhicules en fonction de leur marque :

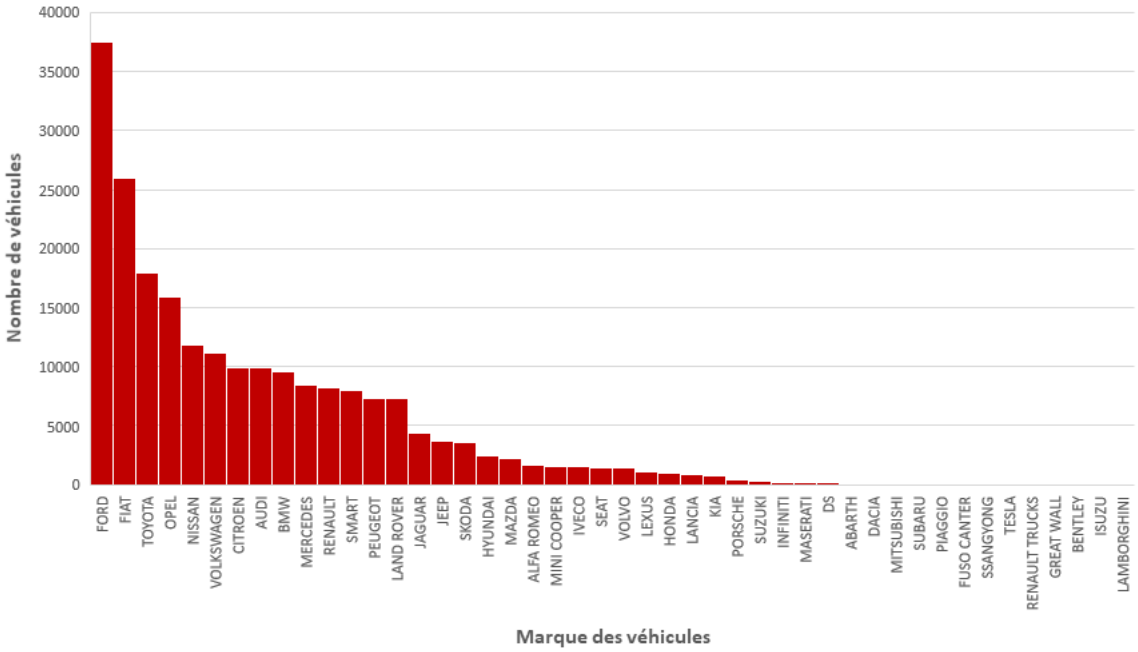

FIGURE 3.2 – Répartition selon la marque du véhicule

Les marques les plus représentées dans le portefeuille sont FORD et FIAT, des marques milieu de gamme.

Le graphique ci-dessous présente la répartition géographique des véhicules en Italie.

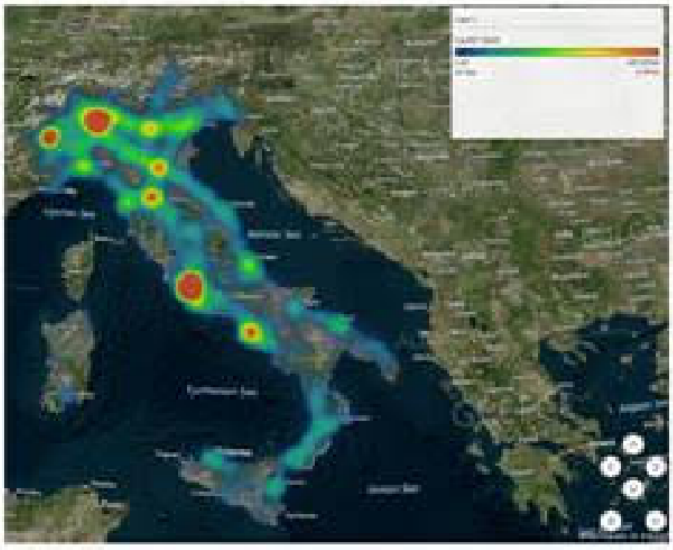

FIGURE 3.3 – Répartition géographique du portefeuille en Italie

Ce graphique montre que la majorité du portefeuille se situe près des grandes villes, notamment Rome et Milan.

Le graphique suivant décrit la répartition du portefeuille selon le domaine d'activité.

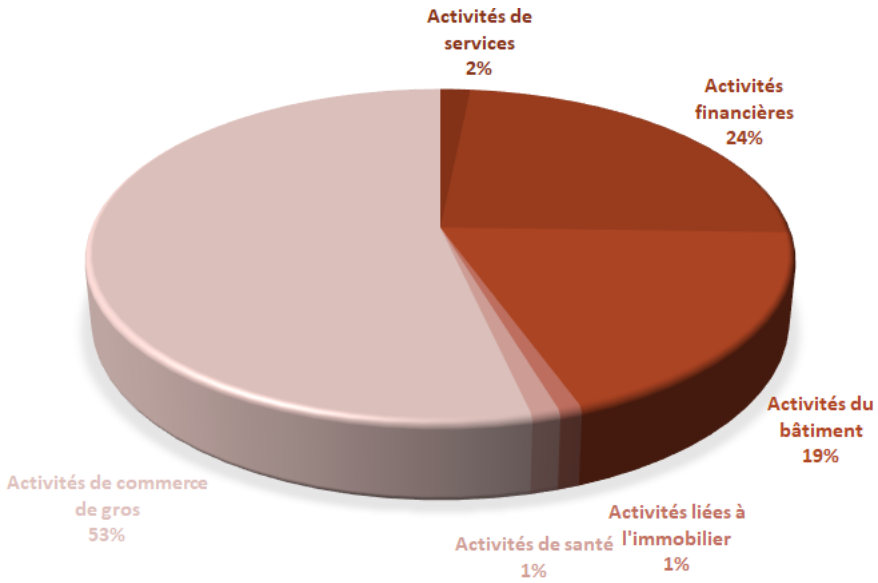

FIGURE 3.4 – Répartition du portefeuille selon le domaine d'activité

Comme le montre le graphique ci-dessus, plus de la moitié du portefeuille est à destination du commerce de gros. Ces données ont été récoltées grâce au domaine d'activité de l'entreprise

qui a loué le véhicule.

L'ensemble de ces variables seront testées ultérieurement dans nos différentes modélisations.

3.1.2 Base sinistre

La base sinistre est la base comportant toutes les informations sur les sinistres : numéro de la police (qui permet de lier les informations avec la base contrat), montant, franchise et responsabilité de chaque sinistre.

Dans cette étude, nous travaillons uniquement sur les sinistres liés à la garantie responsabilité civile : nous ne recenserons donc que les sinistres responsables ou partiellement responsables.

Grâce à cette base, nous pouvons observer la fréquence de sinistres et le coût moyen du portefeuille pour les sinistres matériels et corporels.

3.1.3 Base télématique

Les données recueillies dans cette base proviennent des capteurs télématiques présents dans les véhicules. Elles permettent d'évaluer le comportement des conducteurs en recensant différents paramètres. Concrètement, les capteurs télématiques communiquent trois types d'informations : les données trajets, les données comportementales et les crashs.

La position GPS, la date, la plaque d'immatriculation, l'identifiant du trajet, le type de route et la province forment les données trajets. L'enregistrement de ces données commence dès la mise en route du moteur et s'arrête quand le moteur s'éteint.

Les données comportementales recensent : la position GPS, l'heure, la date, la plaque d'immatriculation et le type d'évènement. Sont considérés comme évènements l'accélération, la décélération, la vitesse angulaire et les mouvements latéraux qui dépassent un certain seuil.

Les crashs partagent les mêmes libellés d'informations que les données comportementales. Cependant, les évènements des crashs correspondent à des évènements de forte intensité qui peuvent être associés à de vrais sinistres. Lorsqu'un évènement crash se produit, le capteur envoie les différentes données de l'instant et celles des 30 secondes qui ont précédé l'accident.

L'ensemble de ces données sont reçues de manière hebdomadaire. Elles représentent annuellement 325 Go de données. Aujourd'hui nous totalisons plus de 124 millions d'heures de conduite et plus de 4,24 milliards de kilomètres, soit 107 000 fois le tour du monde. C'est la raison pour laquelle un prétraitement est nécessaire à la création de la variable score.

3.2 Création de la variable "score de conduite"

Rappelons que le recensement des données télématiques a pour but la création d'une offre de tarification en fonction de la manière dont est conduit le véhicule. Cette distinction de l'offre n'est donc possible que si la manière de conduire est quantifiée grâce aux données captées. Cette manière de conduire est décomposée en niveaux compris entre 0 et 100. Plus le chiffre de score est élevé, meilleure est la conduite. Le score est inversement corrélé avec la probabilité de survenance d'un sinistre.

Trois étapes sont nécessaires à l'obtention de ce score : la phase de prétraitement des données, la phase de modélisation et enfin la phase de scoring.

La phase de prétraitement prend en entrée les données brutes et renvoie une base journalière ayant toutes les informations disponibles (contrat, sinistre et télématique). La phase de modélisation construit le modèle à partir d'un échantillon des données. Cette phase renvoie des fonctions qui permettent la création du score global. La dernière étape est l'application du modèle sur les données permettant l'obtention de la variable score de conduite.

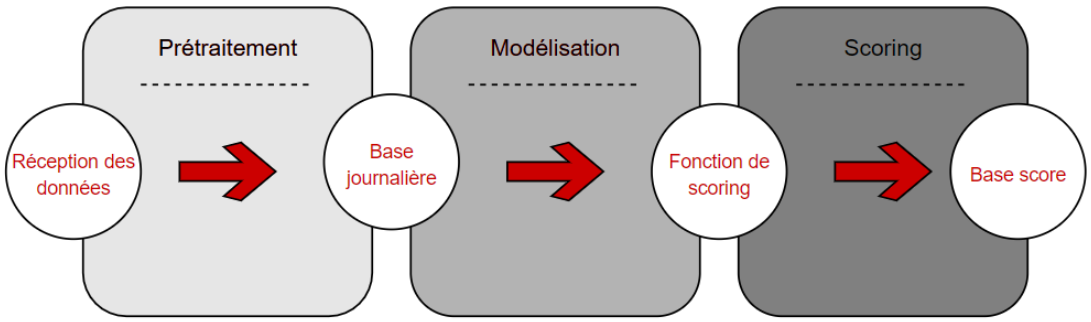

FIGURE 3.5 – Etapes de construction du score

3.2.1 Prétraitement

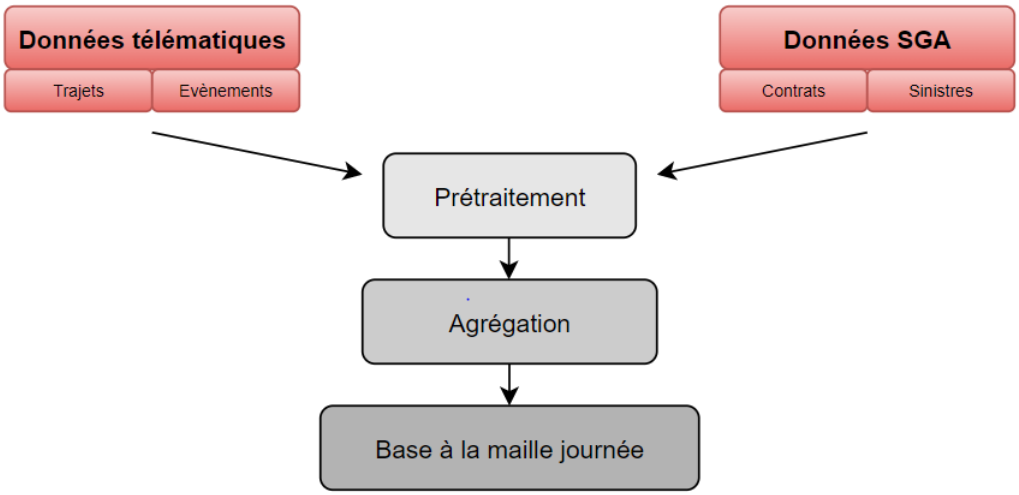

FIGURE 3.6 – Etapes du prétraitement

La première étape du travail sur les données consiste en un nettoyage de la base qui vise à supprimer les doublons, les valeurs manquantes et toutes les aberrations générées par le système.

Dans un second temps, les données sont agrégées trajet par trajet avec les données comportementales.

Enfin, ces données sont assemblées par jour et rattachées aux informations contrats et sinistres.

A la suite de cette étape, nous disposons d'une base contenant, pour chaque jour et chaque véhicule ayant circulé, les kilomètres effectués, le temps de conduite, le nombre d'évènements de chaque type ainsi que les informations contrats et sinistres.

3.2.2 Modélisation

Un modèle statistique est, en règle générale, un instrument qui prend la tendance des observations pendant une période donnée et qui permet de projeter cette tendance sur une autre période, notamment prospective.

Dans notre étude, nous cherchons à modéliser une variable score qui capte le profil de risque d'un individu en fonction de sa conduite.

Pour ce faire, on utilise un modèle de Machine Learning, le modèle XGBoost, combinant plusieurs algorithmes afin d'obtenir un résultat unique issu de tous les autres. Il travaille par

récurrence : il commence par créer un premier modèle de prédiction, donnant à chaque variable un coefficient d'ajout ou de diminution de la probabilité d'avoir un sinistre en fonction de la valeur de la variable. On peut le représenter sous la forme d'une arborescence :

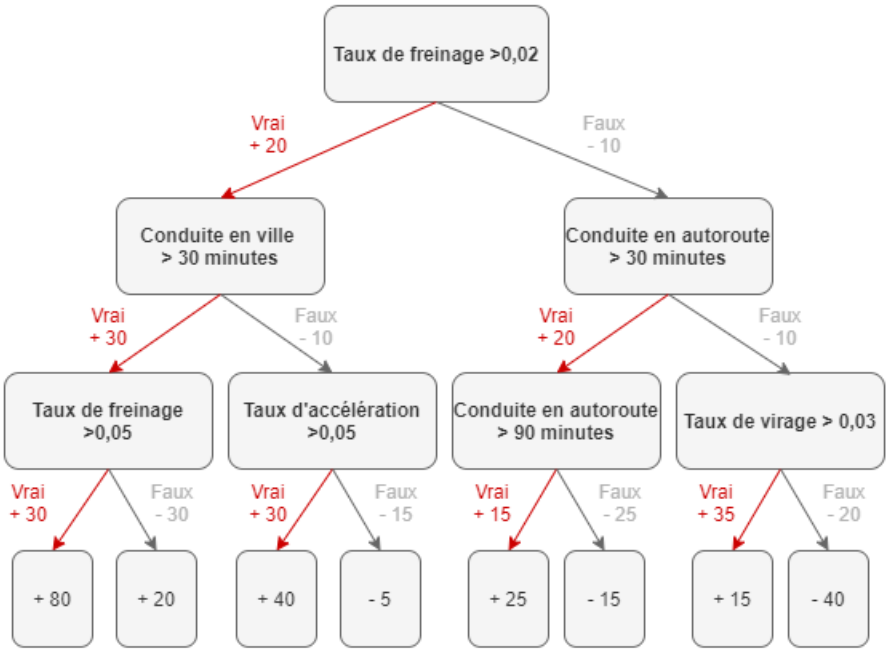

FIGURE 3.7 – Exemple d'arbre XGBoost

Notre modèle nous donne également une mesure de la significativité pour chaque individu. Ensuite nous réalisons un second modèle en prenant en compte les bonnes prédictions du premier.

Le troisième modèle fonctionne de manière similaire et conserve les bonnes prédictions des deux premiers modèles. Il est ainsi de meilleure qualité. Ce schéma se reproduit en boucle à de nombreuses reprises si bien que le modèle obtenu à la fin de l'algorithme est le meilleur possible.

Après avoir appliqué ce modèle, nous obtenons une prédiction de la sinistralité chaque jour, pour chaque véhicule, grâce aux variables explicatives et leurs modalités.

Ce modèle de Machine Learning est très efficace pour faire des prédictions issues de variables non linéaires (par exemple des variables ayant un cap sur la valeur du maximum). Il présente cependant un défaut : l'effet "boîte noire". Cet effet provient du fait qu'il prend les arguments et ressort une prédiction sans expliquer ses différentes actions, ce qui peut le rendre difficile à utiliser. Pour pallier ce défaut, nous avons modifié l'algorithme afin de récupérer l'impact de chaque variable dans chacun des modèles qu'il crée.

Pour cela, nous évaluons l'impact de chaque variable prédictive dans chaque modèle créé lors de l'algorithme XGBoost. Puis nous moyennons ces impacts afin d'obtenir l'impact moyen de chaque variable. L'impact moyen obtenu est l'impact marginal moyen de la variable.

Ces impacts marginaux permettent de créer des sous-scores à chacune des quatre variables comportementales : accélération, décélération, mouvement latéral et vitesse angulaire. Ils sont créés grâce à la formule :

$$Score_{ij} = 100 \times (1 - \frac{Min[Max(Im, q_{5\%}(Im)), q_{95\%}(Im)] - q_{5\%}(Im)}{q_{95\%}(Im) - q_{5\%}(Im)})$$

Avec :

- Im : $impact_{ij}^{event}$
- i : véhicule
- j : journée
- event : évènement intérêt
- $q_{l\%}(x)$: le quantile d'ordre l de la variable x

A présent, nous avons besoin d'outils efficaces pour appliquer le modèle sur des données d'une autre période. Ces outils sont des fonctions bâties grâce à une interpolation linéaire des sous-scores obtenus précédemment. Elles prennent en argument le taux d'évènements par minute et renvoient le sous-score associé.

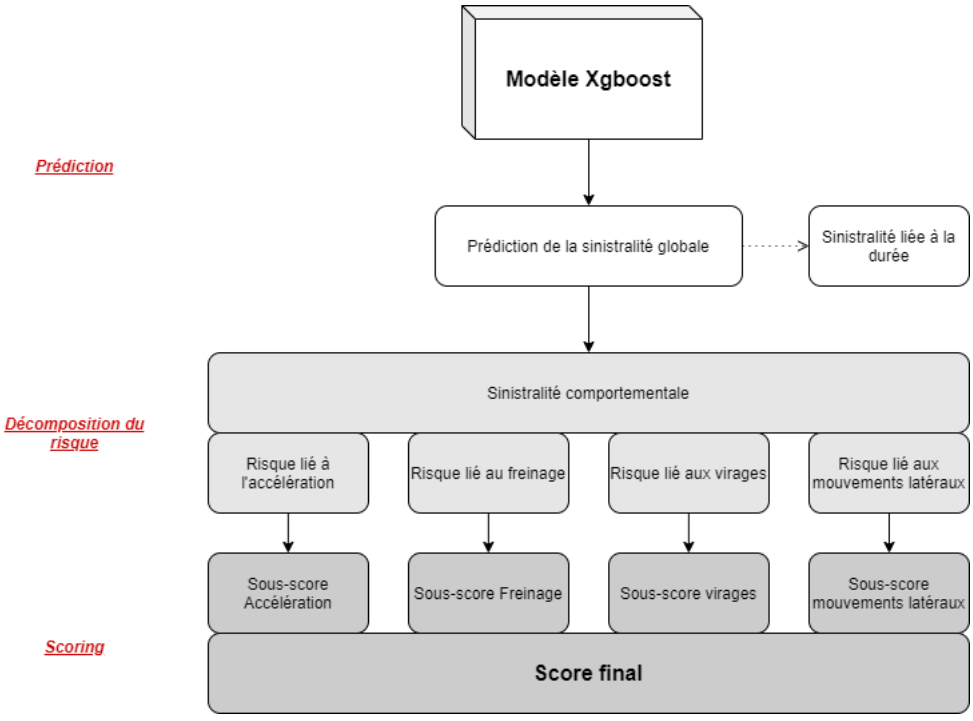

FIGURE 3.8 – Application XGBoost

A l'issue de cette étape, nous disposons de quatre fonctions de scoring associées aux quatre variables de conduite. Elles prennent en argument le nombre d'évènements de la variable et retournent un score du type d'évènement associé. Nous disposons également du poids de chaque type d'évènement.

3.2.3 Scoring

L'étape de scoring est l'étape d'application du modèle de scoring sur les données. Cela permet d'obtenir le score de conduite à partir des données télématiques.

Ce score est obtenu en deux étapes : la première étape est la création des sous-scores, la seconde la création du score global.

Pour chaque véhicule et chaque jour, on calcule les 4 taux d'évènements par minute associés aux 4 évènements. Ces taux d'évènements par minute représentent la conduite du véhicule. Afin de matérialiser la conduite du véhicule sous la forme d'un niveau compris entre 0 et 100, nous appliquons nos fonctions de sous-score précédemment calculées.

Enfin, nous créons un score global de conduite grâce à une moyenne des sous-scores pondérée par le poids de ces derniers.

Nous obtenons ainsi un score de conduite journalier par véhicule. Afin d'obtenir un score de conduite mensuel ou annuel, nous faisons une moyenne des scores du véhicule sur la période désirée, pondérée par le temps conduit chaque jour.

À ce niveau, nous disposons d'un score de conduite annuel par véhicule. Dans la suite du mémoire, nous chercherons le meilleur moyen d'utiliser ce score de conduite dans nos modélisations.

3.3 Evolution du score de conduite

Le score de conduite est une information obtenue quotidiennement, c'est pourquoi il est essentiel de s'intéresser à son évolution dans le temps.

Nous prenons la population de véhicules présents depuis 2017 et la découpons en huit selon les scores de conduite moyens : 0-30, 30-40, 40-50, 50-60, 60-70, 70-80, 80-90 et 90-100.

Observons l'évolution du score mensuel sur chacun de ces huit groupes :

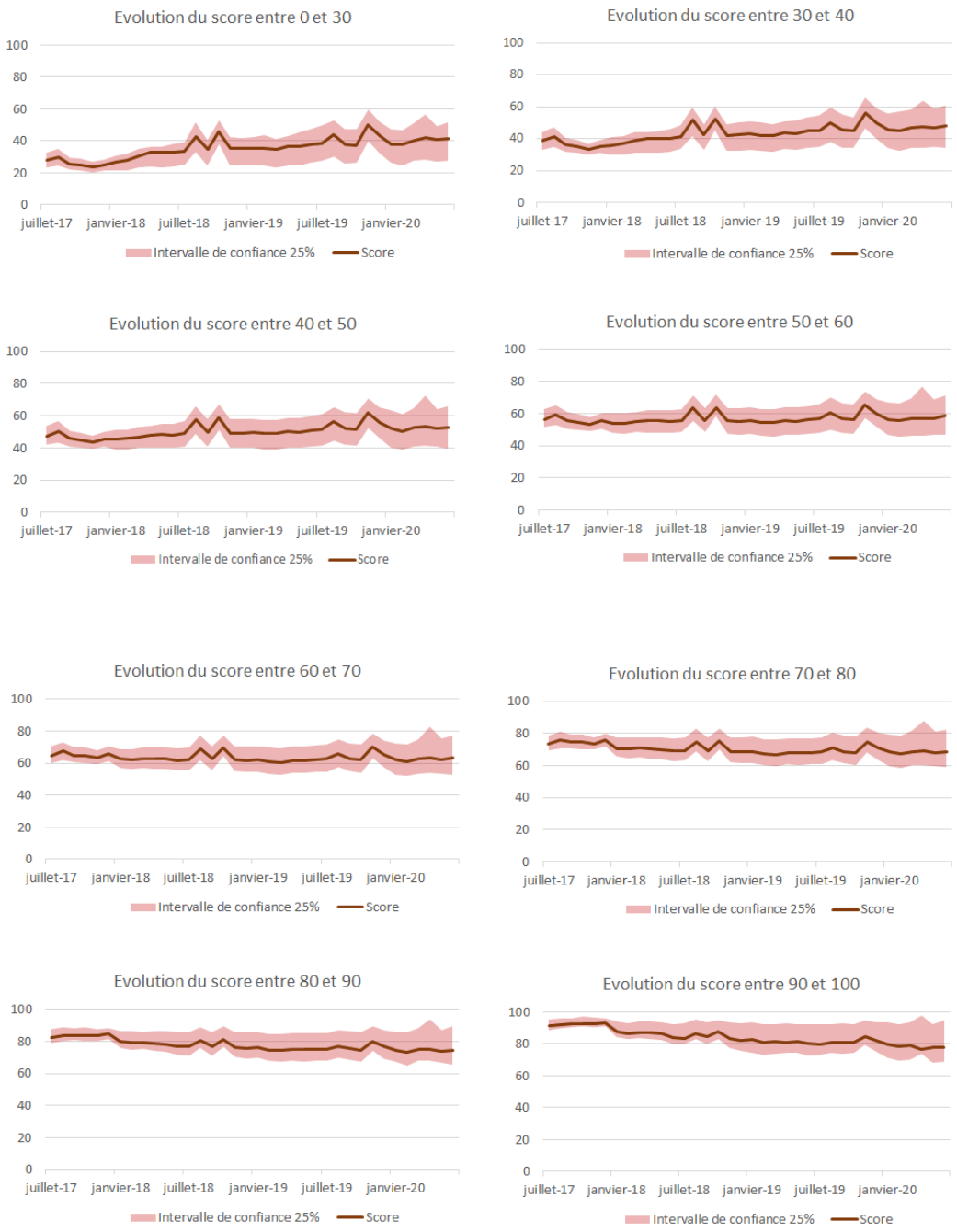

FIGURE 3.9 – Evolution des scores de conduite au cours du temps

Les scores moyens sont stables dans le temps. Un véhicule ayant un score entre 50 et 60 en 2017 aura tendance à rester dans le même intervalle de score en 2018, 2019 et 2020.

Ces observations nous permettent de formuler l'hypothèse selon laquelle les scores obtenus sur les premiers mois sont représentatifs du score de conduite de l'année.

Afin de confirmer cette hypothèse, nous représentons la convergence du score de conduite. Nous calculons le score de conduite moyen sur la période, pondéré par la distance effectuée chaque jour. Puis nous déterminons l'écart moyen entre le score de conduite observé chaque jour et le score de conduite moyen observé sur toute la période.

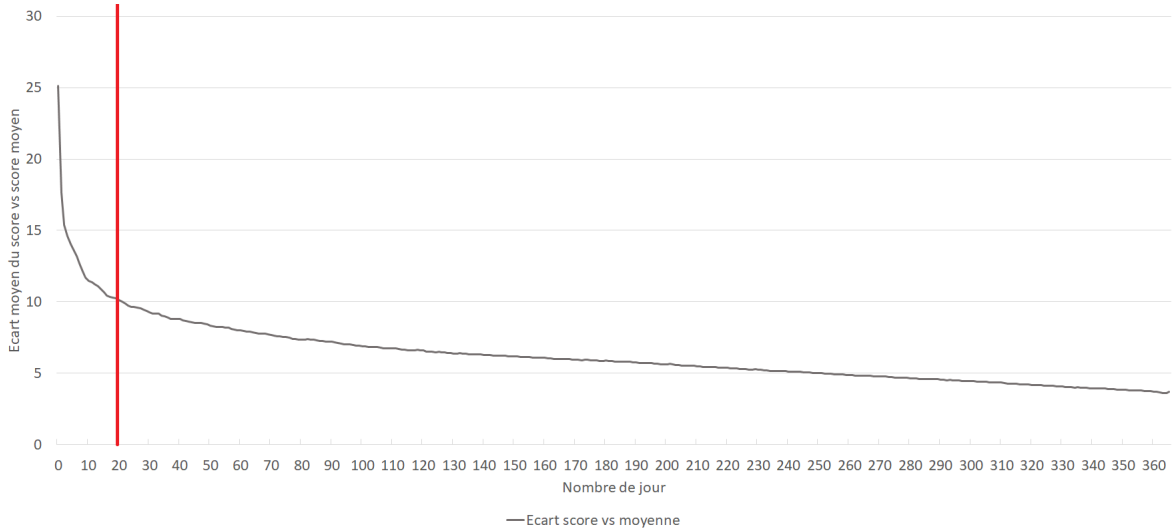

FIGURE 3.10 – Convergence des scores de conduite au cours du temps

Le score de conduite observé après 20 jours est identique à celui observé sur toute la période à +/- 10. Après un an, le score de conduite observé est identique sur toute la période à +/- 3,59. Le score de conduite présente une importante rapidité de convergence. En seulement 20 jours, ce paramètre permet de connaitre le profil approximatif du conducteur.

Deuxième partie

Modélisation GLM

La plupart des modèles de tarification en assurance automobile se basent sur des GLM. Afin d'évaluer l'intérêt des données télématiques dans la modèlisation, nous allons construire un modèle de sinistralité avec les composantes classiques auquel nous allons ajouter l'information télématique par l'intermédiaire du score de conduite décrit précédemment.

Nous présenterons tout d'abord la théorie des GLM. Nous modéliserons ensuite la sinistralité matérielle d'une part et la sinistralité corporelle d'autre part.

Chapitre 4

Présentation théorique des modèles linéaires généralisés

Cette partie introduit les fondements théoriques utiles à la modélisation des impacts de variables explicatives sur la sinistralité. Elle permet notamment de comprendre comment nous utiliserons cette théorie.

Dans un premier temps, nous décrirons les principes de la tarification. Nous rappellerons les principaux éléments ayant une influence sur la tarification. Nous étudierons ensuite la théorie de fréquence-coût de sinistres. Nous décrirons enfin la segmentation tarifaire et détaillerons, dans l'ordre, les cinq étapes de la mise en place d'un modèle de tarification : principes de régression, cas général des GLM, construction d'un modèle, sélection des variables du modèle et mesure de la qualité du modèle.

4.1 Processus de tarification d'un produit d'assurance automobile

La tarification consiste à donner un prix à un produit d'assurance. Pour déterminer le prix d'un produit, l'équipe actuariat va d'abord devoir calculer le montant de la sinistralité probable. Pour établir ce montant, l'équipe a recours aux données historiques du portefeuille concerné ou d'un portefeuille similaire.

Les éventuels besoins en réassurance du produit seront ensuite évalués. Ce travail se fait en collaboration avec les équipes contractuelles qui évaluent précisément les risques liés aux différentes garanties et limites.

La tarification doit bien évidemment tenir compte de la probabilité que le sinistre survienne et du coût moyen du sinistre s'il survient. Mais elle doit intégrer également de nombreux autres paramètres relatifs aux contraintes de fonctionnement de la société d'assurance. Au final, le tarif d'un produit se décompose ainsi :

- **La prime pure** : elle correspond au montant estimé de la charge des sinistres.
- **Les chargements de sécurité** : ils correspondent à une provision financière supplémentaire en cas de sur-sinistralité les premières années du contrat ou d'aggravation inattendue de la sinistralité pour les contrats plus anciens.
- **La gestion de sinistres** : les compagnies d'assurance délèguent la gestion logistique des sinistres à des entreprises spécialisées, ce qui induit un coût supplémentaire pour chaque sinistre.
- **Les chargements pour frais** : ils correspondent aux frais de structure et de fonctionnement de la compagnie d'assurance.
- **La réassurance** : les compagnies d'assurance cèdent, dans certains cas, une partie du risque assuranciel à une autre compagnie.
- **La marge** : elle correspond à la rentabilité du contrat pour l'assureur. Cette partie est celle qui suscite le plus de négociations.
- Les taxes obligatoires sur les produits assuranciels.

Chacun de ces éléments est dépendant des autres à l'exception de la prime pure qui représente la proportion la plus importante de la prime commerciale.

Après le lancement du produit, un suivi est réalisé par l'équipe actuariat pour vérifier la cohérence entre la sinistralité prévue lors de la tarification et la sinistralité réelle. La rentabilité du produit est analysé régulièrement. Si nécessaire, ces différentes observations peuvent conduire à une refonte tarifaire ou à une modification du produit.

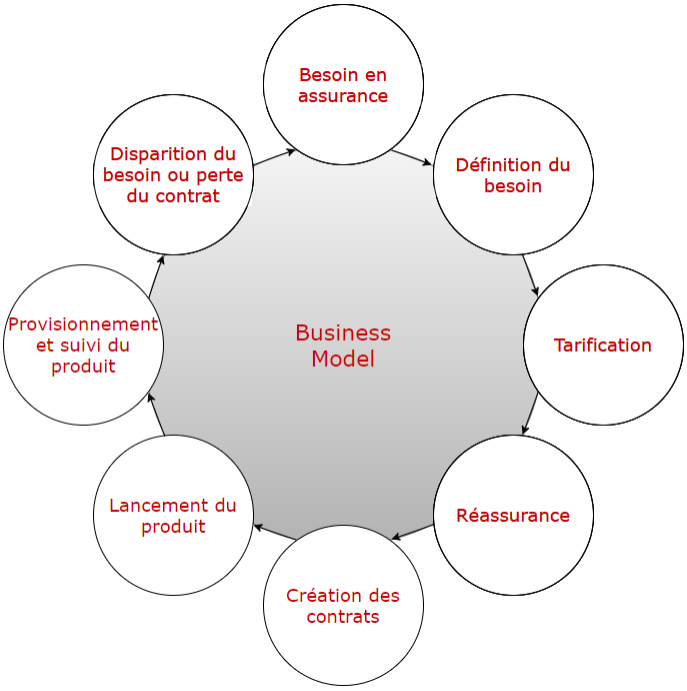

FIGURE 4.1 – Vie d'un produit d'assurance

Dans la suite de ce mémoire, nous choisissons de nous intéresser exclusivement aux modalités de calcul de la prime pure.

4.2 Théorie fréquence-coût

Sur une période donnée, la prime pure correspond à la moyenne des coûts passés et futurs. Elle est donc égale à l'espérance de la charge globale de sinistre.

Dans le cas où il y a plusieurs polices, il est possible de définir la charge globale de sinistre de deux manières différentes : la moyenne pondérée du coût moyen par police ou le coût moyen total. La première écriture correspond au modèle individuel et la seconde au modèle collectif.

Dans cette étude, nous considérons une période donnée unitaire. Pour obtenir le résultat sur plusieurs périodes, il faudra multiplier les résultats par le nombre de périodes.

4.2.1 Modèle individuel

Le modèle individuel fonctionne police par police. Considérons un portefeuille de n polices avec S^i la charge totale des sinistres de la police i.

La charge totale modélisée du portefeuille s'écrit :

$$S^{ind} = \sum_{i=1}^{n} S^i \tag{4.1}$$

Notons PP la prime pure :

$$PP = \mathbb{E}[S^{ind}] = \sum_{i=1}^{n} \mathbb{E}[S^i]$$

Cependant, chaque S_i a souvent sa fonction de densité propre, ce qui rend l'obtention de la fonction de répartition de S^{ind} compliquée. C'est la raison pour laquelle cette modélisation n'est que très rarement utilisée sauf lorsqu'il y a peu de polices dans le modèle.

4.2.2 Modèle collectif

Le modèle collectif travaille sur le portefeuille dans son ensemble. Il considère, en effet, la charge sinistre totale en fonction du montant de chaque sinistre quelle que soit sa police.
On considère N le nombre de sinistres et la suite $M = (M_i)_{i\geq 0}$ des variables aléatoires réelles représentant les montants individuels de sinistre. La charge totale des sinistres du portefeuille est :

$$S^{collectif} = \sum_{i=1}^{N} M_i \tag{4.2}$$

Si on suppose que les M_i sont de même loi, indépendants entre eux et vis à vis de la variable N :

$$\begin{aligned} PP &= \mathbb{E}[S^{collectif}] \\ &= \mathbb{E}[\sum_{i=1}^{N} M_i] \\ &= \mathbb{E}[N] \times \mathbb{E}[M] \end{aligned}$$

On peut alors scinder la prime pure en deux : une partie sur le nombre de sinistres et une partie sur le coût des sinistres.

Il est plus facile de modéliser les deux composantes fréquence et coût moyen séparément. Nous les étudierons une à une dans un premier temps avant de les réunir pour obtenir notre prime pure.

Cette modélisation est efficace lorsque le portefeuille est de taille importante. L'application de l'espérance suppose, en effet, qu'il y ait suffisamment d'observations pour avoir une estimation valable du risque réel. Dans notre cas, le modèle collectif est choisi car il concerne une flotte de véhicules donc un échantillon suffisamment important.

4.3 Présentation de la segmentation tarifaire

Dans le modèle collectif, la prime pure se calcule en multipliant la fréquence de survenance d'un sinistre par le coût moyen des sinistres. Si l'on applique cette logique sur l'ensemble de la population, le prix de l'assurance sera identique quel que soit le modèle du véhicule par exemple. Or le prix des réparations est différent selon la catégorie du véhicule accidenté.

Afin de pallier ce biais, nous mettons en place une segmentation de façon à regrouper dans une même classe les individus ayant un même niveau de risque.

FIGURE 4.2 – Principe de la segmentation

Cette pratique permet d'optimiser la tarification. Cette segmentation permet également à l'assureur de sélectionner préférentiellement les bons risques. Prenons l'exemple d'un marché avec deux assureurs. On le représente par le schéma suivant.

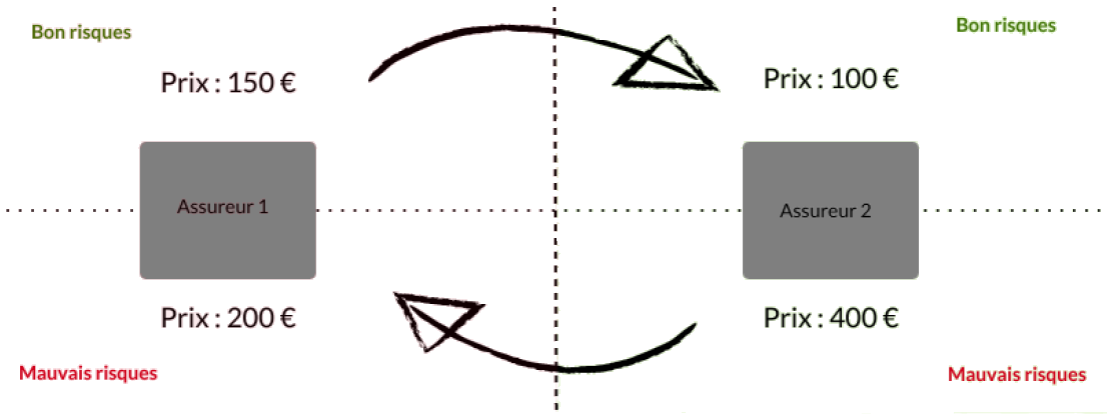

FIGURE 4.3 – Utilité de la segmentation

Si l'on considère que, pour deux produits identiques, le choix des assurés se porte sur le moins onéreux, on constate que l'assureur 1 aura préférentiellement des mauvais risques. L'assureur 2, quant à lui, sélectionnera majoritairement les bons risques. Sur le long terme, l'assureur 2 pourra baisser davantage ses prix car il n'aura que des bons risques, peu onéreux à couvrir, alors que l'assureur 1 se verra dans l'obligation d'augmenter ses prix pour faire face à l'indemnisation des mauvais risques.

On comprend alors toute l'importance d'une bonne segmentation.

4.4 Principe de la régression linéaire

La méthode principale de la segmentation en assurance non-vie repose sur des modèles linéaires.

Ils sont utilisés afin d'évaluer les effets de variables explicatives sur une variable continue dite de réponse. En termes mathématiques, on peut traduire ce modèle par une égalité matricielle :

$$Y = X\beta + \epsilon \tag{4.3}$$

Avec :

- $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix}$, un vecteur de $\mathbb{R}^n$ tel que pour tout i, Y_i est la ième observation de la variable de réponse.

- $X = \begin{pmatrix} 1 & X_1^1 & \cdots & X_1^p \\ \vdots & & \ddots & \\ 1 & X_n^1 & \cdots & X_n^p \end{pmatrix}$, une matrice de $\mathbb{R}^{n\times(p+1)}$ tel que pour tout i $\in [1;n]$ et pour tout j $\in [1;p]$ X_i^j est la ième observation de la jème variable.

- $\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_p \end{pmatrix}$, un vecteur de $\mathbb{R}^{p+1}$ tel que pour tout j, β_j le jème coefficient de la régression. Le β_j correspond à la relation entre la jème variable et la variable réponse.

- $\epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_j \\ \vdots \\ \epsilon_n \end{pmatrix}$, un vecteur de $\mathbb{R}^n$ tel que pour tout j, ϵ_j l'estimation de l'erreur à la jème variable. On parle alors de résidus.

Le but de la modélisation linéaire est de trouver les β qui permettent de minimiser les résidus. Quand des β satisfaisants sont déterminés, il est possible de prendre de nouvelles observations, correspondant à $(1, x_{n+k}^1, \cdots, x_{n+k}^p)$ avec k un entier naturel, et de lui appliquer le modèle afin de trouver un Y_{n+k} estimé.

Cette modélisation est soumise à trois hypothèses :

- Les résidus sont indépendants.
- Les résidus suivent une loi Normale de moyenne nulle et de variance résiduelle.
- Les résidus sont homogènes.

Ces hypothèses sont simples mais contraignantes. Dans la modélisation de la sinistralité automobile, le nombre de sinistres est une variable de comptage qui appartient à $\mathbb{N}$ et suit une distribution de Poisson. La variance du nombre de sinistres est le paramètre de la loi de Poisson. Ce paramètre augmente donc avec le nombre de sinistres, ce qui implique une augmentation de la variance. Ceci permet de démontrer que l'hypothèse d'homogénéité des résidus n'est pas concevable dans notre cas.

De plus, le fait d'avoir des résidus qui suivent une loi Normale implique la possibilité d'avoir

des prédictions négatives alors qu'il est impossible d'avoir un nombre de sinistres négatif.

Les hypothèses du modèle linéaire empêchent de modéliser le nombre de sinistres tel qu'on le souhaiterait, c'est pourquoi nous utiliserons les modèles linéaires généralisés.

4.5 Modèles linéaires généralisés (GLM)

Le principe général des GLM est le même que celui des modèles linéaires : trouver des coefficients tels que $Y = \beta X + \theta$. Cependant, des hypothèses vont être formulées sur la loi suivie par la variable réponse. Le GLM généralise la régression linéaire en permettant au modèle linéaire d'être relié à la variable réponse via une fonction lien. Le but est de "normaliser" les $\beta X + \theta$ afin de respecter les hypothèses des modèles linéaires.

Le modèle GLM s'écrit :

$$g(\mathbb{E}[Y]) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

avec g la fonction lien et p un entier représentant le nombre de variables explicatives.

Afin d'obtenir les vraies prédictions, il est nécessaire d'appliquer l'inverse de la fonction lien

$$\mathbb{E}[Y] = g^{-1}(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p) \tag{4.4}$$

Un modèle GLM se met en place en 4 étapes :

- Le choix d'une fonction lien à partir de la détermination de la loi de distribution de la variable de réponse.
- Le calcul des coefficients β.
- La sélection des variables.
- La vérification du modèle.

4.5.1 Détermination d'une fonction lien

Le choix de la fonction lien s'initie par l'identification de la loi de distribution de la variable de réponse Y. Pour rappel, cette variable est celle à expliquer dans le modèle.

Détermination de la distribution de la variable de réponse

Nous ne travaillerons qu'avec des lois de densité issues de la famille exponentielle. Les fonctions de densité de cette famille sont de la forme :

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

avec a une fonction continue définie sur $\mathbb{R}$, b une fonction définie sur $\mathbb{R}$ deux fois dérivable et c une fonction définie sur $\mathbb{R}^2$. De plus, on a :

$$\mathbb{E}(Y) = b'(\theta) \quad \mathbb{V}(Y) = b''(\theta)a(\varphi) \tag{4.5}$$

Grâce à ces deux équations, il est possible de comparer avec chaque loi de la famille exponentielle celle qui se rapproche le plus de la variable réponse. Dans le tableau ci-dessous sont indiqués les paramètres des différentes distributions de la famille exponentielle.

Distribution de Y_i	θ_i	φ	$a(\varphi)$	$b(\theta_i)$	$c(y_i, \varphi)$
Normale $(\mu_i; \sigma^2)$	μ_i	σ^2	σ^2	$\frac{\theta^2}{2}$	$-\frac{1}{2}\left\{\frac{y_i^2}{\sigma^2} + \log\left(2\pi\sigma^2\right)\right\}$
Poisson (μ_i)	$\log(\mu_i)$	1	1	$\exp(\theta_i)$	$-\log(y_i)$
Binomiale $\frac{1}{m_i}(m_i; \mu_i)$	$\log\left(\frac{\mu_i}{1-\mu_i}\right)$	$\frac{1}{\mu_i}$	$\frac{1}{\mu_i}$	$\log(1 + \exp(\theta_i))$	$\log\begin{pmatrix} m_i \\ m_i y_i \end{pmatrix}$
Gamma $(\mu_i; \alpha)$	$\frac{-1}{\mu_i}$	α^{-1}	α^{-1}	$-\log(-\theta_i)$	$\alpha\log(\alpha y_i) - \log y_i - \log\gamma(\alpha)$

FIGURE 4.4 – Coefficients des différentes distributions de la famille exponentielle

Le but est d'identifier empiriquement les paramètres de la loi Y et de se rapprocher autant que possible d'un des résultats du tableau afin de déterminer la distribution de Y.

Souvent, plusieurs distributions sont viables. Dans ce cas, il est possible d'utiliser un test afin de savoir quelle sera la meilleure loi. Parmi les tests envisageables, deux tests se démarquent par leur efficacité : le Critère d'Information d'AKAIKE (AIC) et le Critère d'Information Bayésien (BIC).

Ces tests s'appuient sur la fonction de vraisemblance qui est définie ainsi :

$$L(\theta|x) = \begin{cases} P_\theta(X = x) = \prod_{i=1}^n P_\theta(X = x_i) & \text{si X est une variable aléatoire discrète} \\ f_\theta(x) = \prod_{i=1}^n f_\theta(x_i) & \text{si X est une variable aléatoire absolument continue} \end{cases}$$

Par souci de simplification, nous noterons L le maximum de la fonction de vraisemblance.

Avec p le nombre de variables explicatives du modèle, le Critère d'Information d'AKAIKE (AIC) se définit ainsi :

$$AIC = 2p - 2\ln(L)$$

Cet estimateur évalue la quantité d'informations perdues par le modèle. Celui qui perd le moins d'informations étant le meilleur, on cherche à avoir le plus petit AIC possible.

Cet estimateur est proportionnel au nombre de variables. Cette pénalisation des modèles avec beaucoup de variables permet, en effet, de limiter le sur-ajustement qui consiste à s'adapter aux valeurs aberrantes ou anormales.

Avec n le nombre d'observations et p le nombre de variables explicatives, le Critère d'Information Bayésien (BIC) se définit ainsi :

$$BIC = \ln(n)p - 2\ln(L)$$

Le BIC est un dérivé de l'AIC. Cet estimateur pénalise plus sévèrement le nombre de paramètres que l'AIC. Comme l'AIC, le meilleur modèle est celui qui a le plus petit BIC.

Détermination de la fonction lien

Dans le cas particulier de la famille exponentielle, chaque loi de distribution représentant Y a une fonction lien qui lui est associée. Le tableau ci-dessous regroupe les distributions les plus utilisées en assurance non-vie avec leurs lois liens associées :

Distribution de Y	Loi lien
Normale	Identité
Poisson	Logarithmique
Binomiale	Logit
Gamma	Inverse

FIGURE 4.5 – Loi lien en fonction de la distribution

En pratique, nous utilisons principalement la loi de Poisson et la loi Binomiale pour les études de fréquence. Ce sont, en effet, deux lois de comptage. Pour l'étude du coût moyen, nous utilisons la loi Gamma ou la loi Normale.

Sachant que $\mathbb{E}[Y] = g^{-1}(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)$, on peut alors trouver des spécificités en fonction de chaque loi lien :

- **Loi lien identité :** $g(x) = x <=> g^{-1}(x) = x$

$$\mathbb{E}[Y] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Dans ce cas on parle de modèle additif.

- **Loi lien logarithme népérien :** $g(x) = ln(x) <=> g^{-1}(x) = e^x$

$$\mathbb{E}[Y] = exp(\beta_0) \times exp(\beta_1 X_1) \times \cdots \times exp(\beta_p X_p)$$

On peut en déduire que la loi lien logarithme népérien implique une modélisation dans laquelle les estimations de chaque variable se multiplient. On parle de modèle multiplicatif.

- **Loi lien logit :** $g(x) = ln(\frac{x}{1-x}) <=> g^{-1}(x) = \frac{e^x}{1+e^x}$

$$\mathbb{E}[Y] = \frac{exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}{1 + exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}$$

- **Loi lien inverse :** $g(x) = 1/x <=> g^{-1}(x) = 1/x$

$$\mathbb{E}[Y] = \frac{1}{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}$$

En pratique, on sélectionne la loi lien en fonction de la distribution de Y et de la nature des données. Si les données suivent une loi Gamma, mais sont uniquement positives, on utilisera une loi logarithme népérien afin d'avoir des prédictions positives. Cette loi est souvent préférée car elle permet une tarification multiplicative, les effets se multipliant mais ne se compensant pas.

Le modèle multiplicatif amplifie cependant le risque d'une répétition d'informations ou d'erreurs d'informations : s'il y a une corrélation entre deux variables du modèle, l'information conjointe va être comptée deux fois par le modèle. De plus, si une catégorie est très peu nombreuse, la modélisation peut être fausse. Si, par exemple, la population de voitures de luxe dans le portefeuille est très faible, le moindre sinistre d'un de ces véhicules aura un impact important non représentatif de la rareté de l'évènement.

4.5.2 Calcul des coefficients β

Cette étape vise l'estimation des coefficients β de la régression. Ces coefficients vont être calculés grâce à la maximisation de la vraisemblance. Dans notre cas, la vraisemblance peut s'écrire :

$$L(\theta_i, \varphi | y_i) = \prod_{i=1}^{n} f_{\theta_i,\varphi} = \prod_{i=1}^{n} \exp\left(\frac{y_i \times \theta_i - b(\theta_i)}{a(\varphi)} + c(y_i, \varphi)\right)$$

Avec :

- i : l'indice du numéro d'observations
- $(\theta_i)_{i\geq 0}$: la variable aléatoire du premier paramètre de la loi de la famille exponentielle
- φ : la variable aléatoire du deuxième paramètre de la loi de la famille exponentielle
- y_i : la i-ème observation de la variable réponse
- n : nombre d'observations

La log-vraisemblance vaut alors :

$$\mathcal{L}_n(\theta|x) = \sum_{i=1}^{n} \left(\frac{y_i \times \theta_i - b(\theta_i)}{a(\varphi)} + c(y_i, \varphi) \right)$$

Nous cherchons à obtenir le maximum de vraisemblance, donc nous cherchons le maximum de la log vraisemblance. Pour ce faire, nous utilisons le Gradient. Il vaut 0 aux points extrêmaux, ce qui nous permet d'obtenir le maximum s'il existe. Il est défini par :

$$\nabla_\beta \mathcal{L}_n(\theta_i, \varphi|y_i) = \left[\frac{\partial \mathcal{L}_n}{\partial \beta_0}(\theta_i, \varphi|y_i), \ldots, \frac{\partial \mathcal{L}_n}{\partial \beta_p}(\theta_i, \varphi|y_i) \right]'$$

avec $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$ et x' désigne la transposée de x.

Soit j ∈ [0,p] avec p le nombre de variables, cherchons si possible β_j tels que $\frac{\partial \mathcal{L}_n}{\partial \beta_j}(\theta_i, \varphi|y_i) = 0$. Or :

$$\begin{aligned}
\frac{\partial \mathcal{L}_n}{\partial \beta_j}(\theta_i, \varphi|y_i) &= \sum_{i=1}^{n} \frac{\partial}{\partial \beta_j} \left(\frac{y_i \times \theta_i - b(\theta_i)}{a(\varphi)} + c(y_i, \varphi) \right) \\
&= \sum_{i=1}^{n} \frac{\partial \theta_i}{\partial \beta_j} \frac{\partial}{\partial \theta_i} \left(\frac{y_i \times \theta_i - b(\theta_i)}{a(\varphi)} + c(y_i, \varphi) \right) \\
&= \sum_{i=1}^{n} \frac{\partial \theta_i}{\partial \beta_j} \left(\frac{y_i - b'(\theta_i)}{a(\varphi)} \right)
\end{aligned}$$

En utilisant les formules (4.4) et (4.5), on peut obtenir

$$\begin{aligned}
\frac{\partial \theta_i}{\partial \beta_j} &= \frac{\partial \mathbb{E}[Y_i]}{\partial \beta_j} \frac{\partial \theta_i}{\partial \mathbb{E}[Y_i]} \\
&= \frac{\partial (g^{-1}(\beta_0 + \beta_1 x_i^1 + \cdots + \beta_p x_i^p))}{\partial \beta_j} \frac{1}{\frac{\partial b'(\theta_i)}{\partial \theta_i}} \\
&= \frac{\partial (g^{-1}(\beta_0 + \beta_1 x_i^1 + \cdots + \beta_p x_i^p))}{\partial \beta_j} \frac{1}{b''(\theta_i)} \\
&= \frac{\partial (g^{-1}(\beta_0 + \beta_1 x_i^1 + \cdots + \beta_p x_i^p))}{\partial \beta_j} \frac{a(\varphi)}{\mathbb{V}(Y_i)}
\end{aligned}$$

Il faut donc résoudre

$$\frac{\partial \mathcal{L}_n}{\partial \beta_j}(\beta) = \sum_{i=1}^{n} \frac{\partial (g^{-1}(\beta_0 + \beta_1 x_i^1 + \cdots + \beta_p x_i^p))}{\partial \beta_j} \frac{(y_i - b'(\theta_i))}{\mathbb{V}(Y_i)} = 0 \tag{4.6}$$

Pour simplifier le calcul, cette équation est résolue par des méthodes itératives implémentées sur le logiciel R ou SAS. La résolution de ce problème suppose le nombre d'observations n supérieur au nombre de variables p.

4.5.3 Sélection des variables

Une fois les premières modélisations réalisées, l'objectif est de fiabiliser au mieux le modèle et de l'améliorer autant que possible en sélectionnant les variables adéquates.

L'optimisation du modèle se déroule en deux étapes :

- la création de modalités pour chaque variable afin de rendre le modèle le plus efficace possible
- la sélection des variables afin de conserver uniquement les variables ajoutant de l'information.

Les variables utilisées peuvent être objectivées de multiples façons : la variable localisation, par exemple, peut être une ville précise, un département ou une région.

Si la variable est trop précise, elle ne revêt que très peu d'informations : si, par exemple, un seul contrat est associé à la ville de Lyon et qu'un sinistre est recensé sur ce contrat, alors la ville de Lyon apparaîtra à tort. Ce biais représente un risque pour le modèle. Au contraire, une segmentation trop large ne permettra pas d'identifier efficacement les différentes zones de risque.

L'optimisation des variables est donc à la fois primordiale et délicate. Par ailleurs, ces variables ne doivent être ni redondantes, ni vectrices d'erreurs tout en conservant le plus d'informations possible.

Afin de détecter les variables les plus pertinentes dans le modèle, nous analysons la p-value associée à chaque modalité. Une variable non-significative (avec une p-value $> 5\%$) sur toutes ces modalités est une variable qui n'apporte aucune information.

Il est aussi possible d'apprécier la qualité des variables en observant la statistique du Wald Chi Square pour chaque modalité. Il se définit de la manière suivante :

$$\text{Wald Chi-Square} = \frac{\text{valeur estimée}(\beta)}{\text{Erreur type}^2}$$

Cet indicateur donne une mesure de la quantité d'informations apportée par chaque modalité dans le modèle. Plus le Wald Chi Square est élevé plus la modalité est estimée avec précision.

Afin d'identifier la meilleure combinaison de variables, il existe des processus de sélection de variables qui créent des modèles GLM de façon répétées :

- **GLM Forward** : cette méthode, connue sous le nom de **'procédure ascendante'**, construit une première modélisation avec une unique variable explicative. La variable choisie par le modèle est celle qui implique le plus petit AIC. En d'autres termes, on construit un premier modèle avec la variable qui apporte le plus d'informations. Dans un deuxième temps, on construit un deuxième modèle avec la première variable et la seconde variable qui fait baisser le plus l'AIC du modèle. On répète cette action jusqu'à ce qu'il ne reste plus que des variables qui font augmenter l'AIC. Ces variables ne seront pas retenues car elles dégradent le modèle.
- **GLM Backward** : cette méthode est connue sous le nom de **'procédure descendante'**. Elle est considérée comme étant la procédure opposée à la précédente. Elle utilise un premier GLM avec toutes les variables, puis supprime la variable créant le plus d'erreurs. On voit ainsi l'AIC diminuer. La procédure réitère cette boucle tant qu'il existe une variable génératrice d'erreurs, c'est-à-dire une variable qui fait augmenter l'AIC quand elle est présente. A la fin de cette procédure, on se retrouve avec un modèle rendu aussi précis que possible.

En pratique, nous construirons notre modèle avec un maximum de variables explicatives différentes, nous supprimerons ensuite les variables non significatives puis nous appliquerons un processus backward.

A l'issue de cette étape, nous obtenons, grâce aux variables sélectionnées, un modèle optimal dont il convient de mesurer la qualité. Plusieurs techniques existent pour l'estimer.

4.5.4 Vérification du modèle

Nous retiendrons quatre méthodes de mesure pour estimer la qualité de notre modèle :

- MSE (Mean Squared Error ou erreur quadratique moyenne) est une mesure souvent utilisée pour sa simplicité et son efficacité. Elle peut, en effet, s'appliquer à tous les modèles prédictifs. Elle est souvent calculée sur une base dite de test :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(observations_i - estimations_i)^2$$

  En pratique, nous cherchons à minimiser cette mesure.

- AIC : cet estimateur permet également de mesurer la qualité d'un modèle.

- BIC : cet estimateur permet également de mesurer la qualité d'un modèle.
- Déviance du modèle : la déviance permet de mesurer l'écart entre les valeurs théoriques du modèle et les valeurs observées. Considérons un modèle parfait, un modèle pour lequel chaque observation prédit la variable réponse, on parle alors de modèle saturé, la déviance s'écrit :

$$D = 2\left(LL_{sature} - LL\right)$$

avec LL_{sature} la log vraisemblance du modèle saturé et LL la log vraisemblance du modèle étudié.

Ces tests permettent de vérifier l'efficacité du modèle en fonction des observations. Cependant, en raison de la nature aléatoire des données, les erreurs dans la modélisation sont inévitables.

Chapitre 5

Modélisation de la sinistralité responsabilité civile matérielle

La modélisation de la sinistralité matérielle utilise des GLM et se décompose en deux parties : la première concerne la fréquence, et la seconde le coût moyen. Nous obtenons donc deux modèles GLM, un pour la fréquence matérielle et un autre pour le coût moyen matériel.

5.1 Modélisation de la fréquence des sinistres matériels

Les modèles GLM expliquant la fréquence de sinistres utilisent des lois de comptage. Nous supposerons que le nombre observé a une distribution de loi de Poisson, et nous utiliserons une loi lien logarithme népérien. Cette modélisation s'effectue en deux étapes : création des modalités, puis mesure de l'ajout d'informations par la variable score de conduite dans la modélisation.

5.1.1 Création des modalités

Certaines variables présentent de nombreuses modalités, ce qui implique une grande volatilité entre elles avec une significativité faible. Il est donc nécessaire d'en regrouper certaines afin de solidifier le modèle et rendre la variable plus significative.

Nous avons testé l'ensemble des variables présentes dans la base contrat mais nous avons retenu les variables suivantes :

- l'année
- l'âge du véhicule en début de couverture
- la province italienne
- le poids du véhicule
- la puissance du véhicule

- la marque du véhicule
- le type de carburant du véhicule
- le nombre de kilomètres au contrat
- le nombre de places dans le véhicule
- la présence en zone aéroportuaire
- le type de véhicule
- le secteur d'activité
- la taille de la flotte

L'année

La variable "année" est une variable qui correspond à l'année de souscription du contrat : elle permet de capter les "bruits" des années, comme l'impact de la Covid 19 par exemple.

Dans notre cas, chaque année démarre au 1^{er} novembre et se termine au 31 octobre de l'année suivante. Ainsi Y6 est la période du premier novembre 2017 au 31 octobre 2018.

Observons l'impact de cette variable :

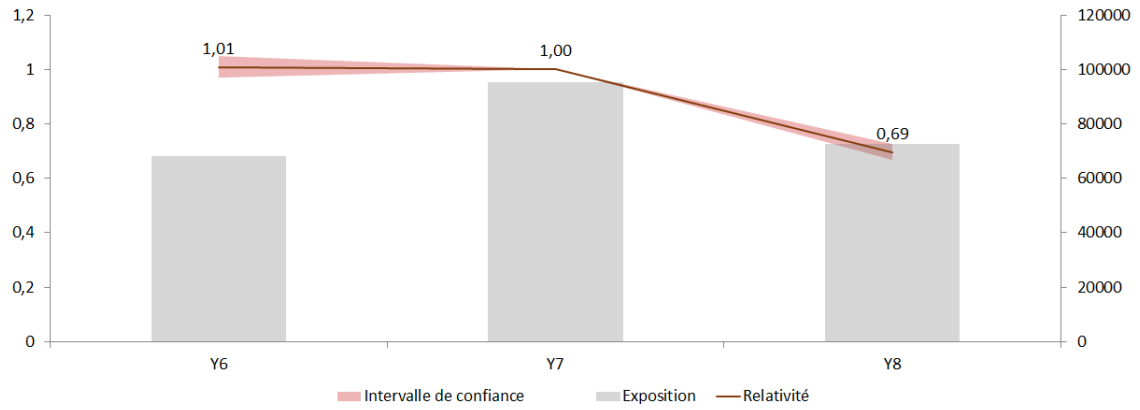

FIGURE 5.1 – Fréquence et relativité par année contrat

On visualise ainsi un impact lié à la Covid 19. En Y6, les véhicules ont 1% plus de probabilité d'avoir un sinistre qu'en Y7. En Y8, les véhicules ont 31% moins de probabilité d'avoir un sinistre qu'en Y7.

L'âge du véhicule en début de couverture

Le nombre d'observations pour chaque âge de véhicule en début de couverture se répartit comme suit :

Age	Nb d'observations
0	48 303
1	80 380
2	58 146
3	25 746
4	5 274
5	395
6	26
7	3
8	2

FIGURE 5.2 – Répartition de l'âge du véhicule en début de couverture

Nous constatons que ce nombre est très faible pour les véhicules de 6,7 et 8 ans en début de couverture.

En appliquant un premier modèle GLM à ces données, on obtient le graphe suivant :

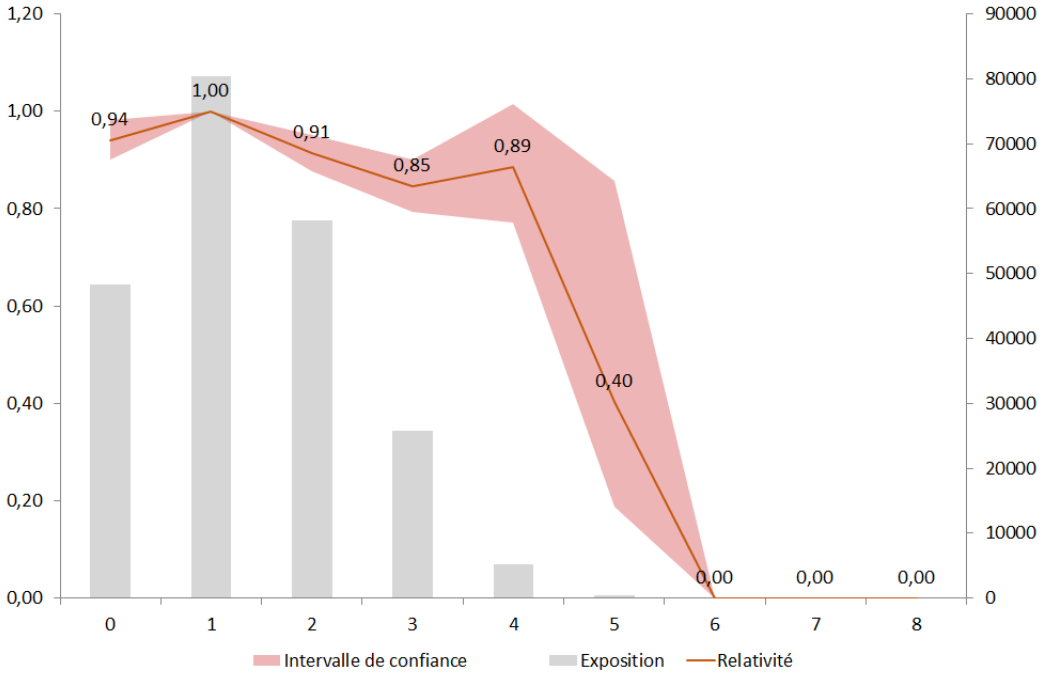

FIGURE 5.3 – Fréquence et relativité par âge de véhicule

La modalité "supérieur à 5" a très peu d'observations et l'intervalle de confiance est assez grand. De plus, lors de la réalisation du GLM d'après la p-value, cette modalité n'est pas significative. De même, pour la p-value du GLM, la modalité "âge de couverture à 0" n'est

pas significative.

Nous regroupons donc les modalités 0, 1 et supérieur à 5 et appliquons une nouvelle modélisation. Nous renouvelons ce travail de regroupement en fonction des résultats obtenus pour conserver finalement 3 modalités : 0-2 ans, 2-3 ans et plus de 3 ans.

Nous obtenons les relativités suivantes :

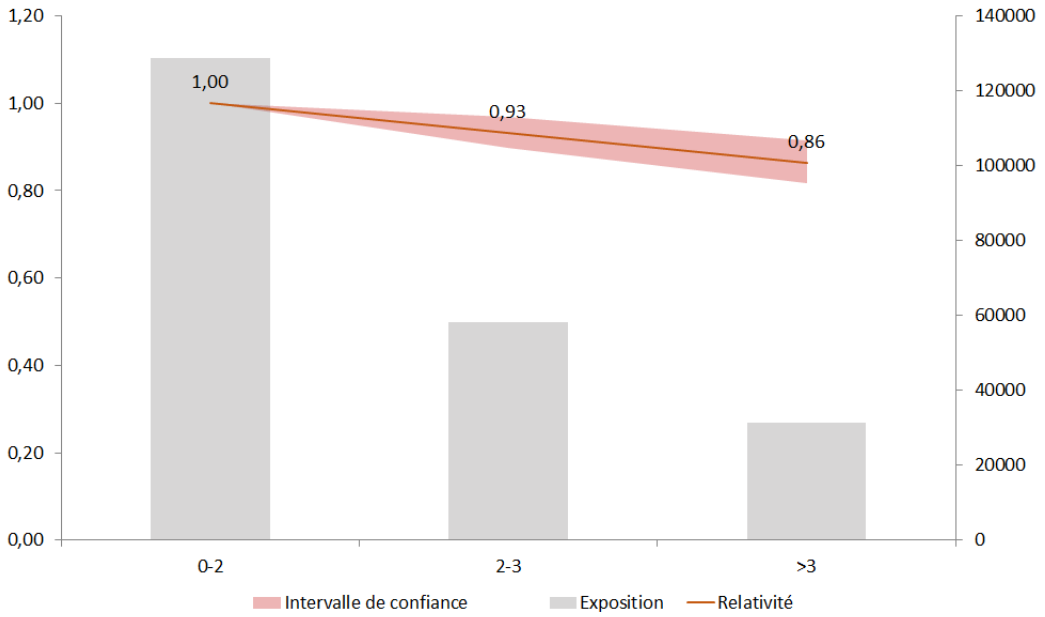

FIGURE 5.4 – Fréquence et relativité par âge de véhicule

La province italienne

Afin de construire les regroupements les plus pertinents, il est nécessaire d'observer la relativité associée à chaque province et le nombre d'observations disponibles pour chacune d'elles.

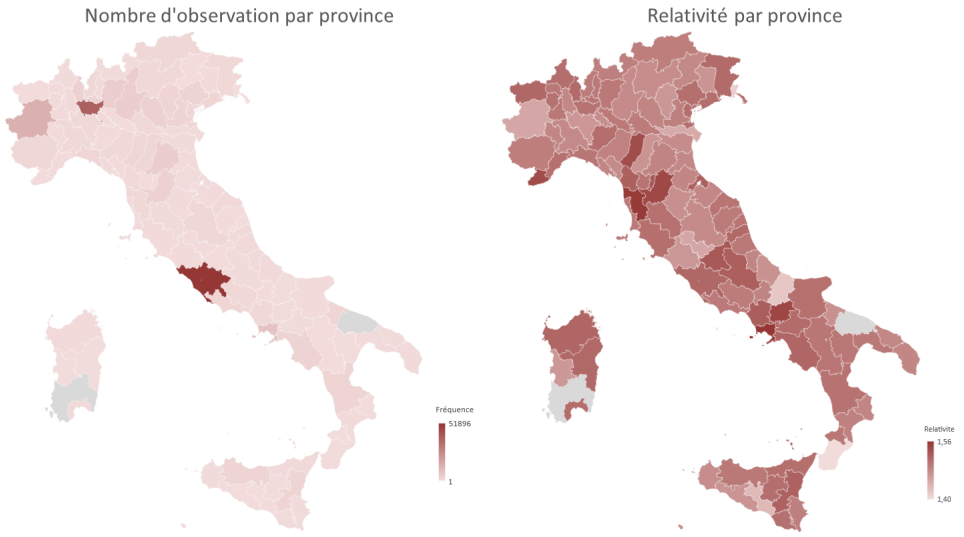

FIGURE 5.5 – Fréquence et relativité par province

Nous regroupons entre elles les provinces qui sont à la fois proches géographiquement, qui ont une relativité comparable ou qui présentent des variables non significatives.

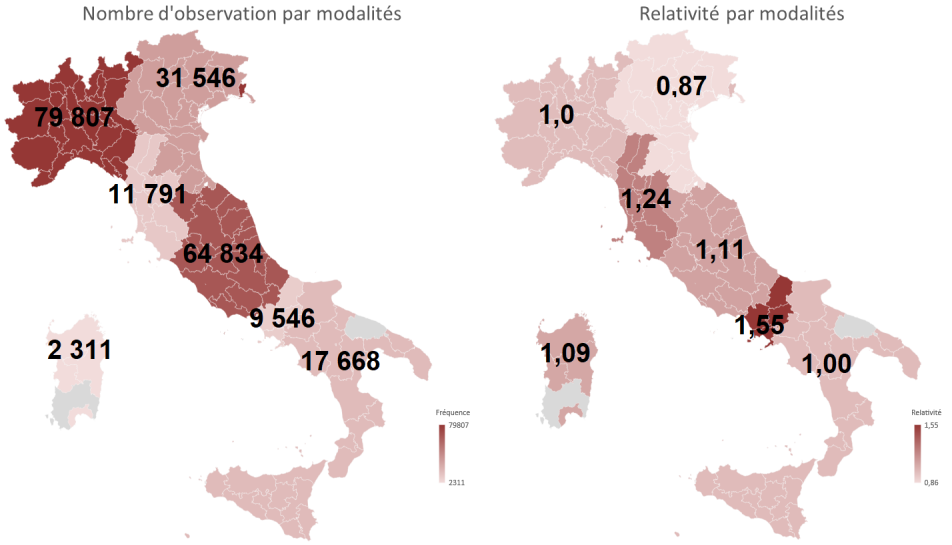

FIGURE 5.6 – Fréquence et relativité par modalités

Nous conserverons dans le modèles ces sept classes.

Le poids et la puissance du véhicule

Dans le domaine de l'assurance automobile, les variables poids et puissance sont généralement corrélées. Afin d'optimiser la modélisation, nous testons trois regroupements différents :

une variable croisée poids-puissance, une variable basée sur le rapport poids-puissance et des variables poids et puissance regroupées.

		Puissance du véhicule (en cheval DIN)							
		0-73	73-115	115-146	146-159	159-174	174-203	203-238	238-650
Poids du véhicule (en kilogramme)	0-1365	1	1	1	1	1	1	1	2
	1365-1566	1	1	1	1	1	1	1	1
	1566-1820	1	1	1	1	1	2	2	2
	1820-1857	2	2	2	2	1	2	1	2
	1857-1966	1	1	1	1	1	1	1	1
	1966-2104	1	2	1	1	1	1	1	1
	2104-2157	2	2	1	1	1	1	1	1
	2157-2244	1	1	1	1	1	1	1	1
	2244-2378	2	2	1	1	1	1	2	2
	2378-2665	1	2	1	1	1	1	1	2
	2665-6000	1	3	3	3	3	3	3	3

FIGURE 5.7 – Lexique de la variable croisée poids-puissance

Pour la variable croisée poids-puissance, nous retiendrons trois catégories, chacune représentée par une couleur dans le graphique ci-dessus.

Pour la variable rapport poids-puissance, nous identifions trois catégories :

- 0 - 18
- 18 - 24,8
- plus de 24,8

Enfin, pour les variables poids et puissance regroupées nous retenons :

— Quatre modalités pour la puissance :
  - 0 - 73 ch
  - 73 - 146 ch
  - 146 - 203 ch
  - plus de 203 ch

— Quatre modalités pour le poids :
  - 0 - 1966 kg
  - 1966 - 2378 kg
  - 2378 - 2665 kg
  - 2665 - 3500 kg
  - plus de 3500 kg

Pour déterminer le meilleur de ces trois regroupements, nous appliquons une modélisation GLM :

Paramètre		Valeur estimée	Wald Chi-Square	Pr > Khi-2	Relativité
Puissance du véhicule	**0-73**	-0,17	39	<10-4	**0,84**
Puissance du véhicule	**73-146**	0,04	1	0,24	**1,04**
Puissance du véhicule	**146-203**	-0,13	6	0,01	**0,88**
Puissance du véhicule	**203-+**	0,00			**1,00**
Poids du véhicule	**0-1966**	0,00			**1,00**
Poids du véhicule	**1966-2378**	0,00	0	0,92	**1,00**
Poids du véhicule	**2378-2665**	0,06	2	0,17	**1,06**
Poids du véhicule	**2665-3500**	0,53	163	<10-4	**1,70**
Poids du véhicule	**3500-+**	1,49	35	<10-4	**4,42**
Rapport poids-puissance	**0-18**	0,00			**1,00**
Rapport poids-puissance	**24,8**	0,01	0	0,54	**1,01**
Rapport poids-puissance	**123,3**	0,03	0	0,51	**1,03**
Variable croisée poids-puissance	2	0,10	9	0,00	**1,11**
Variable croisée poids-puissance	3	0,00			**1,00**
Variable croisée poids-puissance	**1**	0,00			**1,00**
Type de carburant du véhicule	**Electrique et hybride électrique**	-0,14	13	0,00	**0,87**
Type de carburant du véhicule	**GPL et hybride GPL**	0,23	28	<10-4	**1,25**
Type de carburant du véhicule	**Essence et diesel**	0,00			**1,00**
Marque du véhicule	**Francaise**	0,28	111	<10-4	**1,32**
Marque du véhicule	**Luxe 1**	0,23	54	<10-4	**1,26**
Marque du véhicule	**Luxe 2**	0,14	22	<10-4	**1,15**
Marque du véhicule	**Populaires 1**	0,15	44	<10-4	**1,16**
Marque du véhicule	**Populaires 2**	0,00			**1,00**
Province italienne	**BRESCIA**	-0,16	34	<10-4	**0,85**
Province italienne	**CAGLIARI**	0,10	1	0,23	**1,10**
Province italienne	**COSENZA**	0,01	0	0,67	**1,01**
Province italienne	**FIRENZE**	0,13	14	0,00	**1,14**
Province italienne	**NAPOLI**	0,46	177	<10-4	**1,59**
Province italienne	**ROMA**	0,10	23	<10-4	**1,10**
Province italienne	**MILANO**	0,00			**1,00**
Type de véhicule	**Van**	0,25	58	<10-4	**1,28**
Type de véhicule	**Car**	0,00			**1,00**
Scale		1,00			

FIGURE 5.8 – Comparaison des variables poids et puissance

Quand les variables poids et puissance regroupées sont placées en premier, la variable rapport poids-puissance et la variable croisée poids-puissance ne présentent que peu d'intérêt.

En interchangeant la première variable, on constate que les variables croisées poids-puissance et les deux variables séparées poids et puissance apportent la même information, la variable rapport poids-puissance ne présentant alors que peu d'intérêt.

Si on additionne le Wald Chi-square de chacune des trois propositions, il apparaît que la plus pertinente est la variable poids et puissance séparées ; nous retiendrons donc cette dernière.

La marque du véhicule

La construction des modalités de la variable "marque de véhicule" a été réalisée avec des modélisations qui intégraient également les variables poids et puissance afin de minimiser les corrélations entre les marques de véhicule, le poids et la puissance.

Au final, cinq modalités sont retenues pour cette variable :

Francaise	Luxe 1	Luxe 2	Populaires 1	Populaires 2
Peugeot, Renault	Abarth, Audi, Bentley, Bmw, Great Wall, Jaguar, Jeep, Lamborghini, Lexus, Maserati	Infiniti, Isuzu, Iveco, Lancia, Land Rover, Mercedes, Mini Cooper, Smart,Piaggio, Porsche, Tesla	Alfa Romeo, Citroen,Dasia, DS, Ford,Subaru, Volvo	Fiat, Fuso Canter, Honda, Hyundai, Kia, Mazda, Mitsubishi, Nissan, Opel, Renault Trucs, Seat, Skoda, Ssangyong, Suzuki, Toyota, Volkswagen

FIGURE 5.9 – Modalités de la variable "marque de véhicule"

Nous avons découpé les marques en deux : les marques dîtes de "luxe" et les marque plus "populaires". Sur chacun de ces groupes nous avons fait les regroupements qui optimisaient la significativité. Nous avons alors obtenu deux groupes de marques de "luxe" et trois groupes de marques "populaires".

Le type de carburant du véhicule

La variable "type de carburant" a également été construite en intégrant les notions de poids et de puissance du véhicule, elle a été scindée en trois catégories :

- électrique et hybride électrique
- GPL et hybride GPL
- essence et diesel

Avec ces regroupements, nous obtenons les relativités suivantes :

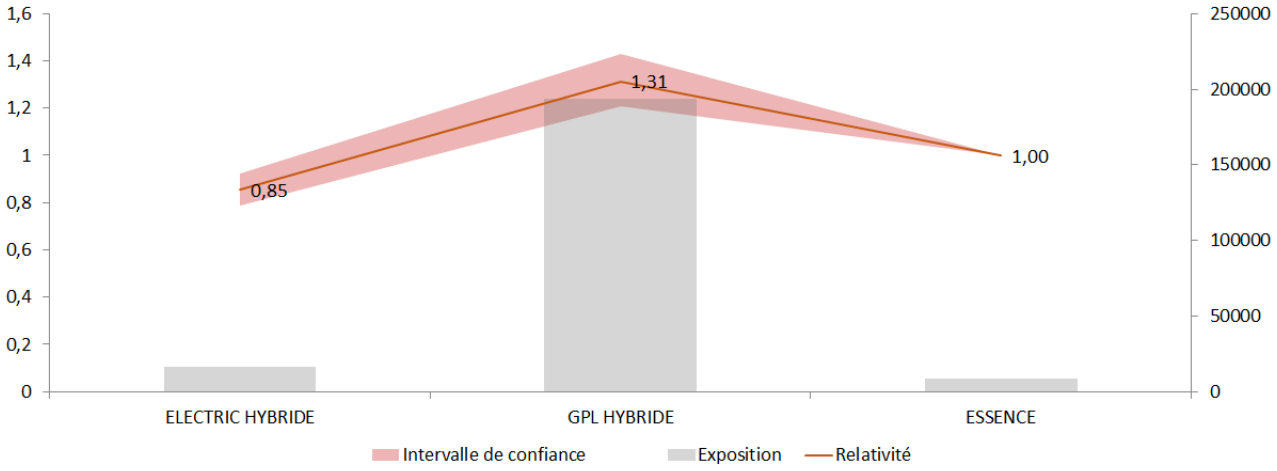

FIGURE 5.10 – Fréquence et relativité par type de carburant

Nous utiliserons ces classes dans le GLM.

Le nombre de kilomètres au contrat

Lors d'une location en leasing automobile, l'entreprise choisi dans son contrat un nombre de kilomètres ainsi qu'une durée de contrat. Grâce à ces deux variables, un nombre de kilomètres contractuel à l'année est déterminé.

Cette variable est regroupée en trois modalités :

- 0 à 19 000 kilomètres par an
- 19 000 à 44 000 kilomètres par an
- plus de 44 000 kilomètres par an

Le nombre de places dans le véhicule

Cette variable a été découpée en quatre modalités :

- 2 places dans le véhicule
- 3 places dans le véhicule
- 4 places dans le véhicule
- plus de 4 places dans le véhicule

La présence en zone aéroportuaire

Cette variable est une variable indicatrice ; elle vaut 1 si le véhicule est présent dans une zone aéroportuaire et 0 dans le cas contraire.

Le type de véhicule

Cette variable contient deux modalités : C si le véhicule est une voiture et V si le véhicule est un van (utilitaire ou voiture van).

Le secteur d'activité

Cette variable représente le secteur d'activité de l'entreprise louant le véhicule. Nous avons catégorisé six secteurs d'activité :

- Les activités de services
- Les activités financières
- Les activités du bâtiment
- Les activités liées à l'immobilier
- Les activités de santé
- Les activités de commerce de gros

La taille de la flotte

En fonction du nombre de véhicules loués par entreprise, nous avons créé une variable fonction de l'exposition par client.
Cette variable a quatre modalités :

- la flotte a entre 0 et 123 d'exposition.
- la flotte a entre 124 et 200 d'exposition.
- la flotte a entre 201 et 400 d'exposition.
- la flotte a entre 401 et 800 d'exposition.

Nous avons ainsi optimisé toutes les variables utilisées.

5.1.2 Mesure de l'ajout du score de conduite

Le but de cette section est d'étudier le modèle GLM avec les variables explicatives présentées précédemment, puis d'y ajouter la variable "score de conduite".

Modèle GLM sans le score de conduite

Dans un premier temps, nous appliquons le modèle GLM fréquence matérielle sans intégrer le score de conduite :

Paramètre		Valeur estimée	Wald Chi-Square	Pr > Khi-2	Relativité
Secteur d'activité	**Activités de services**	-0,16	5	0,02	**0,85**
Secteur d'activité	**Activités financières**	-0,09	15	0,00	**0,92**
Secteur d'activité	**Activités du bâtiment**	-0,13	33	<10-4	**0,88**
Secteur d'activité	**Activités liées à l'immobilier**	0,14	3	0,07	**1,15**
Secteur d'activité	**Activités de santé**	0,26	15	0,00	**1,29**
Secteur d'activité	**Activités de commerce de gros**	0,00			**1,00**
Année	**Y6**	-0,03	2	0,13	**0,97**
Année	**Y7**	0,00			**1,00**
Année	**Y8**	-0,37	321	<10-4	**0,69**
Taille de la flotte	**0-123 d'exposition**	0,00			**1,00**
Taille de la flotte	**124-200 d'exposition**	0,02	0	0,73	**1,02**
Taille de la flotte	**201-400 d'exposition**	0,19	27	<10-4	**1,21**
Taille de la flotte	**>400 d'exposition**	-0,09	9	0,00	**0,91**
Nombre de place dans le véhicule	**0-3 places**	0,51	113	<10-4	**1,67**
Nombre de place dans le véhicule	**4 places**	0,05	2	0,16	**1,06**
Nombre de place dans le véhicule	**>5 places**	0,00			**1,00**
Age du véhicule en début de couverture	**0-2 ans**	0,00			**1,00**
Age du véhicule en début de couverture	**2-3 ans**	-0,04	4	0,05	**0,96**
Age du véhicule en début de couverture	**>3 ans**	-0,04	2	0,20	**0,97**
Nombre de kilomètres au contrat	**0-19000 kilomètres par an**	-0,20	94	<10-4	**0,82**
Nombre de kilomètres au contrat	**19000-44000 kilomètres par an**	0,00			**1,00**
Nombre de kilomètres au contrat	**>44000 kilomètres par an**	0,23	80	<10-4	**1,26**
Puissance du véhicule	**0-73**	-0,14	11	0,00	**0,87**
Puissance du véhicule	**73-146**	0,00			**1,00**
Puissance du véhicule	**146-203**	0,00	0	0,90	**1,00**
Puissance du véhicule	**203-+**	-0,08	3	0,07	**0,92**
Poids du véhicule	**0-1966**	0,00			**1,00**
Poids du véhicule	**1966-2378**	-0,02	1	0,40	**0,98**
Poids du véhicule	**2378-2665**	0,03	1	0,45	**1,03**
Poids du véhicule	**2665-3500**	0,41	89	<10-4	**1,50**
Poids du véhicule	**3500-+**	1,14	25	<10-4	**3,12**
Type de carburant du véhicule	**Electrique et hybride électrique**	-0,14	14	0,00	**0,87**
Type de carburant du véhicule	**GPL et hybride GPL**	0,19	21	<10-4	**1,21**
Type de carburant du véhicule	**Essence et diesel**	0,00			**1,00**
Marque du véhicule	**Francaise**	0,18	34	<10-4	**1,19**
Marque du véhicule	**Luxe 1**	0,28	103	<10-4	**1,32**
Marque du véhicule	**Luxe 2**	0,14	23	<10-4	**1,15**
Marque du véhicule	**Populaires 1**	0,10	19	<10-4	**1,10**
Marque du véhicule	**Populaires 2**	0,00			**1,00**
Province italienne	**BRESCIA**	-0,16	38	<10-4	**0,85**
Province italienne	**CAGLIARI**	0,05	0	0,51	**1,06**
Province italienne	**COSENZA**	0,05	2	0,13	**1,05**
Province italienne	**FIRENZE**	0,13	15	0,00	**1,14**
Province italienne	**NAPOLI**	0,51	211	<10-4	**1,66**
Province italienne	**ROMA**	0,15	53	<10-4	**1,16**
Province italienne	**MILANO**	0,00			**1,00**
Type de véhicule	**Van**	-0,05	1	0,26	**0,95**
Type de véhicule	**Car**	0,00			**1,00**
Scale		1,00			

FIGURE 5.11 – Modèle GLM fréquence matérielle sans le score de conduite

Déviance	82 362
Log Vraissemblance	-55 332
AIC	112 537
BIC	112 918

FIGURE 5.12 – Mesures du modèle GLM fréquence matérielle avec le score de conduite

La variable "type de véhicule" n'est pas significative, mais le même modèle sans cette variable paraît moins bon au regard de l'AIC et du BIC.

L'AIC est de 112 537 et le BIC de 112 918. Ces estimateurs permettent de comparer des modèles entre eux, mais il est compliqué de les interpréter seuls.

Modèle GLM intégrant le score de conduite

Intégrons désormais la variable "tranche de score de conduite" :

Paramètre		Valeur estimée	Wald Chi-Square	Pr > Khi-2	Relativité
Score de conduite	**0-30**	0,83	197	<10-4	**2,29**
Score de conduite	**30-40**	0,49	181	<10-4	**1,63**
Score de conduite	**40-50**	0,31	138	<10-4	**1,37**
Score de conduite	**50-60**	0,10	19	<10-4	**1,11**
Score de conduite	**60-70**	0,00			**1,00**
Score de conduite	**70-80**	-0,18	55	<10-4	**0,83**
Score de conduite	**80-90**	-0,32	104	<10-4	**0,73**
Score de conduite	**90-100**	-0,27	48	<10-4	**0,76**
Secteur d'activité	**Activités de services**	-0,13	3	0,07	**0,88**
Secteur d'activité	**Activités financières**	-0,05	5	0,02	**0,95**
Secteur d'activité	**Activités du bâtiment**	-0,12	29	<10-4	**0,89**
Secteur d'activité	**Activités liées à l'immobilier**	0,16	4	0,04	**1,17**
Secteur d'activité	**Activités de santé**	0,27	17	<10-4	**1,32**
Secteur d'activité	**Activités de commerce de gros**	0,00			**1,00**
Année	**Y6**	0,01	0	0,63	**1,01**
Année	**Y7**	0,00			**1,00**
Année	**Y8**	-0,36	302	<10-4	**0,69**
Taille de la flotte	**0-123 d'exposition**	0,00			**1,00**
Taille de la flotte	**124-200 d'exposition**	0,01	0	0,80	**1,01**
Taille de la flotte	**201-400 d'exposition**	0,17	23	<10-4	**1,19**
Taille de la flotte	**>400 d'exposition**	-0,07	6	0,01	**0,93**
Nombre de place dans le véhicule	**0-3 places**	0,45	84	<10-4	**1,56**
Nombre de place dans le véhicule	**4 places**	0,00	0	0,98	**1,00**
Nombre de place dans le véhicule	**>5 places**	0,00			**1,00**
Age du véhicule en début de couverture	**0-2 ans**	0,00			**1,00**
Age du véhicule en début de couverture	**2-3 ans**	-0,03	2	0,17	**0,97**
Age du véhicule en début de couverture	**>3 ans**	0,01	0	0,72	**1,01**
Nombre de kilomètres au contrat	**0-19000 kilomètres par an**	-0,20	95	<10-4	**0,82**
Nombre de kilomètres au contrat	**19000-44000 kilomètres par an**	0,00			**1,00**
Nombre de kilomètres au contrat	**>44000 kilomètres par an**	0,23	78	<10-4	**1,26**
Puissance du véhicule	**0-73**	-0,11	8	0,01	**0,89**
Puissance du véhicule	**73-146**	0,00			**1,00**
Puissance du véhicule	**146-203**	-0,02	1	0,39	**0,98**
Puissance du véhicule	**203-+**	-0,20	20	<10-4	**0,82**
Poids du véhicule	**0-1966**	0,00			**1,00**
Poids du véhicule	**1966-2378**	0,02	1	0,48	**1,02**
Poids du véhicule	**2378-2665**	0,11	7	0,01	**1,11**
Poids du véhicule	**2665-3500**	0,56	164	<10-4	**1,76**
Poids du véhicule	**3500-+**	1,50	43	<10-4	**4,48**
Type de carburant du véhicule	**Electrique et hybride électrique**	-0,09	6	0,02	**0,92**
Type de carburant du véhicule	**GPL et hybride GPL**	0,20	22	<10-4	**1,22**
Type de carburant du véhicule	**Essence et diesel**	0,00			**1,00**
Marque du véhicule	**Francaise**	0,15	24	<10-4	**1,16**
Marque du véhicule	**Luxe 1**	0,25	79	<10-4	**1,28**
Marque du véhicule	**Luxe 2**	0,10	11	0,00	**1,10**
Marque du véhicule	**Populaires 1**	0,06	9	0,00	**1,07**
Marque du véhicule	**Populaires 2**	0,00			**1,00**
Province italienne	**BRESCIA**	-0,16	34	<10-4	**0,86**
Province italienne	**CAGLIARI**	0,01	0	0,89	**1,01**
Province italienne	**COSENZA**	0,08	7	0,01	**1,09**
Province italienne	**FIRENZE**	0,11	11	0,00	**1,12**
Province italienne	**NAPOLI**	0,58	270	<10-4	**1,78**
Province italienne	**ROMA**	0,16	58	<10-4	**1,17**
Province italienne	**MILANO**	0,00			**1,00**
Type de véhicule	**Van**	0,00	0	0,97	**1,00**
Type de véhicule	**Car**	0,00			**1,00**
Scale		1,00			

FIGURE 5.13 – Modèle GLM fréquence matérielle avec le score de conduite

Déviance	81 513
Log Vraissemblance	-54 908
AIC	111 702
BIC	112 155

FIGURE 5.14 – Mesures du modèle GLM fréquence matérielle avec le score de conduite

Les modalités non-significatives ont changé. On en déduit que la quantité d'informations apportée par chacune des variables a évolué.

L'AIC a diminué de 835 (passant de 112 537 à 111 702). Le BIC a également diminué de 763 (passant de 112 918 à 112 155). D'après ces indicateurs, le modèle prédictif avec la variable score de conduite serait de meilleure qualité.

Relativité de la variable "score de conduite" sur la fréquence matérielle

Observons, désormais la relativité de la variable "score de conduite" :

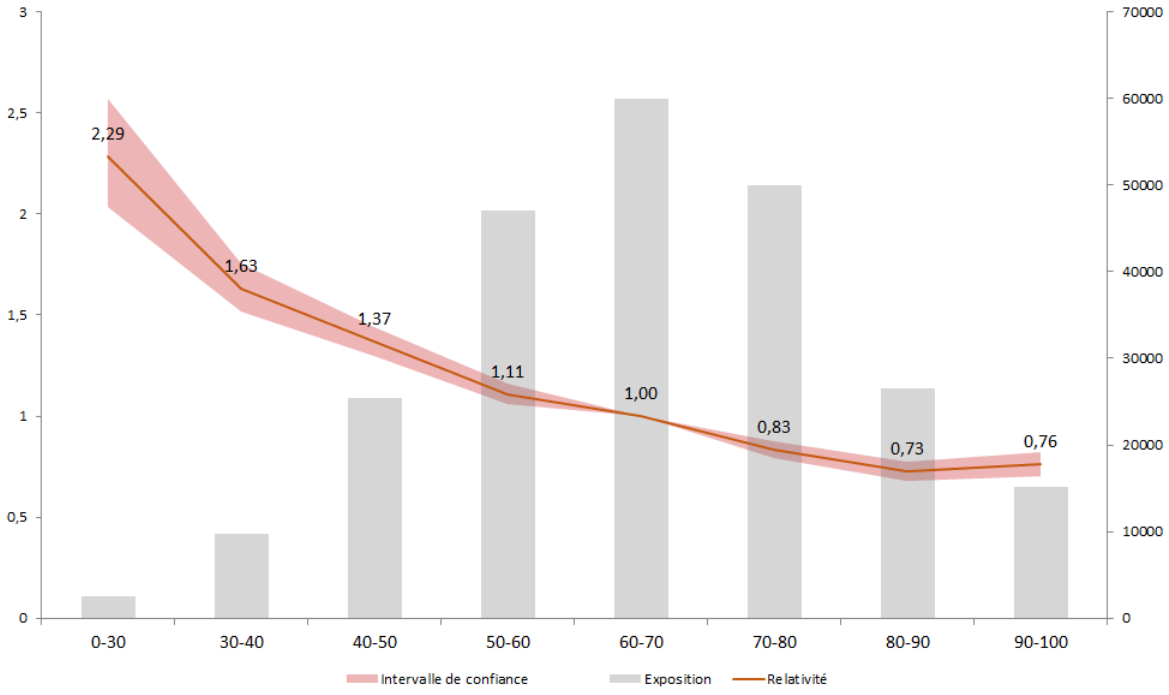

FIGURE 5.15 – Relativité score de conduite sur la fréquence matérielle

Le score de conduite a un très fort impact sur la fréquence matérielle. Par exemple, l'individu dont le score de conduite est compris entre 30 et 40 a 63% de risques supplémentaires d'avoir un sinistre matériel qu'un individu dont le score est compris entre 60 et 70.
Le score de conduite a donc un impact explicatif majeur sur la sinistralité matérielle.

Comportement des autres variables après ajout du score de conduite

Observons à présent le comportement des autres variables avec l'ajout du score de conduite. Pour ce faire, nous analyserons dans un premier temps les écarts de relativité.

Nom de la variable	Nom de la modalité	Relativité GLM sans score	Relativité GLM avec score
Score de conduite	**0-30**		2,29
Score de conduite	**30-40**		1,63
Score de conduite	**40-50**		1,37
Score de conduite	**50-60**		1,11
Score de conduite	**60-70**		1,00
Score de conduite	**70-80**		0,83
Score de conduite	**80-90**		0,73
Score de conduite	**90-100**		0,76
Secteur d'activité	**Activités de services**	0,85	0,88
Secteur d'activité	**Activités financières**	0,92	0,95
Secteur d'activité	**Activités du bâtiment**	0,88	0,89
Secteur d'activité	**Activités liées à l'immobilier**	1,15	1,17
Secteur d'activité	**Activités de santé**	1,29	1,32
Secteur d'activité	**Activités de commerce de gros**	1,00	1,00
Année	**Y6**	0,97	1,01
Année	**Y7**	1,00	1,00
Année	**Y8**	0,69	0,69
Taille de la flotte	**0-123 d'exposition**	1,00	1,00
Taille de la flotte	**124-200 d'exposition**	1,02	1,01
Taille de la flotte	**201-400 d'exposition**	1,21	1,19
Taille de la flotte	**>400 d'exposition**	0,91	0,93
Nombre de place dans le véhicule	**0-3 places**	1,67	1,56
Nombre de place dans le véhicule	**4 places**	1,06	1,00
Nombre de place dans le véhicule	**>5 places**	1,00	1,00
Age du véhicule en début de couverture	**0-2 ans**	1,00	1,00
Age du véhicule en début de couverture	**2-3 ans**	0,96	0,97
Age du véhicule en début de couverture	**>3 ans**	0,97	1,01
Nombre de kilomètres au contrat	**0-19000 kilomètres par an**	0,82	0,82
Nombre de kilomètres au contrat	**19000-44000 kilomètres par an**	1,00	1,00
Nombre de kilomètres au contrat	**>44000 kilomètres par an**	1,26	1,26
Puissance du véhicule	**0-73**	0,87	0,89
Puissance du véhicule	**73-146**	1,00	1,00
Puissance du véhicule	**146-203**	1,00	0,98
Puissance du véhicule	**203-+**	0,92	0,82
Poids du véhicule	**0-1966**	1,00	1,00
Poids du véhicule	**1966-2378**	0,98	1,02
Poids du véhicule	**2378-2665**	1,03	1,11
Poids du véhicule	**2665-3500**	1,50	1,76
Poids du véhicule	**3500-+**	3,12	4,48
Type de carburant du véhicule	**Electrique et hybride électrique**	0,87	0,92
Type de carburant du véhicule	**GPL et hybride GPL**	1,21	1,22
Type de carburant du véhicule	**Essence et diesel**	1,00	1,00
Marque du véhicule	**Francaise**	1,19	1,16
Marque du véhicule	**Luxe 1**	1,32	1,28
Marque du véhicule	**Luxe 2**	1,15	1,10
Marque du véhicule	**Populaires 1**	1,10	1,07
Marque du véhicule	**Populaires 2**	1,00	1,00
Province italienne	**BRESCIA**	0,85	0,86
Province italienne	**CAGLIARI**	1,06	1,01
Province italienne	**COSENZA**	1,05	1,09
Province italienne	**FIRENZE**	1,14	1,12
Province italienne	**NAPOLI**	1,66	1,78
Province italienne	**ROMA**	1,16	1,17
Province italienne	**MILANO**	1,00	1,00
Type de véhicule	**Van**	0,95	1,00
Type de véhicule	**Car**	1,00	1,00

FIGURE 5.16 – Relativité modèle fréquence matérielle avec et sans score de conduite

Certaines variables, comme le poids ou la puissance voient leur relativité modifiée par l'ajout du score de conduite dans le modèle. D'autres variables, en revanche, ne sont pas impactées

par le score de conduite (le nombre de kilomètres au contrat par exemple).

A ce stade, il nous paraît judicieux d'utiliser le Wald Chi-Square. C'est un indicateur qui peut nous permettre de mettre en évidence, de façon comparative, la quantité d'informations de chaque modalité dans les deux modèles :

Nom de la variable	Nom de la modalité	Wald Chi-Square GLM sans score	Wald Chi-Square GLM avec score	Score Impact
Score de conduite	**0-30**		197,03	
Score de conduite	**30-40**		180,59	
Score de conduite	**40-50**		137,67	
Score de conduite	**50-60**		19,35	
Score de conduite	**60-70**			
Score de conduite	**70-80**		55,00	
Score de conduite	**80-90**		103,53	
Score de conduite	**90-100**		48,31	
Secteur d'activité	**Activités de services**	5,17	3,27	-37%
Secteur d'activité	**Activités financières**	14,92	5,46	-63%
Secteur d'activité	**Activités du bâtiment**	33,29	29,13	-12%
Secteur d'activité	**Activités liées à l'immobilier**	3,28	4,12	26%
Secteur d'activité	**Activités de santé**	14,85	16,80	13%
Secteur d'activité	**Activités de commerce de gros**			
Année	**Y6**	2,24	0,23	-90%
Année	**Y7**			
Année	**Y8**	320,58	301,68	-6%
Taille de la flotte	**0-123 d'exposition**			
Taille de la flotte	**124-200 d'exposition**	0,12	0,07	-42%
Taille de la flotte	**201-400 d'exposition**	26,90	22,84	-15%
Taille de la flotte	**>400 d'exposition**	9,37	6,10	-35%
Nombre de place dans le véhicule	**0-3 places**	113,03	84,27	-25%
Nombre de place dans le véhicule	**4 places**	1,95	0,00	-100%
Nombre de place dans le véhicule	**>5 places**			
Age du véhicule en début de couverture	**0-2 ans**			
Age du véhicule en début de couverture	**2-3 ans**	3,70	1,85	-50%
Age du véhicule en début de couverture	**>3 ans**	1,63	0,12	-93%
Nombre de kilomètres au contrat	**0-19000 kilomètres par an**	93,76	94,78	1%
Nombre de kilomètres au contrat	**19000-44000 kilomètres par an**			
Nombre de kilomètres au contrat	**>44000 kilomètres par an**	79,51	77,84	-2%
Puissance du véhicule	**0-73**	11,20	7,79	-30%
Puissance du véhicule	**73-146**			
Puissance du véhicule	**146-203**	0,02	0,74	3600%
Puissance du véhicule	**203-+**	3,19	19,74	519%
Poids du véhicule	**0-1966**			
Poids du véhicule	**1966-2378**	0,71	0,51	-28%
Poids du véhicule	**2378-2665**	0,57	6,81	1095%
Poids du véhicule	**2665-3500**	88,83	164,04	85%
Poids du véhicule	**3500-+**	25,18	43,31	72%
Type de carburant du véhicule	**Electrique et hybride électrique**	13,91	5,59	-60%
Type de carburant du véhicule	**GPL et hybride GPL**	21,11	21,90	4%
Type de carburant du véhicule	**Essence et diesel**			
Marque du véhicule	**Francaise**	33,89	23,53	-31%
Marque du véhicule	**Luxe 1**	102,51	79,28	-23%
Marque du véhicule	**Luxe 2**	23,40	11,33	-52%
Marque du véhicule	**Populaires 1**	19,45	8,84	-55%
Marque du véhicule	**Populaires 2**			
Province italienne	**BRESCIA**	37,52	34,25	-9%
Province italienne	**CAGLIARI**	0,43	0,02	-95%
Province italienne	**COSENZA**	2,31	7,19	211%
Province italienne	**FIRENZE**	14,56	11,20	-23%
Province italienne	**NAPOLI**	211,21	269,72	28%
Province italienne	**ROMA**	53,14	58,17	9%
Province italienne	**MILANO**			
Type de véhicule	**Van**	1,27	0,00	-100%
Type de véhicule	**Car**			

FIGURE 5.17 – Wald Chi-square modèle fréquence matérielle avec et sans score de conduite

Dans le modèle sans le score de conduite, la variable apportant le plus d'informations d'après la mesure Wald Chi-square, est la zone italienne, avec un total d'environ 320 Wald Chi-square.

Dans le modèle avec le score de conduite, ce dernier apporte environ 700 Wald Chi-square, soit plus du double que la variable de la zone italienne dans le modèle sans le score de conduite. Le score de conduite a donc un très fort pouvoir explicatif.

Une deuxième observation concerne l'évolution de l'apport d'informations : les variables peuvent être impactées positivement ou négativement par l'ajout du score de conduite. Par exemple, le poids du véhicule apporte 2 fois plus d'informations avec l'ajout du score de conduite alors que la variable "marque de véhicule" perd 1/3 d'information. Cette perte d'informations s'explique si l'on considère que la variable traduisait, en réalité, une partie du comportement du conducteur, information qui a été captée par le score de conduite. De même pour les individus conduisant des véhicules de luxe : leur tendance à adopter plus fréquemment une conduite sportive a été captée par le score de conduite.

Certaines variables apportent davantage d'informations dans la modélisation car elles induisent une segmentation supplémentaire : parmi les individus dont le score se situe entre 30 et 40, le fait de vivre dans la région de Rome va impliquer une plus grande segmentation par rapport aux individus qui présentent un score plus élevé.

Sachant qu'on ne dispose d'aucune variable liée au conducteur, on peut supposer que le score de conduite s'apparente à une information sur le conducteur. Le score de conduite permet de remplacer l'information sur le style de conduite de chaque conducteur.

Les offres de tarification des flottes automobiles se basent généralement sur un nombre restreint de variables explicatives. La variable score de conduite apporte donc une partie de l'information conducteur qui manque en comparaison des modèles de tarification destinés aux particuliers.

Observons désormais l'apport du score de conduite dans la modélisation du coût moyen matériel.

5.2 Modélisation du coût moyen matériel

Pour modéliser le coût moyen matériel intégrant le score de conduite, nous procéderons en trois étapes : création des modalités, création d'un modèle complet, puis ajout du score de conduite.

La modélisation coût moyen se fait par un GLM utilisant la loi Gamma avec une fonction lien logarithme. La base de travail est une base contenant uniquement les sinistres matériels.

5.2.1 Création des modalités

Nous cherchons à identifier les variables menant à la modélisation du coût moyen matériel la plus précise. Pour ce faire, nous utilisons un processus backward.

Nous avons testé l'ensemble des variables présentes dans la base contrat mais nous avons retenu les variables suivantes :

- l'année
- l'âge du véhicule en début de couverture
- la province italienne
- le secteur d'activité
- le poids du véhicule
- la puissance du véhicule
- la marque du véhicule
- le nombre de kilomètres au contrat
- le type de véhicule
- la franchise

L'année

Cette variable est décomposée de façon identique à la modélisation de la fréquence de la sinistralité matérielle :

- Y6 de novembre 2017 à octobre 2018
- Y7 de novembre 2018 à octobre 2019
- Y8 de novembre 2019 à octobre 2020

L'impact de cette variable est traduit dans le graphique ci-dessous :

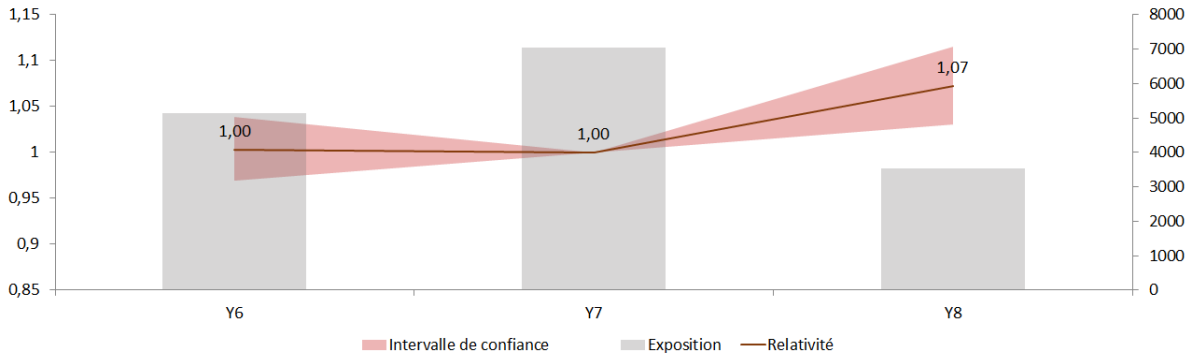

FIGURE 5.18 – Fréquence et relativité par année contrat

Les coûts moyen matériels des années Y6 et Y7 sont constants car d'après le graphique leurs relativités sont égales à 1. A contrario, le coût moyen matériel de l'année Y8 est modélisé avec une hausse de 7%. Nous expliquons cet effet par une hausse des coûts de réparation directement liée aux prix des pièces détachées qui augmentent également.

L'âge du véhicule en début de couverture

Les modalités de cette variable sont les mêmes qu'au chapitre précédent.

Les relativités et la fréquence pour chacune de ces modalités sont les suivantes :

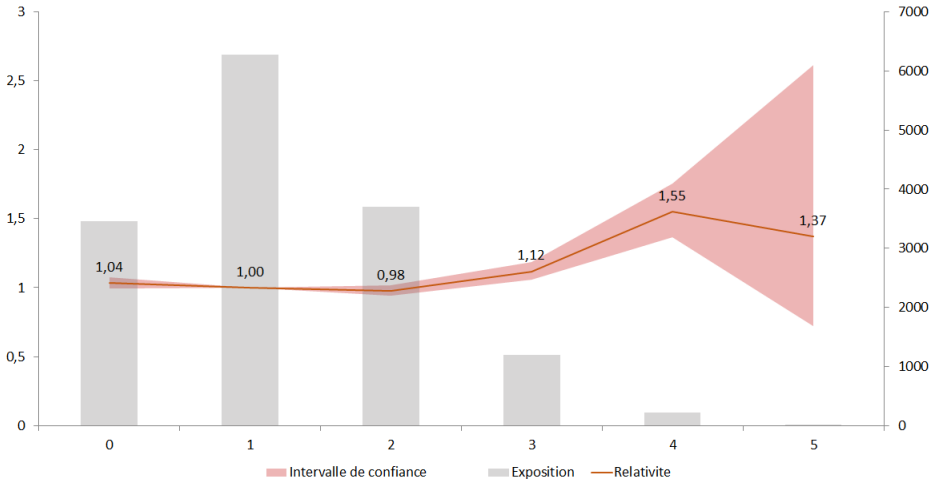

FIGURE 5.19 – Relativité et fréquence variable "âge début de couverture"

Le nombre d'observations pour cette modalité est moins important que lors de la modélisation fréquence. Cette baisse du nombre d'observations implique une moindre certitude dans les relativités et donc un risque d'erreurs plus important.

Nous avons finalement choisi de ne garder que 2 catégories : 0 - 2 et > 3.

La province italienne

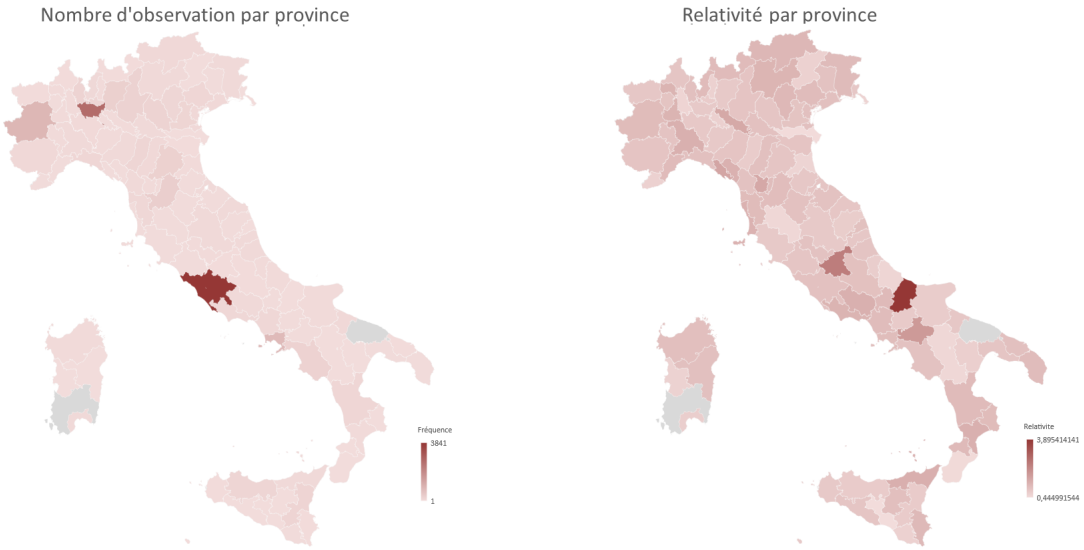

FIGURE 5.20 – Fréquence et relativité par province

Les relativités obtenues sont différentes de celles observées en fréquence, ce qui nous a conduit à réaliser des regroupements différents :

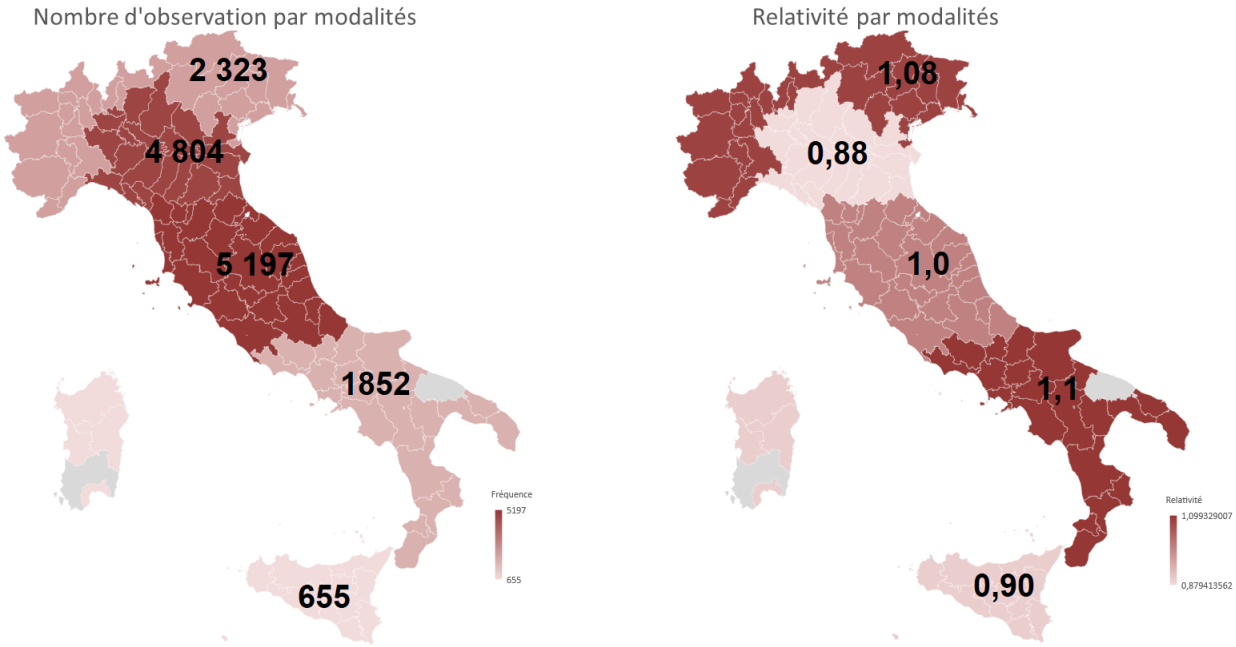

FIGURE 5.21 – Fréquence et relativité par modalité

Nous utiliserons ces classes dans le GLM.

Le secteur d'activité

Les modalités des secteurs d'activité sont les mêmes que pour la fréquence matérielle :

- Les activités de service
- Les activités financières
- Les activités du bâtiment
- Les activités liées à l'immobilier
- Les activités de santé
- Les activités de commerce de gros

Le poids et la puissance du véhicule

Comme au chapitre précédent, nous avons observé les variables poids et puissance ensembles, sous forme de rapport et en tant que variables croisées.

La variable est optimisée quand les éléments poids et puissance sont séparés :

— Poids :

- entre 0 et 1897 kg
- 1898 kg et plus

— Puissance :

- entre 0 et 71 ch
- 72 ch et plus

La marque du véhicule

En tenant en compte du poids et de la puissance des véhicules, nous avons opéré un regroupement des marques de luxe par rapport au modèle précédent. Les arbitrages des catégories populaires ont été légèrement modifiés pour s'adapter à la problématique du coût moyen matériel.

Les modalités ont été regroupées de la façon suivante :

Luxe 1	Populaires 1	Populaires 2
Abarth, Audi, Bmw, Infiniti, Jaguar, Jeep, Land Rover, Lexus, Maserati, Mercedes, Mini Cooper, Smart, Piaggio, Porsche	Dasia, DS, Fuso Canter, Great Wall, Kia, Mitsubishi, Subaru, Suzuki, Toyota Volvo, Volkswagen	Alfa Romeo, Citroen, Fiat, Ford, Honda, Hyundai, Iveco, Lancia, Mazda, Nissan, Opel, Peugeot, Renault, Renault Trucs, Seat, Ssangyong, Skoda

FIGURE 5.22 – Modalités de la variable "marque du véhicule"

Le nombre de kilomètres au contrat

Le nombre de kilomètres au contrat est significatif pour l'étude du coût moyen. Quatre modalités sont retenues :

- de 0 à 9000 kilomètres
- de 9000 à 19000 kilomètres
- de 19000 à 55000 kilomètres
- plus de 55000 kilomètres

Avec ces regroupements, nous pouvons observer les relativités suivantes :

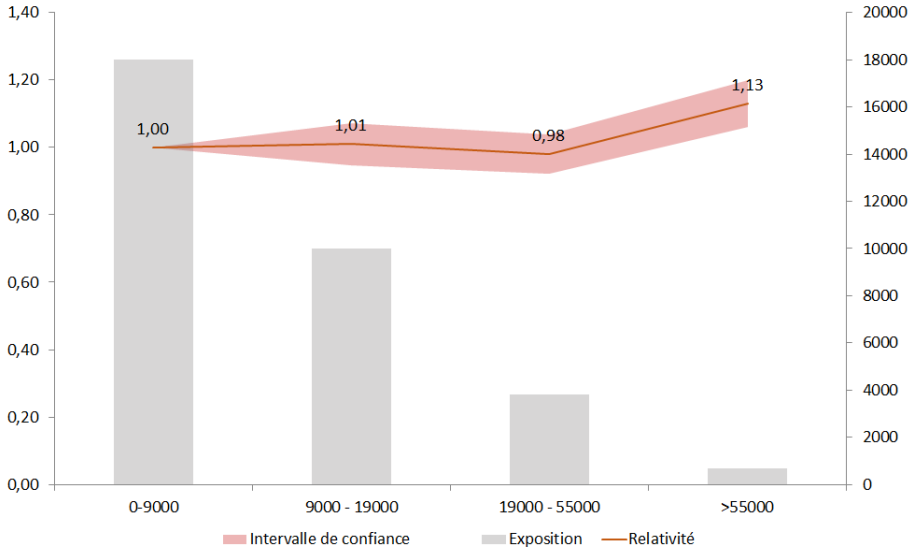

FIGURE 5.23 – Fréquence et relativité par nombre de kilomètres au contrat

Nous utiliserons ces classes dans le GLM.

Le type de véhicule

La variable "type de véhicule" a deux modalités, "car" ou "van".

La franchise

La variable franchise de sinistre est un élément très important de la modélisation du coût moyen. Elle se décompose en trois modalités :

- 0 € de franchise
- 250 € de franchise
- 500 € de franchise.

5.2.2 Mesure de l'ajout du score de conduite

Le but de cette section est d'étudier le modèle GLM avec les variables explicatives présentées précédemment, puis d'y ajouter la variable score de conduite.

Modèle GLM sans le score de conduite

Les divers regroupements effectués nous permettent d'obtenir le modèle de coût moyen matériel suivant :

Paramètre		Valeur estimée	Wald Chi-Square	Pr > Khi-2	Relativité
Secteur d'activité	**Activités de services**	-0,16	6	0,01	**0,85**
Secteur d'activité	**Activités financières**	-0,10	23	<10-4	**0,91**
Secteur d'activité	**Activités du bâtiment**	-0,07	12	0,00	**0,93**
Secteur d'activité	**Activités liées à l'immobilier**	-0,16	5	0,02	**0,85**
Secteur d'activité	**Activités de santé**	-0,21	12	0,00	**0,81**
Secteur d'activité	**Activités de commerce de gros**	0,00			**1,00**
Année	**Y6**	-0,01	0	0,48	**0,99**
Année	**Y8**	0,00			**1,00**
Année	**Y7**	0,06	10	0,00	**1,07**
Age du véhicule en début de couverture	**0-2 ans**	0,00			**1,00**
Age du véhicule en début de couverture	**3 ans**	0,06	5	0,03	**1,06**
Age du véhicule en début de couverture	**>4 ans**	0,33	32	<10-4	**1,39**
Nombre de kilomètres au contrat	**0-9000 kilomètres par an**	0,00			**1,00**
Nombre de kilomètres au contrat	**9000-19000 kilomètres par an**	0,16	9	0,00	**1,17**
Nombre de kilomètres au contrat	**19000-55000 kilomètres par an**	0,12	5	0,02	**1,13**
Nombre de kilomètres au contrat	**>55000 kilomètres par an**	0,27	16	<10-4	**1,31**
Puissance du véhicule	**0-71**	0,13	22	<10-4	**1,13**
Puissance du véhicule	**>71**	0,00			**1,00**
Poids du véhicule	**0-1897**	0,00			**1,00**
Poids du véhicule	**>1897**	-0,05	7	0,01	**0,95**
Marque du véhicule	**Luxe**	-0,05	8	0,00	**0,95**
Marque du véhicule	**Populaire 1**	-0,12	25	<10-4	**0,88**
Marque du véhicule	**Populaire 2**	0,00			**1,00**
Province italienne	**CATANIA**	-0,13	11	0,00	**0,88**
Province italienne	**MILANO**	-0,10	26	<10-4	**0,91**
Province italienne	**NAPOLI**	0,03	2	0,17	**1,04**
Province italienne	**TORINO**	0,06	8	0,00	**1,07**
Province italienne	**ROMA**	0,00			**1,00**
Type de véhicule	**Van**	0,01	0	0,79	**1,01**
Type de véhicule	**Car**	0,00			**1,00**
Franchise	**0**	-0,12	46	<10-4	**0,88**
Franchise	**150**	-0,08	7	0,01	**0,93**
Franchise	**250**	0,00			**1,00**
Scale		1,20			

FIGURE 5.24 – Modèle GLM coût moyen matériel sans le score de conduite

Déviance	14 333
Log Vraissemblance	-131 505
AIC	263 059
BIC	263 250

FIGURE 5.25 – Mesures du modèle GLM coût moyen matériel sans le score de conduite

L'AIC est de 263 059 et le BIC de 263 250 pour ce modèle.

Modèle GLM intégrant le score de conduite

En intégrant le score de conduite on obtient le modèle GLM suivant :

Paramètre		Valeur estimée	Wald Chi-Square	Pr > Khi-2	Relativité
Score de conduite	**0-30**	0,25	23	<10-4	**1,29**
Score de conduite	**30-40**	0,12	14	0,00	**1,13**
Score de conduite	**40-50**	0,09	13	0,00	**1,09**
Score de conduite	**50-60**	0,01	0	0,78	**1,01**
Score de conduite	**60-70**	0,00			**1,00**
Score de conduite	**70-80**	-0,06	7	0,01	**0,94**
Score de conduite	**80-90**	-0,10	12	0,00	**0,90**
Score de conduite	**90-100**	-0,14	15	<10-4	**0,87**
Secteur d'activité	**Activités de services**	-0,16	6	0,01	**0,85**
Secteur d'activité	**Activités financières**	-0,09	21	<10-4	**0,91**
Secteur d'activité	**Activités du bâtiment**	-0,07	13	0,00	**0,93**
Secteur d'activité	**Activités liées à l'immobilier**	-0,17	6	0,02	**0,85**
Secteur d'activité	**Activités de santé**	-0,22	12	0,00	**0,80**
Secteur d'activité	**Activités de commerce de gros**	0,00			**1,00**
Année	**Y6**	0,00	0	0,85	**1,00**
Année	**Y8**	0,00			**1,00**
Année	**Y7**	0,07	12	0,00	**1,07**
Age du véhicule en début de couverture	**0-2 ans**	0,00			**1,00**
Age du véhicule en début de couverture	**3 ans**	0,08	8	0,00	**1,08**
Age du véhicule en début de couverture	**>4 ans**	0,37	41	<10-4	**1,45**
Nombre de kilomètres au contrat	**0-9000 kilomètres par an**	0,00			**1,00**
Nombre de kilomètres au contrat	**9000-19000 kilomètres par an**	0,14	7	0,01	**1,15**
Nombre de kilomètres au contrat	**19000-55000 kilomètres par an**	0,11	4	0,04	**1,12**
Nombre de kilomètres au contrat	**>55000 kilomètres par an**	0,25	14	0,00	**1,28**
Puissance du véhicule	**0-71**	0,12	19	<10-4	**1,12**
Puissance du véhicule	**>71**	0,00			**1,00**
Poids du véhicule	**0-1897**	0,00			**1,00**
Poids du véhicule	**>1897**	-0,03	3	0,08	**0,97**
Marque du véhicule	**Luxe**	-0,06	12	0,00	**0,94**
Marque du véhicule	**Populaire 1**	-0,11	21	<10-4	**0,89**
Marque du véhicule	**Populaire 2**	0,00			**1,00**
Province italienne	**CATANIA**	-0,14	14	0,00	**0,87**
Province italienne	**MILANO**	-0,10	28	<10-4	**0,90**
Province italienne	**NAPOLI**	0,04	3	0,08	**1,05**
Province italienne	**TORINO**	0,05	6	0,02	**1,06**
Province italienne	**ROMA**	0,00			**1,00**
Type de véhicule	**Van**	0,03	2	0,17	**1,03**
Type de véhicule	**Car**	0,00			**1,00**
Franchise	**0**	-0,12	42	<10-4	**0,89**
Franchise	**150**	-0,07	6	0,02	**0,93**
Franchise	**250**	0,00			**1,00**
Scale		1,21			

FIGURE 5.26 – Modèle GLM coût moyen matériel avec le score de conduite

Déviance	14 246
Log Vraissemblance	-131 452
AIC	262 968
BIC	263 212

FIGURE 5.27 – Mesures du modèle GLM coût moyen matériel avec le score de conduite

De même que pour la modélisation en fréquence, l'ajout de la variable score de conduite modifie les p-value de toutes les modalités. Cependant, et contrairement au modèle de fréquence matérielle, toutes les modalités du score de conduite ne sont pas significatives.

L'AIC a diminué à 262 968 soit une baisse de 91. Le BIC a également diminué de 38 à 263 212. D'après l'AIC et le BIC, le modèle prédictif avec la variable score de conduite est meilleur.

Relativité de la variable score de conduite sur le coût moyen matériel

On a donc estimé l'impact du score de conduite sur la modélisation :

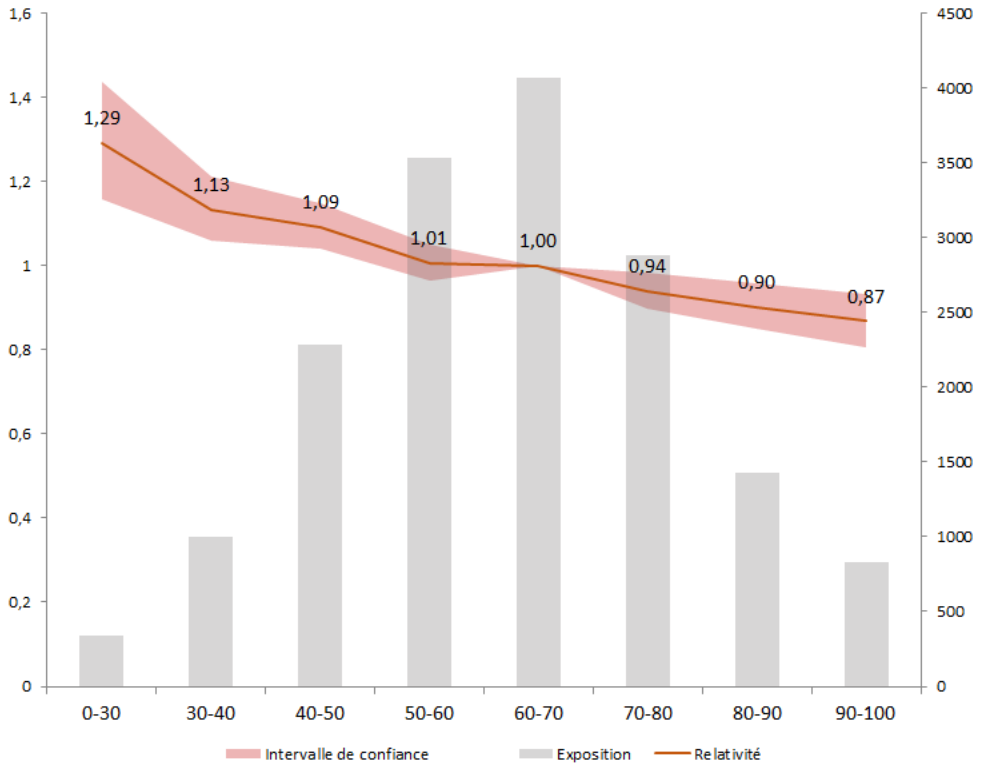

FIGURE 5.28 – Relativité score de conduite sur le coût moyen matériel

Le score de conduite apporte une importante information pour l'explication du coût moyen matériel. Il permet une segmentation supplémentaire car il implique, d'après le graphique, une augmentation de 29% du coût moyen matériel pour les scores de 0 à 30, toutes choses égales par ailleurs.

Comportement des autres variables après ajout du score de conduite

Observons la différence de relativité entre les deux modèles :

Nom de la variable	Nom de la modalité	Relativité GLM sans score	Relativité GLM avec score
Score de conduite	**0-30**		1,29
Score de conduite	**30-40**		1,13
Score de conduite	**40-50**		1,09
Score de conduite	**50-60**		1,01
Score de conduite	**60-70**		1,00
Score de conduite	**70-80**		0,94
Score de conduite	**80-90**		0,90
Score de conduite	**90-100**		0,87
Secteur d'activité	**Activités de services**	0,85	0,85
Secteur d'activité	**Activités financières**	0,91	0,91
Secteur d'activité	**Activités du bâtiment**	0,93	0,93
Secteur d'activité	**Activités liées à l'immobilier**	0,85	0,85
Secteur d'activité	**Activités de santé**	0,81	0,80
Secteur d'activité	**Activités de commerce de gros**	1,00	1,00
Année	**Y6**	0,99	1,00
Année	**Y8**	1,00	1,00
Année	**Y7**	1,07	1,07
Age du véhicule en début de couverture	**0-2 ans**	1,00	1,00
Age du véhicule en début de couverture	**3 ans**	1,06	1,08
Age du véhicule en début de couverture	**>4 ans**	1,39	1,45
Nombre de kilomètres au contrat	**0-9000 kilomètres par an**	1,00	1,00
Nombre de kilomètres au contrat	**9000-19000 kilomètres par an**	1,17	1,15
Nombre de kilomètres au contrat	**19000-55000 kilomètres par an**	1,13	1,12
Nombre de kilomètres au contrat	**>55000 kilomètres par an**	1,31	1,28
Puissance du véhicule	**0-71**	1,13	1,12
Puissance du véhicule	**>71**	1,00	1,00
Poids du véhicule	**0-1897**	1,00	1,00
Poids du véhicule	**>1897**	0,95	0,97
Marque du véhicule	**Luxe**	0,95	0,94
Marque du véhicule	**Populaire 1**	0,88	0,89
Marque du véhicule	**Populaire 2**	1,00	1,00
Province italienne	**CATANIA**	0,88	0,87
Province italienne	**MILANO**	0,91	0,90
Province italienne	**NAPOLI**	1,04	1,05
Province italienne	**TORINO**	1,07	1,06
Province italienne	**ROMA**	1,00	1,00
Type de véhicule	**Van**	1,01	1,03
Type de véhicule	**Car**	1,00	1,00
Franchise	**0**	0,88	0,89
Franchise	**150**	0,93	0,93
Franchise	**250**	1,00	1,00

FIGURE 5.29 – Ecart de relativité entre GLM coût moyen matériel avec ou sans score de conduite

Les relativités ne sont que très peu modifiées par l'apport du score de conduite : on peut donc supposer que l'impact principal du score de conduite dans la modélisation du coût moyen matériel est un ajout d'informations qui n'était pas présent dans les autres variables.

Confirmons notre hypothèse grâce à la comparaison du Wald Chi-square entre les deux modèles :

Nom de la variable	Nom de la modalité	Wald Chi-Square GLM sans score	Wald Chi-Square GLM avec score	Score Impact
Score de conduite	**0-30**		22,56	
Score de conduite	**30-40**		13,63	
Score de conduite	**40-50**		12,61	
Score de conduite	**50-60**		0,08	
Score de conduite	**60-70**			
Score de conduite	**70-80**		7,33	
Score de conduite	**80-90**		12,11	
Score de conduite	**90-100**		15,27	
Secteur d'activité	**Activités de services**	6,06	6,21	2%
Secteur d'activité	**Activités financières**	22,84	20,66	-10%
Secteur d'activité	**Activités du bâtiment**	11,50	12,68	10%
Secteur d'activité	**Activités liées à l'immobilier**	5,17	5,58	8%
Secteur d'activité	**Activités de santé**	11,70	12,43	6%
Secteur d'activité	**Activités de commerce de gros**			
Année	**Y6**	0,49	0,04	-92%
Année	**Y8**			
Année	**Y7**	10,38	12,48	20%
Age du véhicule en début de couverture	**3 ans**			
Age du véhicule en début de couverture	**>4 ans**	4,75	8,26	74%
Age du véhicule en début de couverture	**0-2 ans**	31,81	40,92	29%
Nombre de kilomètres au contrat	**9000-19000 kilomètres par an**			
Nombre de kilomètres au contrat	**19000-55000 kilomètres par an**	8,67	7,18	-17%
Nombre de kilomètres au contrat	**>55000 kilomètres par an**	5,22	4,22	-19%
Nombre de kilomètres au contrat	**0-9000 kilomètres par an**	16,32	14,04	-14%
Puissance du véhicule	**0-71**	22,16	19,33	-13%
Puissance du véhicule	**>71**			
Poids du véhicule	**>1897**			
Poids du véhicule	**0-1897**	7,15	3,12	-56%
Marque du véhicule	**Luxe**	8,36	11,88	42%
Marque du véhicule	**Populaire 1**	24,80	21,23	-14%
Marque du véhicule	**Populaire 2**			
Province italienne	**CATANIA**	11,26	13,73	22%
Province italienne	**MILANO**	26,16	28,34	8%
Province italienne	**NAPOLI**	1,92	3,17	65%
Province italienne	**TORINO**	8,07	5,73	-29%
Province italienne	**ROMA**			
Type de véhicule	**Van**	0,07	1,92	2643%
Type de véhicule	**Car**			
Franchise	**0**	45,70	42,23	-8%
Franchise	**150**	7,06	5,75	-19%
Franchise	**250**			

FIGURE 5.30 – Ecart de Wald Chi-square entre GLM coût moyen matériel avec et sans score de conduite

La somme des Wald Chi-square des variables du modèle GLM sans score de conduite est égale à 297,6 contre 301,1 pour les mêmes variables dans le modèle avec score de conduite. Ce résultat confirme notre observation précédente selon laquelle le score de conduite ne modifie que très légèrement l'information apportée par les autres variables dans l'explication du coût moyen matériel.

De plus, les modalités du score de conduite sont significatives donc cette variable ajoute de l'information supplémentaire sans modifier l'information apportée par les autres variables dans la modélisation du coût moyen matériel. Sachant que nous ne disposons d'aucune information conducteur, nous pouvons supposer que le score de conduite comble en partie ce manque.

Le score de conduite apporte une quantité d'informations non négligeable dans les modélisations de fréquence matérielle et de coût moyen matériel. Dans ces deux modèles, le score de conduite est l'unique information conducteur. Cette variable a un impact positif certain dans nos modélisations de sinistralité matérielle ; analysons désormais son impact dans la modélisation des sinistres corporels.

Chapitre 6

Modélisation de la sinistralité responsabilité civile corporelle

Comme pour la sinistralité matérielle, la sinistralité corporelle sera étudiée d'abord sur sa fréquence puis sur son coût moyen.

La fréquence des sinistres corporels est plus faible que celle des sinistres matériels : il y a donc plus de volatilité sur ces modélisations que sur celles qui concernent la sinistralité matérielle.

6.1 Modélisation de la fréquence des sinistres corporels

Le plan de réalisation de cette modélisation est le même que dans les cas précédents : création de différentes modalités et mesure de l'information apportée par la variable score de conduite dans la modélisation.

Le modèle utilisé est un modèle de Poisson avec une loi lien logarithme népérien.

6.1.1 Création des modalités

Nous avons testé l'ensemble des variables présentes dans la base contrat mais nous avons retenu les variables suivantes :

- l'année
- la province italienne
- le poids du véhicule
- la puissance du véhicule
- la marque du véhicule
- le nombre de kilomètres au contrat

- le nombre de place dans le véhicule
- le type de véhicule
- le secteur d'activité
- la taille de la flotte

L'année

Observons le nombre d'observations par année contrat ainsi que leur relativité :

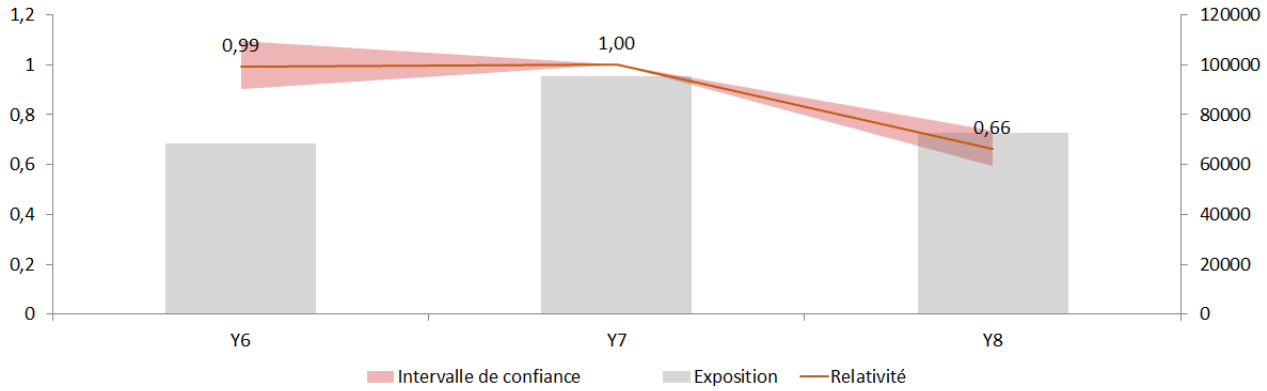

FIGURE 6.1 – Fréquence et relativité par année contrat

De même que pour la sinistralité matérielle, la fréquence de la sinistralité corporelle est réévaluée à la baisse afin d'absorber l'impact de la Covid 19.

La province italienne

Observons le nombre d'observations par province ainsi que leur relativité :

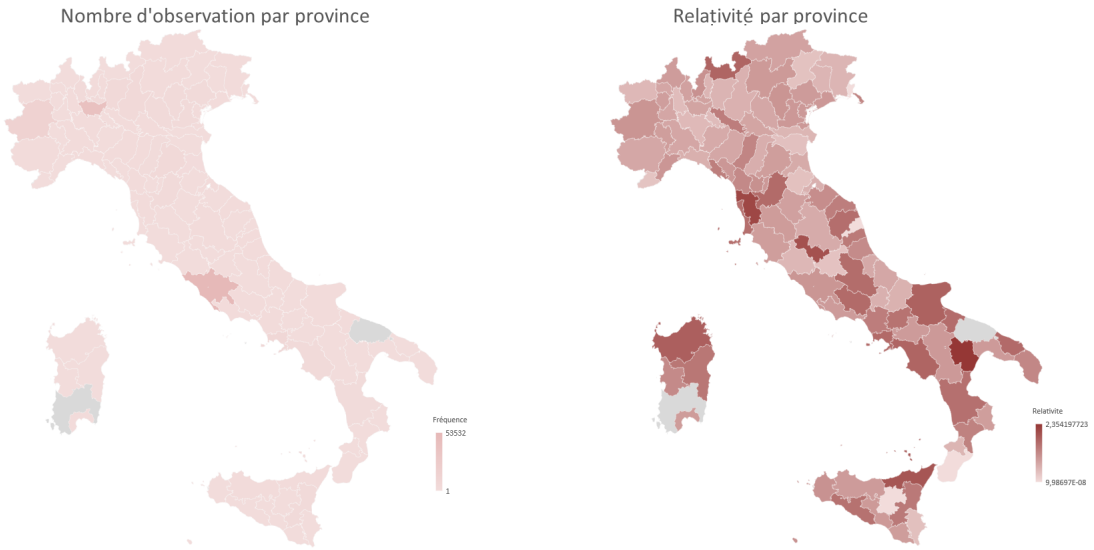

FIGURE 6.2 – Fréquence et relativité par province

Comme pour les modélisations précédentes, nous procédons aux regroupements suivants :

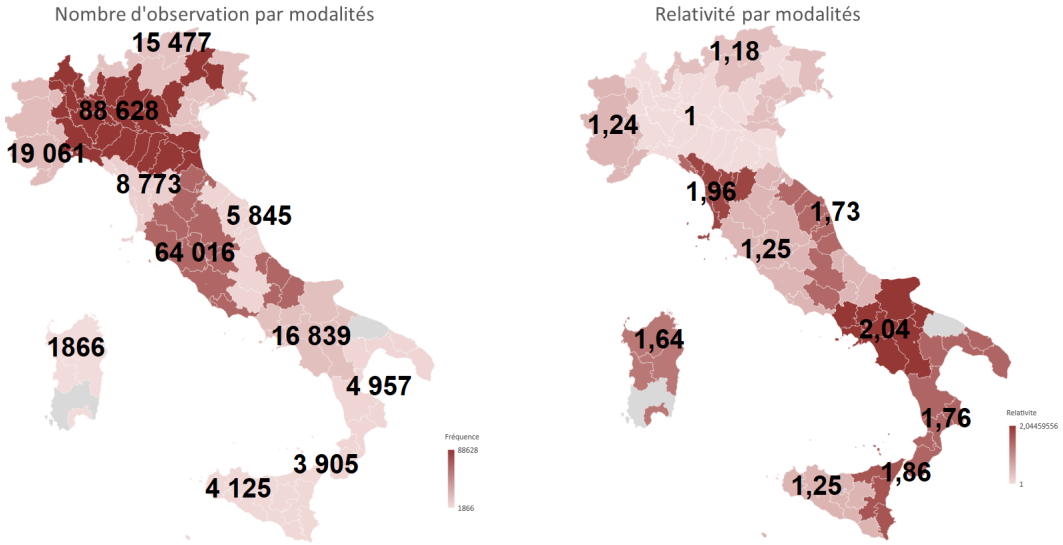

FIGURE 6.3 – Fréquence et relativité par modalité

Nous utiliserons ces classes dans le GLM.

Le poids et la puissance du véhicule

Comme dans les modélisations précédentes, et afin d'optimiser l'utilisation du poids et de la puissance dans nos modèles de fréquence de la sinistralité corporelle, nous avons comparé

l'utilisation de la variable croisée poids et puissance, l'utilisation des variables poids et puissance séparées et enfin le rapport poids puissance.

Il est apparu que les variables poids et puissance séparées ne sont pas significatives et qu'il semblait plus précis d'utiliser dans ce cas la variable rapport poids puissance.

Cette variable a ainsi été scindée en trois modalités :

- rapport compris entre 0 et 15
- rapport compris entre 15 et 22
- rapport supérieur à 22

La marque du véhicule

Cette variable s'est construite en gardant la même logique que précédemment, en identifiant des véhicules "luxe" et d'autres en "populaire".

Cette variable est composée des quatre modalités suivantes :

Luxe 1	Luxe 2	Populaire 1	Populaire 2
Abarth, Audi, Great Wall, Lancia, Lexus, Mini Cooper, Porsche	Bmw, Infiniti, Isuzu, Iveco, Jeep, Jaguar, Maserati, Mercedes, Land Rover, Lamborghini, Smart, Piaggio, Bentley	Alfa Romeo, Fuso Canter, Kia, Mazda, Mitsubishi, Nissan, Renault Trucks, Seat, Skoda, Ssangyong, Subaru	Citroen, Dacia, Ds, Fiat, Ford, Honda, Hyundai, Opel, Peugeot, Suzuki, Renault,

FIGURE 6.4 – Modalités de la variable "marque du véhicule"

Le nombre de kilomètres au contrat

Cette variable a été scindée en six modalités : :

- entre 0 et 9000 kilomètres
- entre 9000 et 19000 kilomètres
- entre 19000 et 25000 kilomètres
- entre 25000 et 40000 kilomètres
- entre 40000 et 55000 kilomètres
- plus de 55000 kilomètres

Le nombre de places dans le véhicule

Le nombre de places donne lieu à la réalisation de trois modalités :

- entre 0 et 2 places
- 3 places
- 4 places et plus

Le secteur d'activité

Les mêmes modalités que pour la fréquence et le coût moyen sont utilisées pour cette variable.

La taille de la flotte

La taille des flottes a été divisée en quatre modalités :

- les très petites flottes ayant moins de 1 d'exposition
- les petites flottes ayant entre 1 et 33 d'exposition
- les grandes flottes qui ont entre 33 et 657 d'exposition
- les flottes de plus de 657 : cette modalité représente les très grosses flottes automobiles, elles sont peu nombreuses mais doivent être caractérisées car elles représentent un très grand nombre de véhicules.

6.1.2 Mesure de l'ajout du score de conduite

Le but de cette section est d'étudier le modèle GLM avec les variables explicatives présentées précédemment, puis d'y ajouter la variable score de conduite.

Modèle GLM sans le score de conduite

Les variables ainsi optimisées nous permettent d'obtenir la modélisation suivante pour la fréquence de sinistres corporels :

Paramètre		Valeur estimée	Wald Chi-Square	Pr > Khi-2	Relativité
Secteur d'activité	**Activités de services**	-0,34	3	0,07	**0,71**
Secteur d'activité	**Activités financières**	-0,13	5	0,02	**0,88**
Secteur d'activité	**Activités du bâtiment**	-0,22	15	0,00	**0,80**
Secteur d'activité	**Activités liées à l'immobilier**	0,21	1	0,27	**1,24**
Secteur d'activité	**Activités de santé**	-0,01	0	0,97	**0,99**
Secteur d'activité	**Activités de commerce de gros**	0,00			**1,00**
Rapport poids-puissance	**0-15**	-0,11	5	0,03	**0,89**
Rapport poids-puissance	**15-22**	0,00			**1,00**
Rapport poids-puissance	**>22**	-0,30	12	0,00	**0,74**
Année	**Y6**	-0,05	1	0,26	**0,95**
Année	**Y7**	0,00			**1,00**
Année	**Y8**	-0,44	67	<10-4	**0,64**
Taille de la flotte	**1 d'exposition**	0,56	50	<10-4	**1,75**
Taille de la flotte	**2-33 d'exposition**	0,00			**1,00**
Taille de la flotte	**34-657 d'exposition**	0,14	7	0,01	**1,14**
Taille de la flotte	**>657 d'exposition**	-0,22	5	0,03	**0,80**
Nombre de place dans le véhicule	**0-2 places**	0,49	32	<10-4	**1,63**
Nombre de place dans le véhicule	**3 places**	0,75	47	<10-4	**2,11**
Nombre de place dans le véhicule	**>4 places**	0,00			**1,00**
Nombre de kilomètres au contrat	**0-9000 kilomètres par an**	0,00			**1,00**
Nombre de kilomètres au contrat	**9000-19000 kilomètres par an**	0,67	13	0,00	**1,95**
Nombre de kilomètres au contrat	**19000-25000 kilomètres par an**	0,82	20	<10-4	**2,27**
Nombre de kilomètres au contrat	**25000-40000 kilomètres par an**	0,86	22	<10-4	**2,36**
Nombre de kilomètres au contrat	**40000-55000 kilomètres par an**	1,17	36	<10-4	**3,21**
Nombre de kilomètres au contrat	**>55000 kilomètres par an**	1,27	37	<10-4	**3,57**
Marque du véhicule	**Luxe 1**	-0,30	9	0,00	**0,74**
Marque du véhicule	**Luxe 2**	-0,04	1	0,47	**0,96**
Marque du véhicule	**Populaire 1**	-0,27	13	0,00	**0,76**
Marque du véhicule	**Populaire 2**	0,00			**1,00**
Province italienne	**BOLZANO**	0,30	13	0,00	**1,35**
Province italienne	**CAGLIARI**	0,22	1	0,35	**1,25**
Province italienne	**CATANIA**	0,71	27	<10-4	**2,04**
Province italienne	**COSENZA**	0,44	17	<10-4	**1,56**
Province italienne	**FIRENZE**	0,63	45	<10-4	**1,88**
Province italienne	**NAPOLI**	0,74	86	<10-4	**2,09**
Province italienne	**PALERMO**	0,09	0	0,56	**1,09**
Province italienne	**ROMA**	0,26	22	<10-4	**1,30**
Province italienne	**TERAMO**	0,63	29	<10-4	**1,87**
Province italienne	**TORINO**	-0,02	0	0,91	**0,98**
Province italienne	**MILANO**	0,00			**1,00**
Type de véhicule	**Van**	-0,04	0	0,67	**0,96**
Type de véhicule	**Car**	0,00			**1,00**
Scale		1,00			

FIGURE 6.5 – Modèle GLM fréquence corporelle sans le score de conduite

Déviance	20 212
Log Vraissemblance	-12 381
AIC	24 878
BIC	25 224

FIGURE 6.6 – Mesures du modèle GLM fréquence corporelle sans le score de conduite

La fréquence de sinistres corporels est relativement faible, moins de 1% en moyenne, et implique la non-significativité de nombreuses variables.

Modèle GLM intégrant le score de conduite

En intégrant la variable "score de conduite", nous obtenons le modèle suivant :

Paramètre		Valeur estimée	Wald Chi-Square	Pr > Khi-2	Relativité
Score de conduite	**0-30**	1,23	100	<10-4	**3,42**
Score de conduite	**30-40**	0,71	68	<10-4	**2,04**
Score de conduite	**40-50**	0,50	54	<10-4	**1,64**
Score de conduite	**50-60**	0,18	9	0,00	**1,20**
Score de conduite	**60-70**	0,00			**1,00**
Score de conduite	**70-80**	-0,11	3	0,09	**0,89**
Score de conduite	**80-90**	-0,40	19	<10-4	**0,67**
Score de conduite	**90-100**	-0,42	13	0,00	**0,66**
Secteur d'activité	**Activités de services**	-0,29	2	0,12	**0,74**
Secteur d'activité	**Activités financières**	-0,09	2	0,13	**0,92**
Secteur d'activité	**Activités du bâtiment**	-0,21	12	0,00	**0,81**
Secteur d'activité	**Activités liées à l'immobilier**	0,24	2	0,21	**1,27**
Secteur d'activité	**Activités de santé**	0,02	0	0,91	**1,02**
Secteur d'activité	**Activités de commerce de gros**	0,00			**1,00**
Rapport poids-puissance	**0-15**	-0,14	7	0,01	**0,87**
Rapport poids-puissance	**15-22**	0,00			**1,00**
Rapport poids-puissance	**>22**	-0,20	5	0,02	**0,82**
Année	**Y6**	-0,01	0	0,89	**0,99**
Année	**Y7**	0,00			**1,00**
Année	**Y8**	-0,41	60	<10-4	**0,66**
Taille de la flotte	**1 d'exposition**	0,53	46	<10-4	**1,70**
Taille de la flotte	**2-33 d'exposition**	0,00			**1,00**
Taille de la flotte	**34-657 d'exposition**	0,13	6	0,02	**1,14**
Taille de la flotte	**>657 d'exposition**	-0,18	3	0,07	**0,83**
Nombre de place dans le véhicule	**0-2 places**	0,31	13	0,00	**1,37**
Nombre de place dans le véhicule	**3 places**	0,72	43	<10-4	**2,06**
Nombre de place dans le véhicule	**>4 places**	0,00			**1,00**
Nombre de kilomètres au contrat	**0-9000 kilomètres par an**	0,00			**1,00**
Nombre de kilomètres au contrat	**9000-19000 kilomètres par an**	0,63	12	0,00	**1,88**
Nombre de kilomètres au contrat	**19000-25000 kilomètres par an**	0,80	19	<10-4	**2,23**
Nombre de kilomètres au contrat	**25000-40000 kilomètres par an**	0,84	21	<10-4	**2,32**
Nombre de kilomètres au contrat	**40000-55000 kilomètres par an**	1,13	34	<10-4	**3,11**
Nombre de kilomètres au contrat	**>55000 kilomètres par an**	1,23	35	<10-4	**3,43**
Marque du véhicule	**Luxe 1**	-0,32	10	0,00	**0,73**
Marque du véhicule	**Luxe 2**	-0,05	1	0,37	**0,95**
Marque du véhicule	**Populaire 1**	-0,27	12	0,00	**0,77**
Marque du véhicule	**Populaire 2**	0,00			**1,00**
Province italienne	**BOLZANO**	0,29	13	0,00	**1,34**
Province italienne	**CAGLIARI**	0,21	1	0,38	**1,23**
Province italienne	**CATANIA**	0,67	24	<10-4	**1,95**
Province italienne	**COSENZA**	0,45	18	<10-4	**1,58**
Province italienne	**FIRENZE**	0,59	40	<10-4	**1,81**
Province italienne	**NAPOLI**	0,80	101	<10-4	**2,22**
Province italienne	**PALERMO**	0,12	1	0,45	**1,12**
Province italienne	**ROMA**	0,24	19	<10-4	**1,27**
Province italienne	**TERAMO**	0,65	31	<10-4	**1,91**
Province italienne	**TORINO**	-0,01	0	0,97	**0,99**
Province italienne	**MILANO**	0,00			**1,00**
Type de véhicule	**Van**	0,06	0	0,53	**1,06**
Type de véhicule	**Car**	0,00			**1,00**
Scale		1,00			

FIGURE 6.7 – Modèle GLM fréquence corporelle avec le score de conduite

Déviance	19 947
Log Vraissemblance	-12 249
AIC	24 627
BIC	25 045

FIGURE 6.8 – Mesures du modèle GLM fréquence corporelle avec le score de conduite

La variable "secteur d'activité" a perdu une partie importante de son pouvoir explicatif. Le secteur d'activité reflétait principalement la dangerosité en terme d'apparition des sinistres corporels.

Cette information télématique estimée par l'intermédiaire du score de conduite impacte les indicateurs à la baisse : 251 pour l'AIC et 179 pour le BIC. D'après ces indicateurs, la variable "score de conduite" permet donc une modélisation plus précise.

Relativité de la variable "score de conduite" sur la fréquence corporelle

La relativité de la variable "score de conduite" est la suivante :

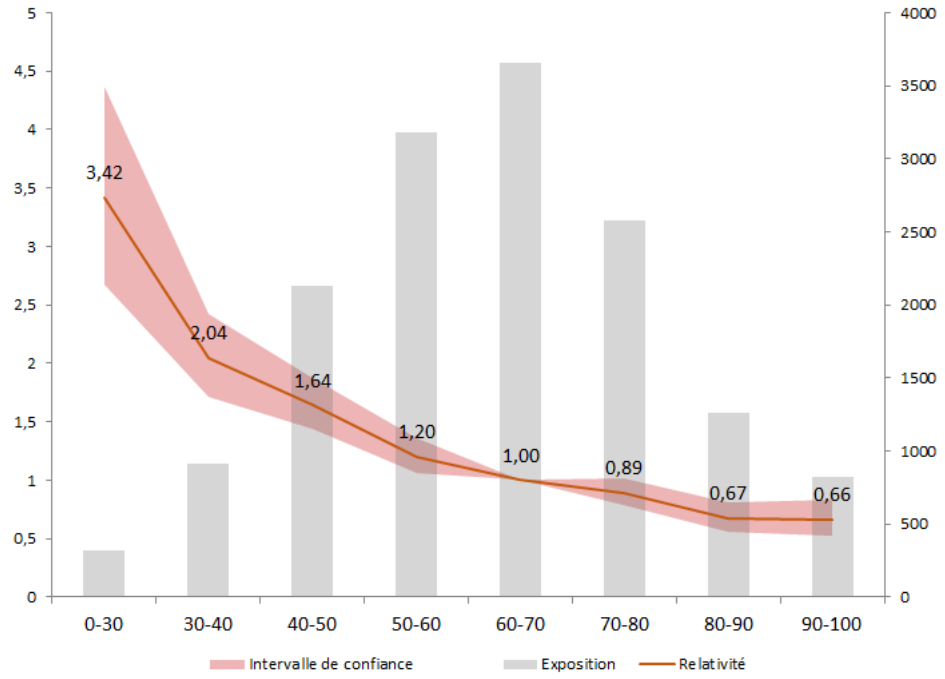

FIGURE 6.9 – Relativité score de conduite sur la fréquence de sinistres corporels

Le score de conduite est très segmentant pour la modélisation de la fréquence de sinistralité corporelle : plus le score de conduite est haut plus la probabilité d'avoir un sinistre corporel est faible.

Comportement des autres variables après ajout du score de conduite

Les différences des relativités entre les deux modèles sont estimées grâce au Wald Chi-square :

Nom de la variable	Nom de la modalité	Relativité GLM sans score	Relativité GLM avec score
Score de conduite	**0-30**		3,42
Score de conduite	**30-40**		2,04
Score de conduite	**40-50**		1,64
Score de conduite	**50-60**		1,00
Score de conduite	**60-70**		0,89
Score de conduite	**70-80**		0,89
Score de conduite	**80-90**		0,67
Score de conduite	**90-100**		0,66
Secteur d'activité	**Activités de services**	0,71	0,74
Secteur d'activité	**Activités financières**	0,88	0,92
Secteur d'activité	**Activités du bâtiment**	0,80	0,81
Secteur d'activité	**Activités liées à l'immobilier**	1,24	1,27
Secteur d'activité	**Activités de santé**	0,99	1,02
Secteur d'activité	**Activités de commerce de gros**	1,00	1,00
Rapport poids-puissance	**0-15**	0,89	0,87
Rapport poids-puissance	**15-22**	1,00	1,00
Rapport poids-puissance	**>22**	0,74	0,82
Année	**Y6**	0,95	0,99
Année	**Y7**	1,00	1,00
Année	**Y8**	0,64	0,66
Taille de la flotte	**1 d'exposition**	1,75	1,70
Taille de la flotte	**2-33 d'exposition**	1,00	1,00
Taille de la flotte	**34-657 d'exposition**	1,14	1,14
Taille de la flotte	**>657 d'exposition**	0,80	0,83
Nombre de place dans le véhicule	**0-2 places**	1,63	1,37
Nombre de place dans le véhicule	**3 places**	2,11	2,06
Nombre de place dans le véhicule	**>4 places**	1,00	1,00
Nombre de kilomètres au contrat	**0-9000 kilomètres par an**	1,00	1,00
Nombre de kilomètres au contrat	**9000-19000 kilomètres par an**	1,95	1,88
Nombre de kilomètres au contrat	**19000-25000 kilomètres par an**	2,27	2,23
Nombre de kilomètres au contrat	**25000-40000 kilomètres par an**	2,36	2,32
Nombre de kilomètres au contrat	**40000-55000 kilomètres par an**	3,21	3,11
Nombre de kilomètres au contrat	**>55000 kilomètres par an**	3,57	3,43
Marque du véhicule	**Luxe 1**	0,74	0,73
Marque du véhicule	**Luxe 2**	0,96	0,95
Marque du véhicule	**Populaire 1**	0,76	0,77
Marque du véhicule	**Populaire 2**	1,00	1,00
Province italienne	**BOLZANO**	1,35	1,34
Province italienne	**CAGLIARI**	1,25	1,23
Province italienne	**CATANIA**	2,04	1,95
Province italienne	**COSENZA**	1,56	1,58
Province italienne	**FIRENZE**	1,88	1,81
Province italienne	**NAPOLI**	2,09	2,22
Province italienne	**PALERMO**	1,09	1,12
Province italienne	**ROMA**	1,30	1,27
Province italienne	**TERAMO**	1,87	1,91
Province italienne	**TORINO**	0,98	0,99
Province italienne	**MILANO**	1,00	1,00
Type de véhicule	**Van**	0,96	1,06
Type de véhicule	**Car**	1,00	1,00

FIGURE 6.10 – Ecart de relativité entre GLM fréquence corporelle avec et sans score de conduite

La majorité des relativités varie avec l'intégration de la variable "score de conduite" dans le modèle de fréquence corporelle. La variable "score de conduite" capte un certain nombre d'informations présentes dans les autres variables.

Les écarts de Wald Chi-square sont les suivants :

Nom de la variable	Nom de la modalité	Wald Chi-Square GLM sans score	Wald Chi-Square GLM avec score	Score Impact
Score de conduite	**0-30**		100,05	
Score de conduite	**30-40**		68,17	
Score de conduite	**40-50**		53,98	
Score de conduite	**50-60**			
Score de conduite	**60-70**		2,86	
Score de conduite	**70-80**		2,86	
Score de conduite	**80-90**		18,94	
Score de conduite	**90-100**		12,93	
Secteur d'activité	**Activités de services**	3,22	2,44	-24%
Secteur d'activité	**Activités financières**	5,10	2,32	-55%
Secteur d'activité	**Activités du bâtiment**	14,55	12,32	-15%
Secteur d'activité	**Activités liées à l'immobilier**	1,24	1,54	24%
Secteur d'activité	**Activités de santé**	0,00	0,01	
Secteur d'activité	**Activités de commerce de gros**			
Rapport poids-puissance	**0-15**	4,59	7,20	57%
Rapport poids-puissance	**15-22**			
Rapport poids-puissance	**>22**	12,37	5,22	-58%
Année	**Y6**	1,29	0,02	-98%
Année	**Y7**			
Année	**Y8**	67,28	59,71	-11%
Taille de la flotte	**1 d'exposition**	49,79	45,54	-9%
Taille de la flotte	**2-33 d'exposition**			
Taille de la flotte	**34-657 d'exposition**	6,57	5,80	-12%
Taille de la flotte	**>657 d'exposition**	4,96	3,21	-35%
Nombre de place dans le véhicule	**0-2 places**	32,31	13,21	-59%
Nombre de place dans le véhicule	**3 places**	47,24	43,19	-9%
Nombre de place dans le véhicule	**>4 places**			
Nombre de kilomètres au contrat	**0-9000 kilomètres par an**			
Nombre de kilomètres au contrat	**9000-19000 kilomètres par an**	13,39	12,04	-10%
Nombre de kilomètres au contrat	**19000-25000 kilomètres par an**	19,89	19,05	-4%
Nombre de kilomètres au contrat	**25000-40000 kilomètres par an**	21,50	20,63	-4%
Nombre de kilomètres au contrat	**40000-55000 kilomètres par an**	36,24	34,24	-6%
Nombre de kilomètres au contrat	**>55000 kilomètres par an**	37,23	34,96	-6%
Marque du véhicule	**Luxe 1**	8,88	9,96	12%
Marque du véhicule	**Luxe 2**	0,52	0,79	52%
Marque du véhicule	**Populaire 1**	12,99	12,25	-6%
Marque du véhicule	**Populaire 2**			
Province italienne	**BOLZANO**	13,27	12,76	-4%
Province italienne	**CAGLIARI**	0,86	0,76	-12%
Province italienne	**CATANIA**	27,36	23,79	-13%
Province italienne	**COSENZA**	17,44	18,25	5%
Province italienne	**FIRENZE**	44,78	39,71	-11%
Province italienne	**NAPOLI**	86,15	100,81	17%
Province italienne	**PALERMO**	0,33	0,58	76%
Province italienne	**ROMA**	22,35	19,11	-14%
Province italienne	**TERAMO**	29,33	31,14	6%
Province italienne	**TORINO**	0,01	0,00	-100%
Province italienne	**MILANO**			
Type de véhicule	**Van**	0,18	0,40	122%
Type de véhicule	**Car**			

FIGURE 6.11 – Ecart de Wald Chi-square entre GLM fréquence corporelle avec et sans score de conduite

Le score de conduite implique une diminution de l'information ajoutée par les autres variables.

On constate également que le score de conduite est une des variables qui apporte le plus d'informations à la modélisation. Le modèle avec le score de conduite est le modèle le plus précis.

Analysons désormais l'impact de l'ajout de cette variable sur le coût moyen corporel.

6.2 Modélisation du coût moyen corporel

Cette modélisation se fait au moyen d'un GLM utilisant la loi Gamma avec une loi lien logarithmique.

Cette modélisation est rendue délicate par la volatilité des données : la base utilisée recense 4 267 observations contre 236 322 en fréquence. La question de l'efficacité de la modélisation par rapport à une modélisation "flat" avec une unique valeur prédite sera donc étudiée.

Le coût des sinistres corporels étant parfois très important, les schémas d'assurance font souvent appel à des réassureurs.

Ces montants très élevés vont biaiser nos modélisations en surévaluant les variables explicatives induites dans ces sinistres. La moyenne des sinistres corporels est de 12 391 € avec un écart type à 75 804 : les observations seront donc très volatiles. Cette volatilité est expliquée par des sinistres corporels dont les montants sont excessivement élevés. Nous noterons notamment la présence de quatre sinistres dont les montants sont de plus d'un million d'euros.
Prenons l'exemple, d'un individu conduisant une Citroën qui aurait un sinistre corporel rendant une tierce personne définitivement paralysée. Le sinistre serait estimé à 600 000 €. Le modèle aurait alors une tendance à surévaluer le coût moyen des sinistres corporels des véhicules de la marque Citroën : le modèle serait biaisé.

Nous allons alors écrêter le montant des sinistres corporels.

6.2.1 Choix de l'écrêtement

Pour déterminer un seuil de modélisation, nous allons utiliser deux méthodes : l'estimateur de Hills et la moyenne des excès.

Estimateur de Hills

Notons $X_{j,n}$ la j-ième plus grande observation d'une nième sous-base des observations construites par un tirage aléatoire sans remise de la base initiale. L'estimateur de Hills est défini de la manière suivante :

$$\hat{\xi}_{k,n} = \frac{1}{k-1} \sum_{j=1}^{k-1} \ln \left(\frac{X_{j,n}}{X_{k,n}} \right), \text{ avec k un entier naturel}$$

Nous cherchons le montant de $X_{l,n}$ tel que pour tout k supérieur à l, l'estimateur de Hills est stable.

En appliquant cet estimateur sur notre base de données, nous obtenons les résultats suivants :

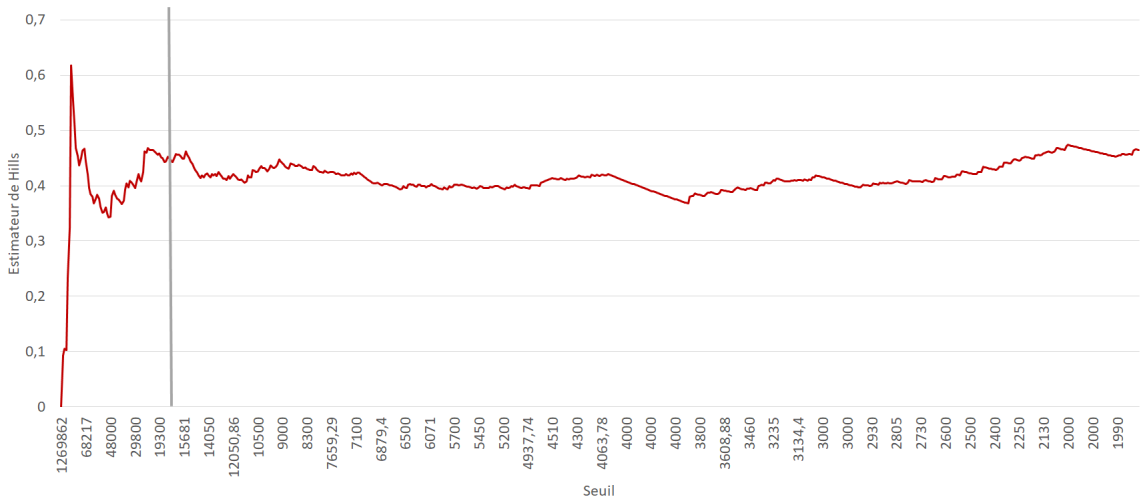

FIGURE 6.12 – Estimateur de Hills

Nous cherchons sur le graphique ci-dessus le montant seuil à partir duquel l'estimateur de Hills est stable. Ce seuil peut être placé entre 15000 et 22000 €. Afin de minimiser la partie aléatoire, nous placerons ce seuil à 17000 €.

Moyenne des excès

La fonction moyenne des excès permet de repérer les plages de linéarité et donc d'observer à partir de quel seuil les montants biaisent nos observations. Elle est définie par :

$$e(u) = \mathbb{E}[X - u | X > u] \tag{6.1}$$

avec u un certain seuil donné et X la variable aléatoire observée.

Nous cherchons à obtenir le seuil à partir duquel la fonction moyenne des excès est linéaire.

En appliquant cette fonction sur notre base de données, nous obtenons le graphique suivant :

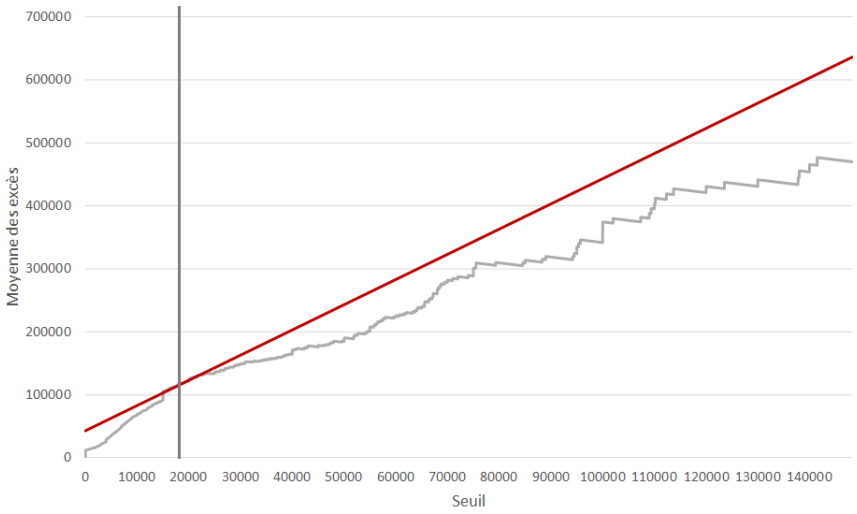

FIGURE 6.13 – Moyenne des excès

Nous cherchons le montant seuil à partir duquel la moyenne des excès est linéaire. Nous estimons ce seuil à 17 000 €.

Cette méthode confirme les résultats obtenus avec l'estimateur de Hills : le seuil conservé sera donc de 17 000 €.

Ceci permet de plafonner 299 sinistres, soit 14,27 % de notre base de données. Le nouveau coût moyen obtenu est de 4 530 € avec un écart type de 4 706. L'écart type des observations est donc bien réduit permettant de diminuer la volatilité.

Cette méthode induit une baisse du coût total des sinistres corporels de 23 270 460,16 € à 19 329 609,27 €. Cet écart de 3 940 850,89 € devra ensuite être reporté sur tous les contrats de manière équivalente.

6.2.2 Création des modalités

Nous avons testé l'ensemble des variables présentes dans la base contrat mais nous avons retenu les variables suivantes :

- l'année
- le secteur d'activité
- l'âge du véhicule en début de couverture
- le nombre de kilomètres au contrat
- la marque du véhicule
- la province

- la type de véhicule

L'année

Le nombre d'observations par année de contrat et leur relativité sont les suivantes :

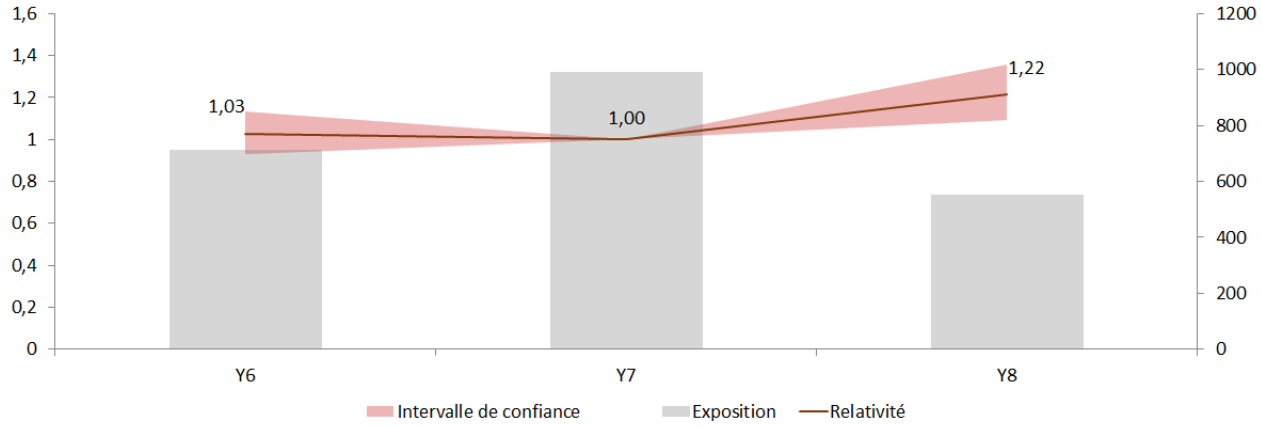

FIGURE 6.14 – Fréquence et relativité par année contrat

Le secteur d'activité

Cette variable a les mêmes modalités que sur les modélisations précédentes.

L'âge du véhicule en début de couverture

L'âge en début de couverture comporte deux modalités : entre 0 et 2 ans, et 3 ans et plus.

Le nombre de kilomètres au contrat

Nous avons créé deux modalités :

- petits conducteurs : de 0 à 9000 kilomètres,
- les autres conducteurs (plus de 9 000 kilomètres) pour lesquels le risque est homogène.

La marque du véhicule

Pour la variable "marque de véhicule" nous avons opéré les regroupements suivants :

Luxe 1	Luxe 2	Populaire 1	Populaire 2
Abarth, Audi, Bmw, Great Wall, Jaguar, Jeep, Lancia, Porsche, Tesla	Land Rover, Infiniti, Iveco, Lexus, Maserati, Mercedes, Mini Cooper, Smart, Piaggio	Fiat, Ford, Fuso Canter, Hyundai, Kia, Opel, Peugeot, Renault, Renault Trucks, Seat, Ssangyong, Subaru, Suzuki	Alfa Romeo, Citroen, Dacia, Ds, Honda, Mazda, Mitsubishi, Nissan, Skoda, Toyota, Volvo,

FIGURE 6.15 – Modalités de la variable "marque du véhicule"

La province

Observons les différentes relativités selon la province italienne.

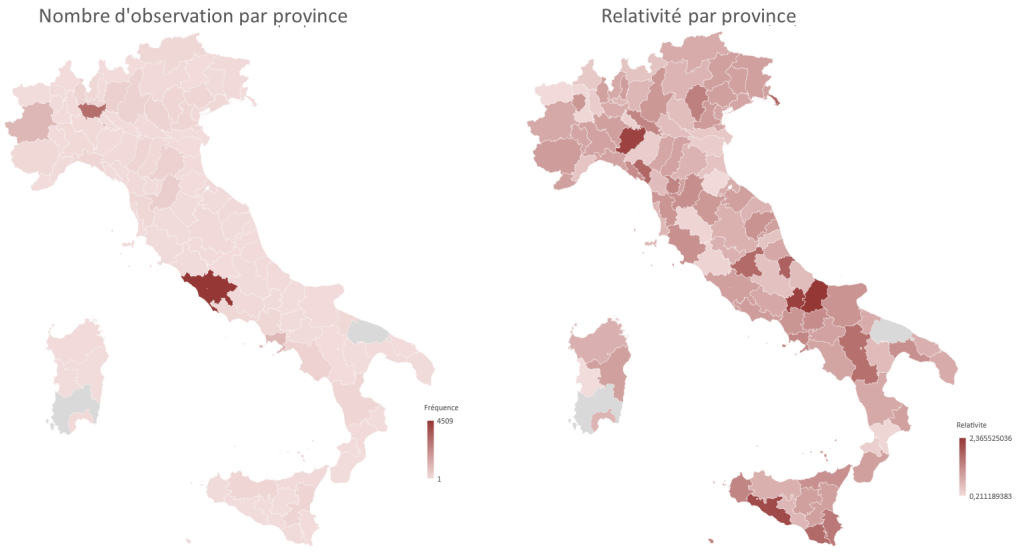

FIGURE 6.16 – Nombre d'observations et relativité par province

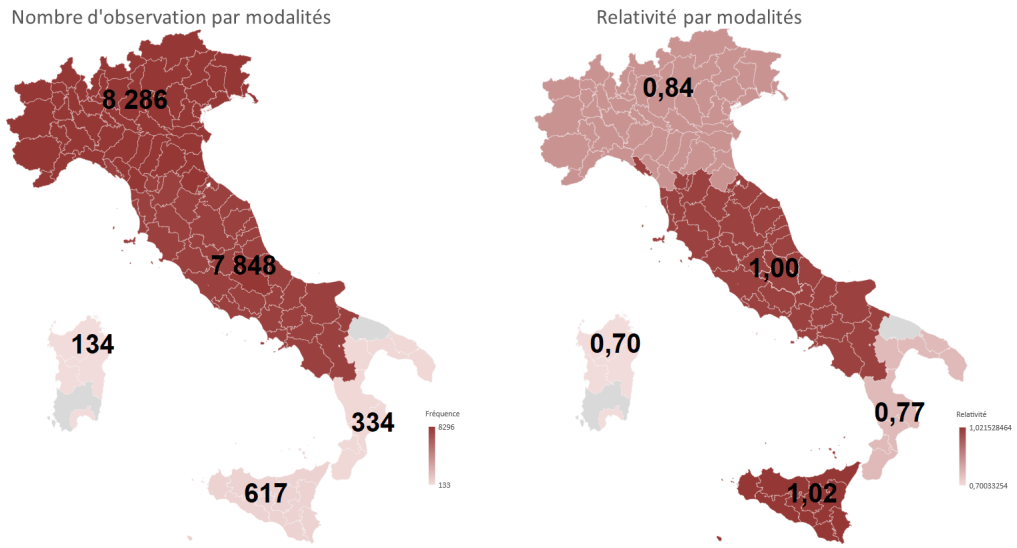

FIGURE 6.17 – Nombre d'observations et relativité par modalité

Nous utiliserons ces classes dans le GLM.

Le type de véhicule

Nous avons conservé les modalités "car" et "van".

6.2.3 Mesure de l'ajout du score de conduite

Le but de cette section est d'étudier le modèle GLM avec les variables explicatives présentées précédemment, puis d'y ajouter la variable "score de conduite".

Modèle GLM sans le score de conduite

Le GLM de coût moyen corporel sans score de conduite est le suivant :

Paramètre		Valeur estimée	Wald Chi-Square	Pr > Khi-2	Relativité
Secteur d'activité	**Activités de services**	-0,17	1	0,38	**0,85**
Secteur d'activité	**Activités financières**	-0,03	0	0,57	**0,97**
Secteur d'activité	**Activités du bâtiment**	0,02	0	0,72	**1,02**
Secteur d'activité	**Activités liées à l'immobilier**	-0,06	0	0,74	**0,94**
Secteur d'activité	**Activités de santé**	0,13	1	0,47	**1,14**
Secteur d'activité	**Activités de commerce de gros**	0,00			**1,00**
Année	**Y6**	0,02	0	0,73	**1,02**
Année	**Y7**	0,00			**1,00**
Année	**Y8**	0,19	13	0,00	**1,21**
Age du véhicule en début de couverture	**0-2 ans**	0,00			**1,00**
Age du véhicule en début de couverture	**>3 ans**	-0,01	0	0,89	**0,99**
Zone aéroportuaire	**0**	0,00			**1,00**
Nombre de kilomètres au contrat	**0-9000 kilomètres par an**	0,09	0	0,57	**1,09**
Nombre de kilomètres au contrat	**>9000 kilomètres par an**	0,00			**1,00**
Marque du véhicule	**Luxe 1**	-0,26	11	0,00	**0,77**
Marque du véhicule	**Luxe 2**	0,02	0	0,78	**1,02**
Marque du véhicule	**Populaire 1**	-0,19	11	0,00	**0,83**
Marque du véhicule	**Populaire 2**	0,00			**1,00**
Province italienne	**CAGLIARI**	-0,37	3	0,09	**0,69**
Province italienne	**COSENZA**	-0,30	5	0,02	**0,74**
Province italienne	**MILANO**	-0,13	9	0,00	**0,88**
Province italienne	**PALERMO**	0,01	0	0,90	**1,01**
Province italienne	**ROMA**	0,00			**1,00**
Type de véhicule	**Van**	-0,01	0	0,82	**0,99**
Type de véhicule	**Car**	0,00			**1,00**
Scale		1,09			

FIGURE 6.18 – Modèle GLM coût moyen corporel sans le score de conduite

Déviance	2 230
Log Vraissemblance	-20 056
AIC	40 151
BIC	40 258

FIGURE 6.19 – Mesures du modèle GLM coût moyen corporel sans le score de conduite

Malgré le cap effectué sur les données, peu de variables restent significatives.

Modèle GLM intégrant le score de conduite

En intégrant les tranches de score de conduite nous obtenons le modèle suivant :

Paramètre		Valeur estimée	Wald Chi-Square	Pr > Khi-2	Relativité
Score de conduite	**0-30**	0,23	3	0,07	**1,26**
Score de conduite	**30-40**	0,09	1	0,29	**1,10**
Score de conduite	**40-50**	0,10	2	0,14	**1,10**
Score de conduite	**50-60**	-0,03	0	0,59	**0,97**
Score de conduite	**60-70**	0,00			**1,00**
Score de conduite	**70-80**	0,08	2	0,21	**1,09**
Score de conduite	**80-90**	0,12	2	0,18	**1,13**
Score de conduite	**90-100**	-0,02	0	0,87	**0,98**
Secteur d'activité	**Activités de services**	-0,16	1	0,41	**0,85**
Secteur d'activité	**Activités financières**	-0,04	0	0,53	**0,97**
Secteur d'activité	**Activités du bâtiment**	0,02	0	0,70	**1,02**
Secteur d'activité	**Activités liées à l'immobilier**	-0,09	0	0,62	**0,91**
Secteur d'activité	**Activités de santé**	0,13	1	0,47	**1,14**
Secteur d'activité	**Activités de commerce de gros**	0,00			**1,00**
Année	**Y6**	0,03	0	0,60	**1,03**
Année	**Y7**	0,00			**1,00**
Année	**Y8**	0,20	13	0,00	**1,22**
Age du véhicule en début de couverture	**0-2 ans**	0,00			**1,00**
Age du véhicule en début de couverture	**>3 ans**	0,00	0	0,99	**1,00**
Zone aéroportuaire	**0**	0,00			**1,00**
Nombre de kilomètres au contrat	**0-9000 kilomètres par an**	0,09	0	0,57	**1,09**
Nombre de kilomètres au contrat	**>9000 kilomètres par an**	0,00			**1,00**
Marque du véhicule	**Luxe 1**	-0,24	10	0,00	**0,79**
Marque du véhicule	**Luxe 2**	0,02	0	0,80	**1,02**
Marque du véhicule	**Populaire 1**	-0,18	11	0,00	**0,83**
Marque du véhicule	**Populaire 2**	0,00			**1,00**
Province italienne	**CAGLIARI**	-0,35	3	0,11	**0,71**
Province italienne	**COSENZA**	-0,30	5	0,02	**0,74**
Province italienne	**MILANO**	-0,12	8	0,01	**0,88**
Province italienne	**PALERMO**	0,03	0	0,82	**1,03**
Province italienne	**ROMA**	0,00			**1,00**
Type de véhicule	**Van**	-0,02	0	0,68	**0,98**
Type de véhicule	**Car**	0,00			**1,00**
Scale		1,10			

FIGURE 6.20 – Modèle GLM coût moyen corporel avec le score de conduite

Déviance	2 221
Log Vraissemblance	-20 051
AIC	40 155
BIC	40 032

FIGURE 6.21 – Mesures du modèle GLM coût moyen corporel avec le score de conduite

Relativité de la variable "score de conduite" sur le coût moyen corporel

Le score de conduite ne semble pas très significatif ; analysons cependant son impact :

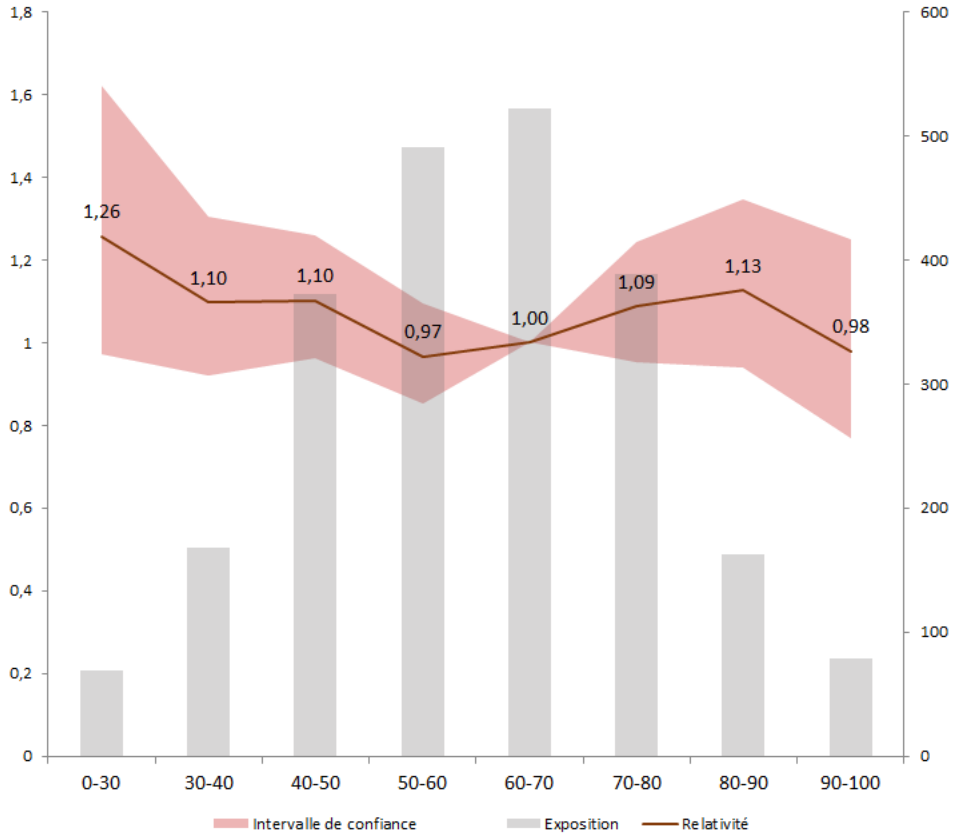

FIGURE 6.22 – Relativité score de conduite sur le coût moyen corporel

Peu de variables sont significatives dans notre modèle de coût moyen corporel et le score de conduite ne l'est pas non plus.

Par soucis de prudence dans nos modélisations, nous préférerons utiliser une modélisation flat, c'est à dire une modélisation qui a une unique valeur prédite. Le coût moyen corporel de tous les contrats est prédit dans ce cas à 12 391 €.

6.3 Conclusion de la modélisation GLM

Au vu des études menées, il apparaît que le score de conduite apporte de l'information sur les modèles GLM matériels (fréquence et coût moyen) et celui de la fréquence corporelle.

Dans le modèle de fréquence matérielle, le score de conduite permet de capter une partie de l'information déjà présente. Il apporte également une partie de l'information qui n'était pas présente dans la base de données (notamment du fait d'absence d'informations conducteurs).

Dans le modèle de coût moyen matériel, le score de conduite apporte autant d'informa-

tions que la moitié de l'information fournie par les autres variables.

Dans le modèle de fréquence corporelle, le score de conduite permet de capter une partie de l'information déjà présente dans presque toutes les variables. C'est la variable qui apporte le plus d'informations dans ce modèle.

Dans le modèle de coût moyen corporel, aucune variable ne se révèle être significative si bien que nous décidons de prédire le coût moyen corporel par la moyenne des coûts des sinistres corporels qui s'élève à 12 391 €.

Si l'on compare l'information de notre base de données à celle disponible lors d'une location retail, nous constatons qu'elle est faible : le score de conduite vient combler une partie de ce manque dans nos bases de données.

Les relativités des modèles GLM avec et sans score de conduite sont résumées ainsi :

Fréquence matérielle		
Score de conduite	**0-30**	2,29
Score de conduite	**30-40**	1,63
Score de conduite	**40-50**	1,37
Score de conduite	**50-60**	1,11
Score de conduite	**60-70**	1,00
Score de conduite	**70-80**	0,83
Score de conduite	**80-90**	0,73
Score de conduite	**90-100**	0,76
Secteur d'activité	**Activités de services**	0,88
Secteur d'activité	**Activités financières**	0,95
Secteur d'activité	**Activités du bâtiment**	0,89
Secteur d'activité	**Activités liées à l'immobilier**	1,17
Secteur d'activité	**Activités de santé**	1,32
Secteur d'activité	**Activités de commerce de gros**	1,00
Année	**Y6**	1,01
Année	**Y7**	1,00
Année	**Y8**	0,69
Taille de la flotte	**0-123 d'exposition**	1,00
Taille de la flotte	**124-200 d'exposition**	1,01
Taille de la flotte	**201-400 d'exposition**	1,19
Taille de la flotte	**>400 d'exposition**	0,93
Nombre de place dans le véhicule	**0-3 places**	1,56
Nombre de place dans le véhicule	**4 places**	1,00
Nombre de place dans le véhicule	**>5 places**	1,00
Age du véhicule en début de couverture	**0-2 ans**	1,00
Age du véhicule en début de couverture	**2-3 ans**	0,97
Age du véhicule en début de couverture	**>3 ans**	1,01
Nombre de kilomètres au contrat	**0-19000 kilomètres par an**	0,82
Nombre de kilomètres au contrat	**19000-44000 kilomètres par an**	1,00
Nombre de kilomètres au contrat	**>44000 kilomètres par an**	1,26
Puissance du véhicule	**0-73**	0,89
Puissance du véhicule	**73-146**	1,00
Puissance du véhicule	**146-203**	0,98
Puissance du véhicule	**203-+**	0,82
Poids du véhicule	**0-1966**	1,00
Poids du véhicule	**1966-2378**	1,02
Poids du véhicule	**2378-2665**	1,11
Poids du véhicule	**2665-3500**	1,76
Poids du véhicule	**3500-+**	4,48
Type de carburant du véhicule	**Electrique et hybride électrique**	0,92
Type de carburant du véhicule	**GPL et hybride GPL**	1,22
Type de carburant du véhicule	**Essence et diesel**	1,00
Marque du véhicule	**Francaise**	1,16
Marque du véhicule	**Luxe 1**	1,28
Marque du véhicule	**Luxe 2**	1,10
Marque du véhicule	**Populaires 1**	1,07
Marque du véhicule	**Populaires 2**	1,00
Province italienne	**BRESCIA**	0,86
Province italienne	**CAGLIARI**	1,01
Province italienne	**COSENZA**	1,09
Province italienne	**FIRENZE**	1,12
Province italienne	**NAPOLI**	1,78
Province italienne	**ROMA**	1,17
Province italienne	**MILANO**	1,00
Type de véhicule	**Van**	1,00
Type de véhicule	**Car**	1,00
Scale		

Coût moyen matériel		
Score de conduite	**0-30**	1,29
Score de conduite	**30-40**	1,13
Score de conduite	**40-50**	1,09
Score de conduite	**50-60**	1,01
Score de conduite	**60-70**	1,00
Score de conduite	**70-80**	0,94
Score de conduite	**80-90**	0,90
Score de conduite	**90-100**	0,87
Secteur d'activité	**Activités de services**	0,85
Secteur d'activité	**Activités financières**	0,91
Secteur d'activité	**Activités du bâtiment**	0,93
Secteur d'activité	**Activités liées à l'immobilier**	0,85
Secteur d'activité	**Activités de santé**	0,80
Secteur d'activité	**Activités de commerce de gros**	1,00
Année	**Y6**	1,00
Année	**Y8**	1,00
Année	**Y7**	1,07
Age du véhicule en début de couverture	**0-2 ans**	1,00
Age du véhicule en début de couverture	**3 ans**	1,08
Age du véhicule en début de couverture	**>4 ans**	1,45
Nombre de kilomètres au contrat	**0-9000 kilomètres par an**	1,00
Nombre de kilomètres au contrat	**9000-19000 kilomètres par an**	1,15
Nombre de kilomètres au contrat	**19000-55000 kilomètres par an**	1,12
Nombre de kilomètres au contrat	**>55000 kilomètres par an**	1,28
Puissance du véhicule	**0-71**	1,12
Puissance du véhicule	**>71**	1,00
Poids du véhicule	**0-1897**	1,00
Poids du véhicule	**>1897**	0,97
Marque du véhicule	**Luxe**	0,94
Marque du véhicule	**Populaire 1**	0,89
Marque du véhicule	**Populaire 2**	1,00
Province italienne	**CATANIA**	0,87
Province italienne	**MILANO**	0,90
Province italienne	**NAPOLI**	1,05
Province italienne	**TORINO**	1,06
Province italienne	**ROMA**	1,00
Type de véhicule	**Van**	1,03
Type de véhicule	**Car**	1,00
Franchise	**0 €**	0,89
Franchise	**150 €**	0,93
Franchise	**250 €**	1,00
Scale		

FIGURE 6.23 – Récapitulatif des relativités des modèles GLM matériels

Fréquence corporelle		
Score de conduite	**0-30**	3,42
Score de conduite	**30-40**	2,04
Score de conduite	**40-50**	1,64
Score de conduite	**50-60**	1,20
Score de conduite	**60-70**	1,00
Score de conduite	**70-80**	0,89
Score de conduite	**80-90**	0,67
Score de conduite	**90-100**	0,66
Secteur d'activité	**Activités de services**	0,74
Secteur d'activité	**Activités financières**	0,92
Secteur d'activité	**Activités du bâtiment**	0,81
Secteur d'activité	**Activités liées à l'immobilier**	1,27
Secteur d'activité	**Activités de santé**	1,02
Secteur d'activité	**Activités de commerce de gros**	1,00
Rapport poids-puissance	**0-15**	0,87
Rapport poids-puissance	**15-22**	1,00
Rapport poids-puissance	**>22**	0,82
Année	**Y6**	0,99
Année	**Y7**	1,00
Année	**Y8**	0,66
Taille de la flotte	**1 d'exposition**	1,70
Taille de la flotte	**2-33 d'exposition**	1,00
Taille de la flotte	**34-657 d'exposition**	1,14
Taille de la flotte	**>657 d'exposition**	0,83
Nombre de place dans le véhicule	**0-2 places**	1,37
Nombre de place dans le véhicule	**3 places**	2,06
Nombre de place dans le véhicule	**>4 places**	1,00
Nombre de kilomètres au contrat	**0-9000 kilomètres par an**	1,00
Nombre de kilomètres au contrat	**9000-19000 kilomètres par an**	1,88
Nombre de kilomètres au contrat	**19000-25000 kilomètres par an**	2,23
Nombre de kilomètres au contrat	**25000-40000 kilomètres par an**	2,32
Nombre de kilomètres au contrat	**40000-55000 kilomètres par an**	3,11
Nombre de kilomètres au contrat	**>55000 kilomètres par an**	3,43
Marque du véhicule	**Luxe 1**	0,73
Marque du véhicule	**Luxe 2**	0,95
Marque du véhicule	**Populaire 1**	0,77
Marque du véhicule	**Populaire 2**	1,00
Province italienne	**BOLZANO**	1,34
Province italienne	**CAGLIARI**	1,23
Province italienne	**CATANIA**	1,95
Province italienne	**COSENZA**	1,58
Province italienne	**FIRENZE**	1,81
Province italienne	**NAPOLI**	2,22
Province italienne	**PALERMO**	1,12
Province italienne	**ROMA**	1,27
Province italienne	**TERAMO**	1,91
Province italienne	**TORINO**	0,99
Province italienne	**MILANO**	1,00
Type de véhicule	**Van**	1,06
Type de véhicule	**Car**	1,00
Scale		

FIGURE 6.24 – Récapitulatif des relativités du modèle GLM de fréquence corporelle

A ce stade de notre étude, nous avons montré l'importance de la variable score de conduite dans l'explication de la sinistralité de notre portefeuille. Sachant que cette variable capte une information disponible à posteriori, nous allons étudier à présent la qualité de modélisation de la sinistralité par une modélisation Bonus-Malus.

Troisième partie

Modélisation Bonus-Malus

Nous avons vu précédemment l'apport de la variable "score de conduite" dans les modélisations GLM. Nous cherchons à présent une autre manière de modéliser la sinistralité en utilisant cette information du score de conduite. Cette dernière est une information "à posteriori", au même titre que l'historique de la sinistralité. La modélisation la plus courante utilisant de l'information "à posteriori" est la modélisation Bonus-Malus. Ainsi, dans cette partie, nous allons nous attacher à créer un Bonus-Malus qui dépendra du score de conduite.

Dans un premier temps, nous introduirons la notion de Bonus-Malus et sa création dans son cadre le plus général. Nous créerons ensuite une modélisation Bonus-Malus qui intégrera la variable "score de conduite", puis construirons un modèle Bonus-Malus dit "composé", dépendant de la sinistralité et du score de conduite. Enfin, nous comparerons ces différentes modélisations.

Chapitre 7

Modélisation du Bonus-Malus en fonction de la sinistralité

Nous allons, dans un premier temps, appliquer la théorie de Bonus-Malus de flotte automobile afin de créer une échelle Bonus-Malus qui dépende de l'historique de sinistres. Nous la comparerons ensuite aux modélisations Bonus-Malus qui utilisent le score de conduite.

7.1 Théorie du Bonus-Malus en flotte automobile

Le Bonus-Malus est une modélisation à posteriori dans laquelle l'assuré voit sa prime évoluer à l'issue d'une période en fonction des évènements survenus pendant cette période. En France, les règles d'évolution de la prime pour les particuliers sont encadrées de la façon suivante :

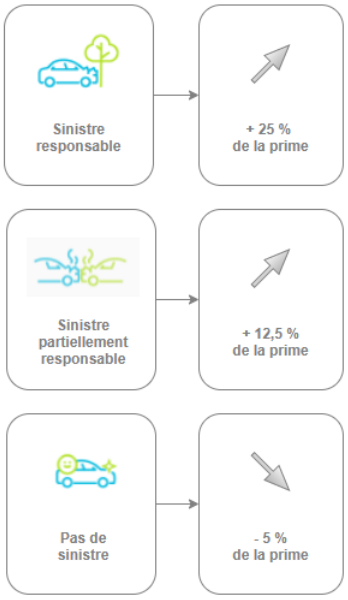

FIGURE 7.1 – Fonctionnement du Bonus-Malus automobile

Le Bonus-Malus permet, en réalité, d'augmenter les primes des conducteurs qui présentent une fréquence importante d'accidents et, inversement, de diminuer progressivement celles des conducteurs n'ayant pas d'accident. Cette modélisation revêt donc un aspect pédagogique en responsabilisant les conducteurs sur leur conduite et leur éventuelle sinistralité.

Les modélisations GLM et celle Bonus-Malus sont deux types de modélisations différentes et, d'une certaine manière, elles se complètent : l'une utilise les informations présentes à la souscription du contrat, alors que l'autre utilise rétrospectivement des informations récoltées pendant la période de couverture.

7.1.1 Explication théorique de la construction du Bonus-Malus

Le Bonus-Malus permet d'introduire l'historique des sinistres dans la modélisation. En effet, après une période de distribution, l'assureur dispose des informations sur la sinistralité de ses assurés. Ainsi, au bout d'un certain nombre de périodes d'observation, la fréquence observée se rapproche de la fréquence réelle.

Par exemple, prenons un conducteur observé j ayant une fréquence prédite par un GLM initial de 12 %. Supposons qu'il passe quotidiennement par des routes moins dangereuses que la moyenne, il obtient alors une fréquence de sinistres réelle N_j inférieure à 12%. Le Bonus-Malus sert à capter cette information et permet de faire baisser sa fréquence modélisée pour se rapprocher de sa fréquence réelle.

En observant sa fréquence de l'année 1 notée N_1^j jusqu'à celle de l'année I notée N_I^j par rapport à celles du portefeuille $(N_1\ ,\ N_I)$, on peut alors affiner l'estimation de la charge sinistre de l'individu notée S^j.

On appelle fréquence crédibilisée la fréquence obtenue après la modélisation Bonus-Malus. Elle dépend des fréquences observées et modélisées.

Hypothèses

Dans cette étude, l'historique de sinistre permettra uniquement une modélisation de la fréquence. Nous nous trouvons ainsi dans l'hypothèse de Bischel :

Hypothèse de Bischel. *Pour tout profil de risque θ et année j ,*

$$\mathbb{E}[S_j|(\Theta = \theta)] = P \times \mathbb{E}[N_j|(\Theta = \theta)]$$

où

- Θ *la variable aléatoire correspondant au profil de risque*
- P *une constante dépendant uniquement de la puissance du véhicule*

- $\mathbb{E}[S_j|\Theta = \theta]$ *l'espérance de la charge de sinistre du profil de risque* θ *pour l'année* j
- $\mathbb{E}[N_j|\Theta = \theta]$ *l'espérance de la fréquence de sinistres du profil de risque* θ *pour l'année* j*, elle dépend uniquement du conducteur*

Dans un premier temps, nous nous placerons dans le cadre d'une police d'assurance (notée j) ayant un profil de risque indépendant des années, si bien que son exposition reste constante et unitaire sur chaque période (une période, étant égale à une année).

Le profil de risque j est observé sur I années et nous souhaitons prédire la charge sinistre de l'année I+1, c'est à dire $\mathbb{E}[S_{I+1}]$.

Base théorique

Par définition du modèle collectif présenté dans la partie 4.2.2, la charge sinistre moyenne de l'année I+1 est égale à la fréquence de sinistres moyenne de l'année I+1 ($\mathbb{E}[N_{I+1}]$) multipliée par le coût moyen de l'année I+1 ($\mathbb{E}[X_{I+1}]$).

On peut le traduire mathématiquement :

$$\mathbb{E}[S_{I+1}] = \mathbb{E}[N_{I+1}] \times \mathbb{E}[X_{I+1}] \tag{7.1}$$

Nous supposons le coût moyen des sinistres indépendant des années :

$$\mathbb{E}[S_{I+1}|(N_1 = n_1, \cdots, N_I = n_I)] = \mathbb{E}[N_{I+1}|(N_1 = n_1, \cdots, N_I = n_I)] \times \mathbb{E}[X_{I+1}] \tag{7.2}$$

Posons une variable de structure Λ de fonction de densité continue h. Nous pouvons supposer que pour toute année i, la fréquence de sinistres N_i suit une loi de Poisson de paramètre Λ. Ce qui nous conduit à

$$\mathbb{E}[N_i] = \mathbb{E}[\Lambda] \qquad \text{et} \qquad \mathbb{V}(N_i) = \mathbb{E}[\Lambda] + \mathbb{V}(\Lambda) \tag{7.3}$$

On peut également supposer que les variables aléatoires $N_1|\Lambda, \cdots, N_I|\Lambda$ sont mutuellement indépendantes et de loi de Poisson de paramètre Λ. D'où :

$$\mathbb{E}[N_i|\Lambda] = \mathbb{E}[\Lambda] \tag{7.4}$$

Coefficient de Réduction-Majoration

On pose la variable C_{I+1}, appelée coefficient de réduction-majoration (CRM) de l'année I+1 telle que :

$$C_{I+1} = 100 \times \frac{\mathbb{E}[S_{I+1}|(N_1, \cdots, N_I)]}{\mathbb{E}[S_{I+1}]}$$

Nous pouvons donc définir la charge sinistre de l'année I+1 sachant l'historique de sinistre de la manière suivante :

$$\boxed{\mathbb{E}[S_{I+1}|(N_1,\cdots,N_I)] = \frac{C_{I+1}\times\mathbb{E}[S_{I+1}]}{100}}$$

Nous montrons ensuite que

$$\begin{aligned}
C_{I+1} &= 100\times\frac{\mathbb{E}[S_{I+1}|(N_1,\cdots,N_I)]}{\mathbb{E}[S_{I+1}]} \\
&= 100\times\frac{\mathbb{E}[N_{I+1}|(N_1,\cdots,N_I)]\times\mathbb{E}[X_{I+1}]}{\mathbb{E}[N_{I+1}]\times\mathbb{E}[X_{I+1}]} && \text{d'après (7.1) et (7.2)} \\
&= 100\times\frac{\mathbb{E}[N_{I+1}|(N_1,\cdots,N_I)]}{\mathbb{E}[N_{I+1}]} \\
&= 100\times\frac{\mathbb{E}[\Lambda|(N_1,\cdots,N_I)]}{\mathbb{E}[\Lambda]} && \text{d'après (7.3) et (7.4)}
\end{aligned}$$

Dans cette nouvelle équation, nous déplaçons le problème d'étude sur la variable de structure.

Loi de la variable de structure

Notons $h^{N_1=n_1,\cdots,N_I=n_I}$ la fonction de densité continue de Λ sachant les fréquences des sinistres des années 1 à I respectivement égaux à n_1, ... n_I, et rappelons que h est la fonction de densité continue de Λ.

D'après le théorème de Bayes :

$$h^{N_1=n_1,\cdots,N_I=n_I}(\lambda) = \frac{h(\lambda)\times P(N_1=n_1,\cdots,N_I=n_I|\Lambda=\lambda)}{\int_0^{+\infty} P\left(N_1=n_1,\ldots,N_I=n_I|\Lambda=\lambda\right)\times h(\lambda)d\lambda}$$

Rappelons que pour tout i $\in [1;\mathrm{I}], N_i|\Lambda$ suit une loi de Poisson de paramètre Λ. Par indépendance, nous avons :

$$\begin{aligned}
P(N_1=n_1,\cdots,N_I=n_I|\Lambda=\lambda) &= \prod_{i=1}^{I} P\left(N_i=n_i|\Lambda=\lambda\right) \\
&= \prod_{i=1}^{I}\frac{\lambda^{n_i}e^{-\lambda}}{n_i!} \\
&= \frac{e^{-\lambda I}\lambda^{\sum_{i=1}^{I}n_i}}{\prod_{i=1}^{I}n_i!}
\end{aligned}$$

donc

$$h^{N_1=n_1,\cdots,N_I=n_I}(\lambda) = \frac{h(\lambda)e^{-\lambda I}\lambda^{\sum_{i=1}^{I}n_i}}{\int_0^{+\infty}e^{-\lambda I}\lambda^{\sum_{i=1}^{I}n_i}h(\lambda)d\lambda}$$

On voit que $h^{N_1=n_1,\cdots,N_I=n_I}$ ne dépend que de la somme des n_i et donc de la sinistralité passée. On en conclue qu'il en est de même pour le CRM.

L'objectif est de modéliser la loi de Λ. Pour ce faire, une loi est souvent utilisée : la loi Gamma.

Cas particulier de la loi Gamma

On suppose que Λ suit une loi Gamma de paramètre r $\in \mathrm{R}^+$ et $\beta \in \mathrm{R}^+$.

Nous définissons la loi Gamma par la fonction de densité suivante :

$$f(x) = x^{r-1}\frac{\beta^r\,\mathrm{e}^{-\beta\,x}}{\Gamma(r)} \text{ pour } x > 0.$$

avec Γ la fonction gamma $\Gamma : z \mapsto \int_0^{+\infty} t^{z-1}\,\mathrm{e}^{-t}\,\mathrm{d}t$ pour z>0.

Par définition :

$$\mathbb{E}[\Lambda] = \frac{r}{\beta} \text{ et } \mathbb{V}(\Lambda) = \frac{r}{\beta^2}$$

Ainsi nous obtenons :

$$\begin{aligned} h^{N_1=n_1,\cdots,N_I=n_I}(\lambda) &= \frac{\frac{\lambda^{r-1}\mathrm{e}^{-\lambda\beta}\beta^r}{\Gamma(r)}e^{-\lambda I}\lambda^{\sum_{i=1}^I n_i}}{\int_0^{+\infty} e^{-\lambda I}\lambda^{\sum_{i=1}^I n_i}\frac{\lambda^{r-1}\mathrm{e}^{-\lambda\beta}\beta^r}{\Gamma(r)}d\lambda} \\ &= \mathrm{e}^{-\lambda(I+\beta)}\lambda^{r-1+\sum_{i=1}^I n_i} \times \frac{\frac{\beta^r}{\Gamma(r)}}{\int_0^{+\infty} e^{-\lambda I}\lambda^{\sum_{i=1}^I n_i}\frac{\lambda^{r-1}\mathrm{e}^{-\lambda\beta}\beta^r}{\Gamma(r)}d\lambda} \\ &\propto \mathrm{e}^{-\lambda(I+\beta)}\lambda^{r-1+\sum_{i=1}^I n_i} \end{aligned}$$

avec $\propto$ désignant un effet de proportionnalité.

On remarque que $h^{N_1=n_1,\cdots,N_I=n_I}(\lambda)$ est proportionnel à une densité de loi Gamma de paramètres $r+\sum_{i=1}^I n_i$ et $I+\beta$. On en conclut que $\Lambda|(N_1=n_1,\cdots,N_I=n_I)$ suit la loi Gamma de paramètres $r+\sum_{i=1}^I n_i$ et $I+\beta$.

D'où

$$\mathbb{E}[\Lambda|(N_1=n_1,\cdots,N_I=n_I)] = \frac{r+\sum_{i=1}^I n_i}{I+\beta}$$

On a donc

$$\begin{aligned} C_{I+1} &= 100 \times \frac{\mathbb{E}[\Lambda|(N_1, \cdots, N_I)]}{\mathbb{E}[\Lambda]} \\ &= 100 \times \frac{\frac{(r+\sum_{i=1}^{I} N_i)}{(I+\beta)}}{\frac{r}{\beta}} \\ &= 100 \times \frac{1 + \frac{\sum_{i=1}^{I} N_i}{r}}{1 + \frac{I}{\beta}} \end{aligned}$$

Si on se place à l'année i dans le cadre d'un portefeuille contenant L_i polices ayant un profil de risque similaire, avec une variable de structure qui suit la loi Gamma de paramètres r et β, alors la variable de structure de la flotte suit la loi Gamma de paramètres $\mathbb{E}[L] \times r$ et β.

7.1.2 Facteur de crédibilité

La théorie de la crédibilité donne des éléments de réponse plus précis sur le degré d'importance à donner à l'historique de chaque police. En pratique, en effet, on ne peut pas donner la même importance à l'historique de sinistralité d'un véhicule unique qu'à une flotte de plusieurs centaines de véhicules.

Dans ce cas, il est essentiel de se pencher sur l'importance donnée à l'historique. On parle ici du principe de la crédibilité.

Cette crédibilité se traduit par le facteur de crédibilité Z^j, pour une police j :

$$S^j_{I+1} = Z^j \times \mathbb{E}[S^j_{I+1}|(N^j_1, \cdots, N^j_I)] + (1 - Z^j) \times \mathbb{E}[S] \tag{7.5}$$

avec $\mathbb{E}[S]$ la charge sinistre moyenne du portefeuille sur toutes les années.

Deux cas de figure peuvent être envisagés :

- **Z=1** on parle de crédibilité totale. L'historique de sinistre suffit à lui seul pour permettre une modélisation complète.
- $\mathbf{Z} \in [0, 1[$ on parle alors de crédibilité partielle. L'historique ne suffit pas, un poids est donné à la charge sinistre moyenne du portefeuille.

Attachons-nous à déterminer une expression explicite du facteur de crédibilité.

D'après (7.5) nous avons :

$$Cov(S^j_{I+1}, (S^j_{I+1}|(N^j_1, \cdots, N^j_I))) = Z^j \times Cov((S^j_{I+1}|(N^j_1, \cdots, N^j_I)), (S^j_{I+1}|(N^j_1, \cdots, N^j_I)))$$

D'après le théorème de la covariance totale :

$$Z^j = \frac{E[Cov(S^j_{I+1}, (S^j_{I+1}|(N^j_1, \cdots, N^j_I))|\Theta)] + Cov(E[S^j_{I+1}|\Theta], E[S^j_{I+1}|(N^j_1, \cdots, N^j_I)|\Theta])}{\mathbb{V}(S^j_{I+1}|(N^j_1, \cdots, N^j_I))}$$
$$= \frac{0 + \mathbb{V}((S^j_{I+1}|(N^j_1, \cdots, N^j_I))|\Theta)}{\mathbb{V}(S^j_{I+1}|(N^j_1, \cdots, N^j_I))}$$

On a donc :

$$\boxed{Z^j = \frac{\mathbb{V}((S^j_{I+1}|(N^j_1, \cdots, N^j_I))|\Theta)}{\mathbb{V}(S^j_{I+1}|(N^j_1, \cdots, N^j_I))}}$$

Dans le cas où Λ suit une loi Gamma, avec W^j l'exposition totale de la police j sur toutes les années observées, on a :

$$\boxed{Z^j = \frac{W^j}{W^j + \frac{\mathbb{V}(S^j_{I+1}|(N^j_1, \cdots, N^j_I))}{\mathbb{V}(\Theta) - \mathbb{V}(S^j_{I+1}|(N^j_1, \cdots, N^j_I))/I}}}$$

On notera également :

$$Z^j = \frac{W^j}{W^j + K}$$

avec K le "coefficient de crédibilité", donné par $K = \frac{\mathbb{V}(S^j_{I+1}|(N^j_1, \cdots, N^j_I))}{\mathbb{V}(\Theta) - \mathbb{V}(S^j_{I+1}|(N^j_1, \cdots, N^j_I))/I}$.

7.1.3 Modèle de Bühlmann-Straub

Dans notre étude, nous utiliserons le modèle de Bühlmann-Straub, qui généralise la formule du coefficient de crédibilité, sur une flotte de véhicules constituée de L clients, chaque client étant une police (noté j) ayant une exposition totale W^j. L'historique étudié sera de I années.

La fréquence de sinistres crédibilisée de l'année I+1 étant :

$$N^j_{I+1} = Z^j \hat{N}^j_{I+1} + \left(1 - Z^j\right) \mu_0 \tag{7.6}$$

où

- $\hat{N}^j_{I+1} = \frac{\sum_{k=1}^I W^j_k \times N^j_k}{\sum_{k=1}^I W^j_k}$
- W^j_i l'exposition de la police j sur l'année i
- $\mu_0 = \mathbb{E}[N]$ la fréquence moyenne de sinistre observée sur toutes les périodes
- $Z^j = \frac{W^j}{W^j + K}$, avec K=$\frac{\sigma^2}{\tau^2}$

Le modèle donne également des estimateurs pour σ et τ. Nous pouvons estimer σ, une mesure du risque interne au risque individuel par :

$$\widehat{\sigma^2} := \frac{1}{I(L-1)} \sum_{i=1}^{I} \sum_{j=1}^{L} W_i^j \left(N_i^j - N_i\right)^2$$

et nous pouvons estimer τ, une mesure de l'hétérogénéité du portefeuille par :

$$\hat{\tau}^2 := \frac{W}{W^2 - \sum_{i=1}^{I} W_i^2} \left\{ \sum_{i=1}^{I} W_i \left(N_i - \mathbb{E}[N]\right)^2 - (I-1)\widehat{\sigma^2} \right\}$$

avec W l'exposition totale du portefeuille et W_i l'exposition totale de l'année i.

7.1.4 Application aux flottes de véhicules

Appliquons le modèle de Bühlmann-Straub sur un cas simple : un portefeuille composé de trois flottes automobiles avec un historique de trois années.

Prenons le portefeuille suivant :

Exposition	Année 1	Année 2	Année 3	Total
Client 1	10	20	15	45
Client 2	150	175	200	525
Client 3	1200	1300	1500	4000
Total	1360	1495	1715	4570

FIGURE 7.2 – Exemple de portefeuille d'application du modèle de Bühlmann-Straub

Considérons la sinistralité suivante :

Nombre de sinistres	Année 1	Année 2	Année 3	Total
Client 1	1	3	3	7
Client 2	9	20	18	47
Client 3	250	190	180	620
Total	260	213	201	674

FIGURE 7.3 – Sinistralité du portefeuille d'application du modèle Bühlmann-Straub

Nous obtenons ainsi la fréquence de sinistres observée suivante :

Fréquence de sinistres	Année 1	Année 2	Année 3	Total
Client 1	10%	15%	20%	16%
Client 2	6%	11%	9%	9%
Client 3	21%	15%	12%	16%
Total	19%	14%	12%	15%

FIGURE 7.4 – Fréquence de sinistres observée du portefeuille d'application du modèle Bühlmann-Straub

Calculons les estimateurs $\widehat{\sigma^2}$ et $\widehat{\tau}$:

$$\widehat{\sigma^2} := \frac{1}{3(3-1)} \sum_{i=1}^{3} \sum_{j=1}^{3} W_i^j \left(N_i^j - N_i\right)^2 = 0,57$$

$$\hat{\tau}^2 := \frac{4575}{4575^2 - \sum_{i=1}^{3} W_i^2} \left(\sum_{i=1}^{3} W_i \left(N_i - 15\%\right)^2 - (3-1)\right) = 1,01E-03$$

On a ainsi K=$\frac{\sigma^2}{\tau^2}$=568.

On obtient alors les résultats suivants :

	Z^j	N_j^{I+1}	C_{I+1}
Client 1	7%	15%	100 %
Client 2	48%	12%	81 %
Client 3	88%	15%	104 %

FIGURE 7.5 – Principaux résultats de l'exemple d'application du modèle de Bühlmann-Straub

La fréquence crédibilisée donne ainsi la fréquence réelle modélisée pour chaque client. Ainsi, pour le client 2, la fréquence modélisée est de 100 % - 81 % = 19% inférieure à la fréquence moyenne du portefeuille. On pourra donc appliquer à la prime pure de ce client une réduction tarifaire de 19%. En raisonnant de manière similaire, on pourra appliquer une majoration tarifaire de 4% à la prime pure du client 3.

Pour la suite, nous utiliserons ce modèle de Bühlmann-Straub sur nos données afin d'obtenir une échelle de Bonus-Malus fonction de la sinistralité, et nous la comparerons avec celle obtenue en fonction du score de conduite.

7.2 Création d'une échelle Bonus-Malus évoluant en fonction de la sinistralité

Nous commencerons par appliquer le modèle de Bühlmann-Straub sur 2 années de contrat, puis nous construirons une échelle Bonus-Malus. Enfin, nous étudierons la pertinence de cette échelle et l'améliorerons le cas échéant.

7.2.1 Application de la théorie

La pandémie liée à la Covid 19, par l'intermédiaire des mesures mises en place par le gouvernement italien pour y faire face, a impacté de façon sensible l'année Y8, comme en témoigne le graphique des relativités suivant :

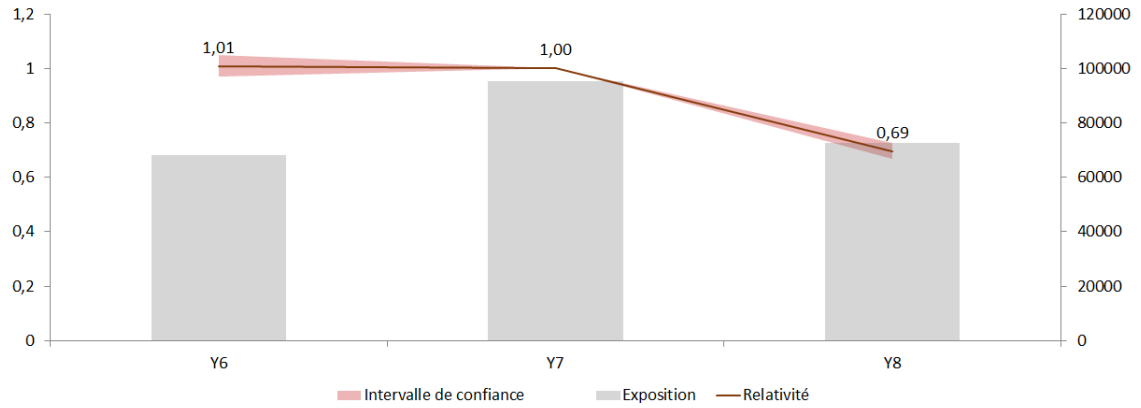

FIGURE 7.6 – Fréquence et relativité par année contrat

Nous avons donc décidé de calibrer notre modèle sur les années Y6 et Y7.

Nous appliquons le modèle de Bühlmann-Straub sur nos données en nous restreignant aux clients ayant plus de dix années de contrat sur Y6 et Y7.

Le modèle de Bühlmann-Straub donne les indicateurs suivants :

σ	τ	K
0,13	3,99E-03	31,36

FIGURE 7.7 – Indicateurs du modèle de Bühlmann-Straub

Nous pouvons observer la boîte à moustaches du coefficient CRM obtenu :

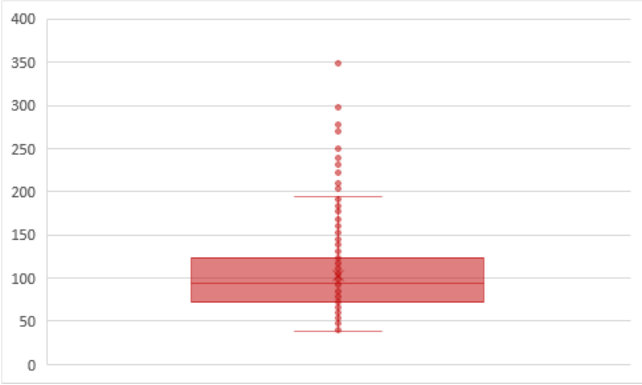

FIGURE 7.8 – Boîte à moustaches CRM Bonus-Malus en fonction de la sinistralité

Le premier quartile étant inférieur à 50 et le troisième quartile proche de 200, on constate une très grande variation de fréquence. On peut en déduire que 25% des clients vont voir leur prime doubler. A contrario, 25% des clients vont voir leur prime divisée par deux.

7.2.2 Echelle Bonus-Malus

Construisons à présent une échelle à partir de ce CRM. Nous commençons par regrouper les CRM obtenus précédemment en différentes classes en appliquant une classification ascendante hiérarchique. Cette classification a pour but d'obtenir un nombre limité de classes, moins de quinze, tout en minimisant les écarts de modélisation dus aux regroupements. Concrètement, nous cherchons à obtenir une fréquence crédibilisée par groupes de clients en utilisant celle calculée pour chaque client.

Pour ce faire, nous avons utilisé la procédure "cluster" du logiciel SAS. Afin d'optimiser le nombre de classes, nous pouvons observer la statistique du pseudo t au carré.

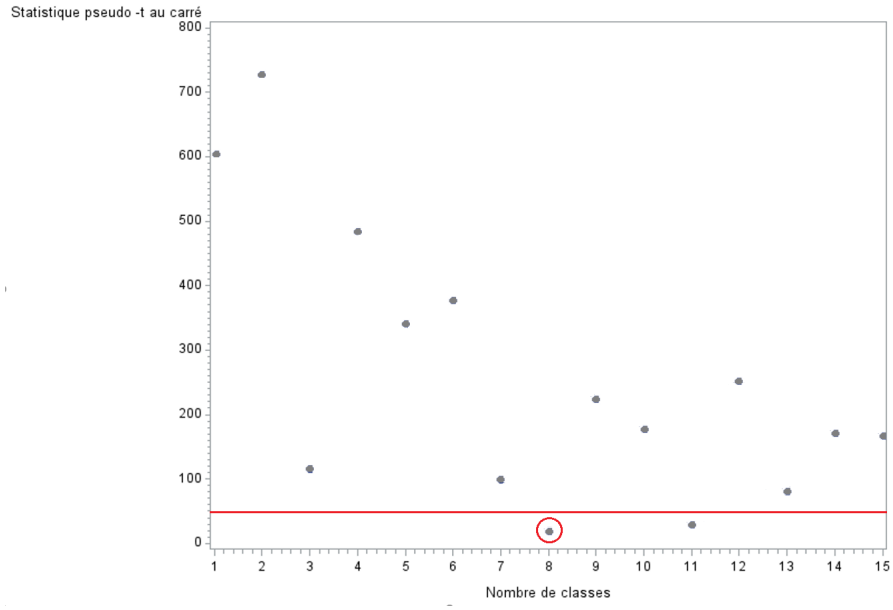

FIGURE 7.9 – Pseudo t au carré des CRM en fonction de la sinistralité

Ce graphique montre que 8 ou 11 classes permettent de minimiser au mieux les écarts. Afin de rester le plus précis possible, nous gardons celui avec le plus petit pseudo t soit huit classes. Les statistiques sur ces classes sont les suivantes :

Classe CRM	Fréquence Maximum	CRM Minimum	CRM Moyen	CRM Maximum	CRM Ecart type
1	7,0%	39	58	71	8
2	8,9%	71	77	83	3
3	11,5%	83	93	103	6
4	15,9%	104	112	120	5
5	24,7%	121	134	156	10
6	32,1%	160	175	194	10
7	45,5%	204	214	231	10
8	68,8%	240	271	349	31

FIGURE 7.10 – Statistiques des classes de CRM pour la modélisation Bonus-malus traditionnelle

Ces classes, triées dans l'ordre croissant de CRM, nous permettent de créer une échelle Bonus-Malus, avec, pour chacune , un intervalle de fréquence et un CRM associé.

Par prudence, nous choisissons de garder le CRM maximal de chaque classe. Nous pouvons ainsi construire une échelle Bonus-Malus en considérant l'évolution tarifaire d'une année sur

l'autre de la manière suivante :

$$\text{"évolution tarifaire"} = (\text{CRM}_{max} - 100)/100$$

Nous obtenons l'échelle suivante :

Fréquence	CRM Maximum	Evolution tarifaire	Ecart d'évolution entre classes
0% - 7,0%	71	-29%	
7,0% - 8,9%	83	-17%	12%
8,9% - 11,5%	103	+ 3%	20%
11,5% - 15,9%	120	+ 20%	17%
15,9% - 24,7%	156	+ 56%	36%
24,7% - 32,1%	194	+ 94%	38%
32,1% - 45,5%	231	+ 131%	37%
45,5% - 68,8%	349	+ 249%	118%

FIGURE 7.11 – Echelle en fonction de la sinistralité

Ces résultats confirment la volatilité observée avec la boîte à moustaches : l'étendue de l'échelle est [71 ; 349]. En pratique, il n'est pas possible d'augmenter la prime de 249% ou de la diminuer de 29%.

Afin de réduire cet écart, nous créons une deuxième échelle en réduisant l'étendue à [90 ; 160]. Pour ce faire, nous appliquons à chaque étage la fonction suivante :

$$y = \frac{x - 71}{349 - 71} \times (160 - 90) + 90$$

avec y le CRM obtenu après modification et x le CRM initial.

Nous obtenons cette nouvelle échelle :

Fréquence	CRM Maximum	Evolution tarifaire	Ecart d'évolution entre classes
0% - 7,0%	90	-10%	
7,0% - 8,9%	93	-7%	3%
8,9% - 11,5%	98	-2%	5%
11,5% - 15,9%	103	+ 3%	5%
15,9% - 24,7%	112	+ 12%	9%
24,7% - 32,1%	121	+ 21%	9%
32,1% - 45,5%	130	+ 30%	9%
45,5% - 68,8%	160	+ 60%	30%

FIGURE 7.12 – Echelle en fonction de la sinistralité, seconde version

Nous obtenons ainsi une échelle Bonus-Malus réaliste et exploitable qui dépend de la fréquence de sinistre.

Nous allons désormais construire la même échelle en intégrant le score de conduite, puis nous les comparerons.

Chapitre 8

Modélisation Bonus-Malus évoluant en fonction du score de conduite

Nous commencerons par adapter la variable "score de conduite" à l'utilisation du modèle de Bühlmann-Straub, puis nous appliquerons ce modèle pour construire l'échelle Bonus-Malus.

8.1 Transformation du score de conduite en fréquence de sinistres

La modélisation Bonus-Malus repose sur un équilibre entre la fréquence historique et la fréquence moyenne de la flotte. Le score de conduite étant une variable sans formule théorique, il n'est pas possible d'en déduire un lien direct avec la fréquence de sinistres.

On reprend le graphique de l'impact du score de conduite des modélisations GLM sur la fréquence matérielle :

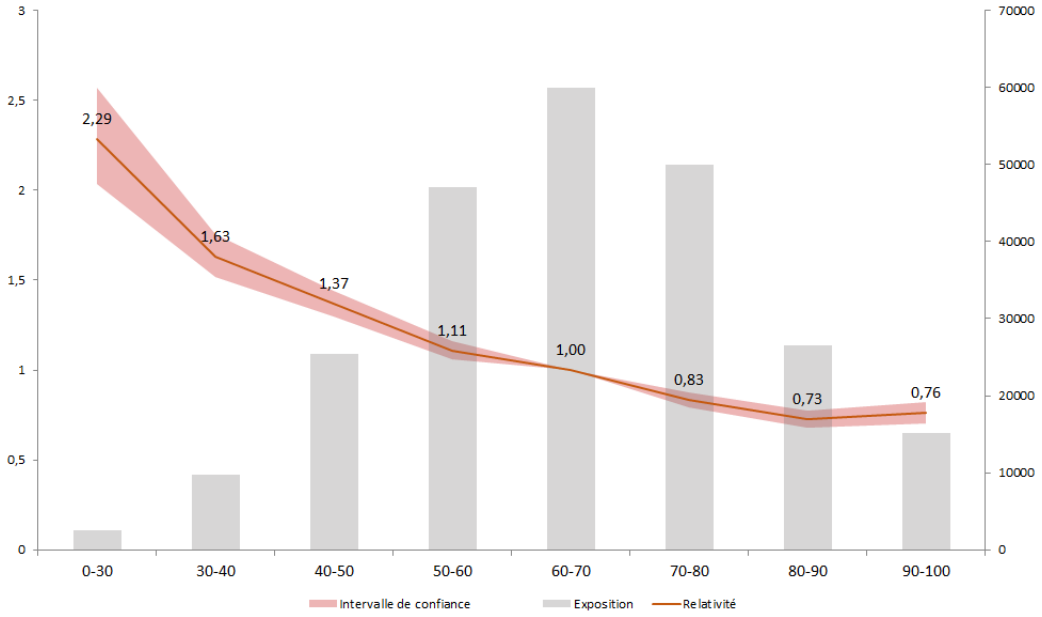

FIGURE 8.1 – Relativité du score de conduite sur la fréquence matérielle

Il existe bien une corrélation entre le score de conduite et la fréquence de sinistres car la relativité représentée sur le graphique précédent diminue en fonction des tranches de score. On représente donc la fréquence de sinistres en fonction du score de conduite sur les données Y6 et Y7 de chaque véhicule :

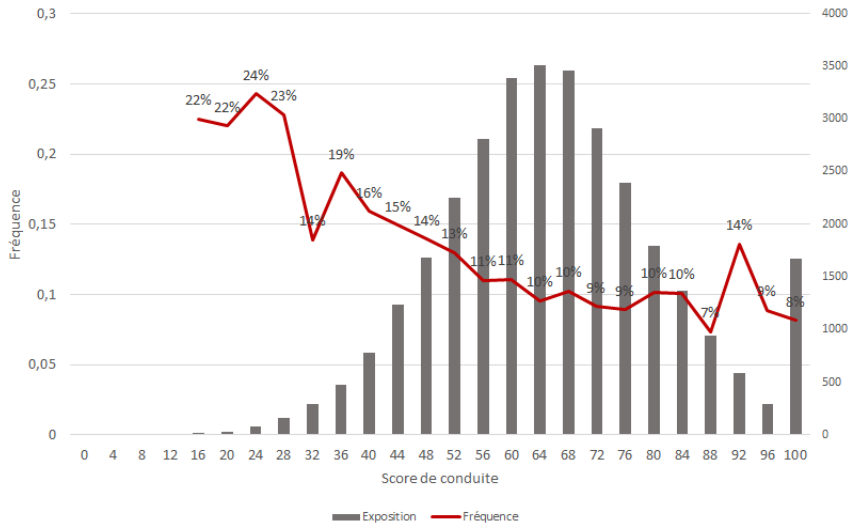

FIGURE 8.2 – Fréquence de sinistres en fonction du score de conduite

La courbe de la fréquence en fonction du score semble se rapprocher de la courbe d'une fonction géométrique.

Posons f une fonction géométrique de la forme :

$$f(x) = a^{b\times(x+c)} + d$$

avec a,b,c et d des réels.

Pour déterminer les quatre paramètres, nous calculons les écarts entre la fonction et l'observation : $Exposition \times (x - f(x))^2$, et nous cherchons à minimiser ces écarts. Grâce au solveur d'Excel, nous obtenons les coefficients optimaux. Par prudence, nous avons ajouté une condition supplémentaire : nous décidons que la fréquence moyenne théorique pondérée par l'exposition doit être supérieure ou égale à la fréquence moyenne du portefeuille. On obtient la fonction suivante :

$$f(x) = 0,97^{1,14\times(x+33,78)} + 0,077$$

Nous obtenons ainsi la courbe suivante :

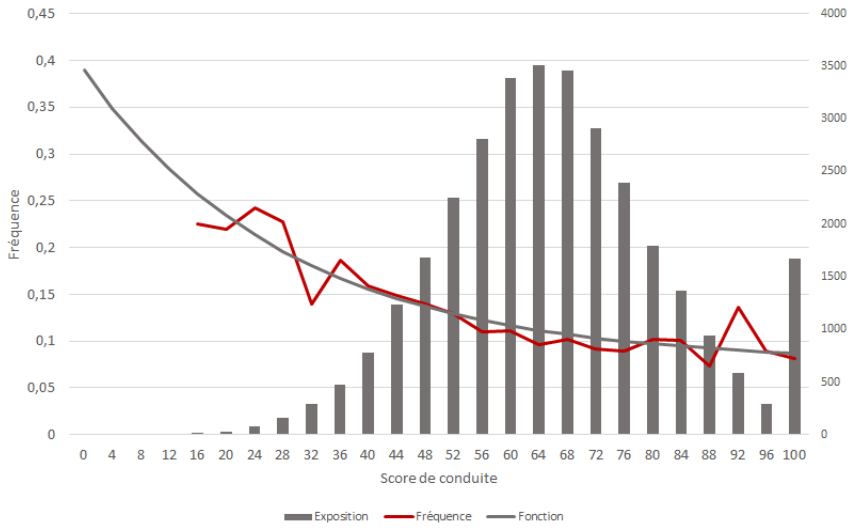

FIGURE 8.3 – Courbe géométrique de tendance

Cette fonction géométrique nous permet de traduire le score de conduite en fréquence de sinistres. Nous pouvons alors appliquer la théorie du Bonus-Malus.

8.2 Application de la théorie du Bonus-Malus

Nous réalisons les mêmes étapes que lors de la création de l'échelle Bonus-Malus en fonction de la sinistralité. Nous commençons donc par appliquer le modèle de Bühlmann-Straub et obtenons les indicateurs suivants :

σ	τ	K
4,88E-04	4,04E-05	12,06

FIGURE 8.4 – Indicateurs du modèles de Bühlmann-Straub pour le modèle Bonus-Malus score de conduite

Nous observons maintenant la boîte à moustaches du coefficient CRM obtenu :

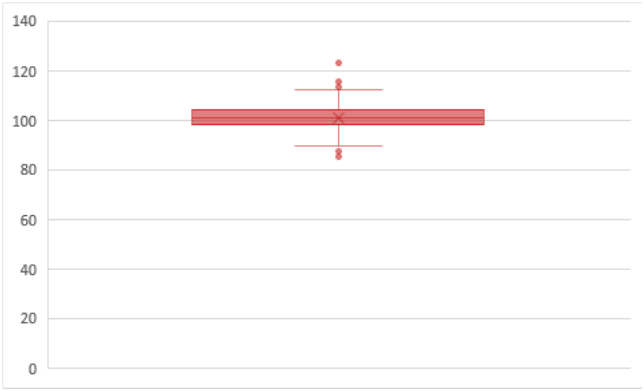

FIGURE 8.5 – Boîte à moustaches CRM Bonus-Malus en fonction du score de conduite

L'étendue observée dans ce cas est de 85 à 121 : elle est donc plus faible que celle de la sinistralité. L'évolution tarifaire serait donc plus faible entre les différentes classes.

8.3 Echelle Bonus-Malus

Observons la statistique du pseudo t au carré.

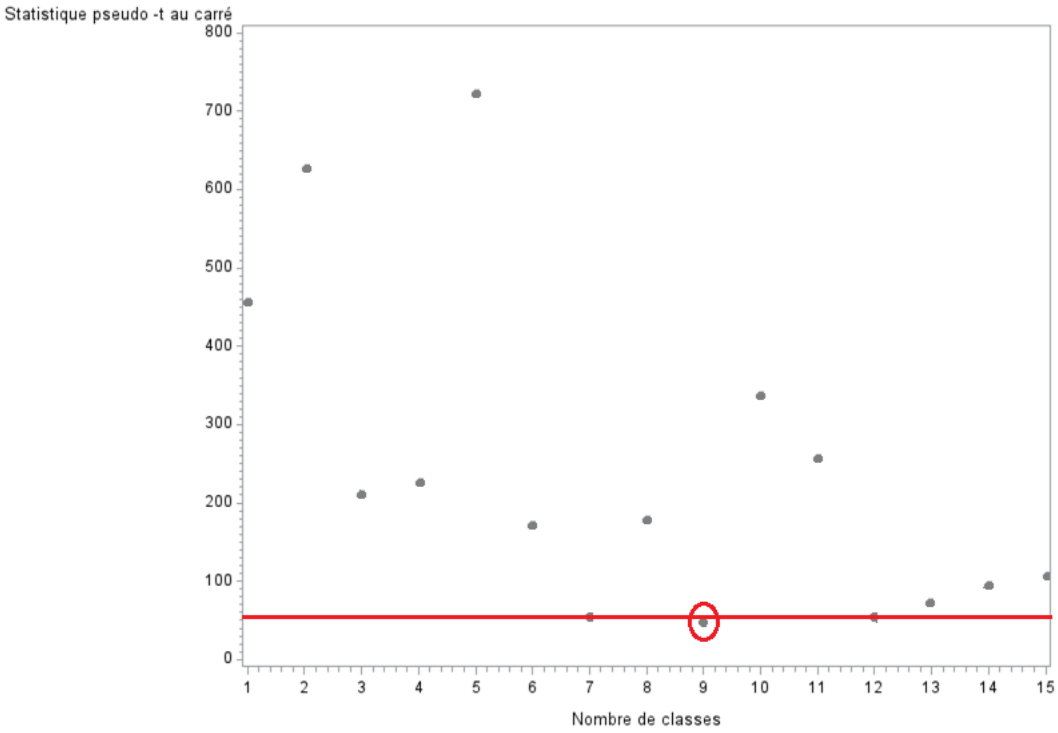

FIGURE 8.6 – Pseudo t au carré CRM en fonction du score de conduite

Afin de minimiser au mieux les écarts et de garder le plus d'évolution possible dans l'échelle, nous décidons de conserver neuf classes :

Classe CRM	Maximum Score de conduite	CRM Minimum	CRM Moyen	CRM Maximum	CRM Ecart type
1	90	85	89	91	1,6
2	84	91	93	95	1,2
3	73	95	97	98	0,7
4	69	98	100	101	0,9
5	65	101	103	104	0,8
6	62	104	105	106	0,6
7	60	106	107	109	0,8
8	56	110	111	113	0,8
9	54	114	116	123	2,8

FIGURE 8.7 – Statistique des classes de CRM Bonus-Malus fonction du score de conduite

Les classes obtenues sont cohérentes, il n'y a pas d'enchevêtrement de CRM.

Pour créer l'échelle Bonus-Malus nous décidons de conserver le CRM maximum de chaque classe, et pouvons ainsi construire une échelle Bonus-Malus en considérant l'évolution tarifaire d'une année sur l'autre de la manière suivante :

$$\text{"évolution tarifaire"} = (\text{CRM}_{max} - 100)/100$$

Nous obtenons l'échelle suivante :

Score de conduite	CRM Maximum	Evolution tarifaire	Ecart d'évolution entre classes
100 - 84	91	-9%	
84 - 90	95	-5%	4%
73 - 84	98	-2%	3%
69 - 73	101	+ 1%	3%
65 - 69	104	+ 4%	3%
62 - 65	106	+ 6%	2%
60 - 62	109	+ 9%	3%
56 - 60	113	+ 13%	4%
00 - 56	123	+ 23%	10%

FIGURE 8.8 – Echelle en fonction du score de conduite

Nous obtenons une échelle Bonus-Malus qui dépend du score de conduite. Les écarts entre les classes proches sont faibles (de l'ordre de 3%), à l'exception de la plus petite tranche qui représente les mauvais conducteurs et donc les "mauvais" risques.

Ces profils de "mauvais" risques sont caractérisés par les éléments télématiques relatifs à la conduite du véhicule : freinages et accélérations brusques, mouvements latéraux et virages brutaux. Cependant, il est possible qu'un individu ayant une conduite sportive, et donc un score de conduite plutôt élevé, ait une fréquence réelle inférieure à la moyenne. Pour éviter ces profils où le score de conduite est décorrélé de la sinistralité, nous proposons la construction d'une échelle Bonus-Malus qui dépendra conjointement de la fréquence historique et du score de conduite.

Chapitre 9

Modélisation Bonus-Malus "combinée" en fonction du score de conduite et de la sinistralité

Le score de conduite étant une variable issue d'une modélisation, elle n'a pas de formule explicite et a ainsi un effet "boîte noire". Cet effet empêche d'avoir la certitude qu'un véhicule avec un score de conduite élevé ait une fréquence réelle élevée. C'est pourquoi nous avons décidé de construire un Bonus-Malus qui dépendra à la fois du score de conduite et des fréquences historiques.

Dans un premier temps, nous nous attacherons à adapter la théorie de la modélisation de Bühlmann-Straub à notre cas, puis nous regrouperons les polices en classes de risques et enfin, nous construirons une échelle Bonus-Malus.

9.1 Adaptation de la théorie

Dans la théorie du Bonus-Malus, on attribue une certaine confiance à l'historique du client par rapport à la fréquence globale du portefeuille. Dans cette modélisation nous avons deux variables supplémentaire à la fréquence moyenne du portefeuille : la fréquence historique et le score de conduite historique.

Rappelons que la fréquence crédibilisée donnée par le modèle de Bühlmann-Straub (formule (7.6)) est de la forme :

$$N^j = Z^j X^j + \left(1 - Z^j\right) \mu_0$$

Dans notre cas nous avons deux variables au lieu d'une, nous avons alors identifié deux manières d'adapter la forme de la fréquence crédibilisée :

- $N^j = Z_1^j X^j + Z_2^j D^j + \left(1 - Z_1^j - Z_2^j\right) \mu_0$
  avec D^j la fréquence en fonction du score de conduite historique de la police j, Z_1^j

le facteur de crédibilité relatif à la sinistralité, et Z_2^j le facteur de crédibilité relatif au score de conduite. Cette formule ne se base plus sur la fréquence moyenne du portefeuille.

- $N^j = Z^j X^j + (1 - Z^j) D^j$
  avec D^j la fréquence en fonction du score de conduite historique de la police j.

Ne parvenant pas à résoudre le problème de façon théorique, nous cherchons à le résoudre de manière empirique. Nous cherchons ainsi à trouver les coefficients de crédibilité optimaux dans les deux formes possibles de la modélisation.

Nous construisons vingt bases de travail représentant chacune 80% de la base totale et 20 bases test qui correspondent aux données non présentes dans chacune des bases de travail.

Grâce à ces bases, nous calculons, pour chaque modélisation les coefficients de crédibilité optimaux, la base test nous donnant la fréquence à atteindre et la base de travail représentant nos fréquences historiques. Nous pouvons ainsi, sur chaque modélisation, trouver le coefficient de crédibilité minimisant les écarts : Exposition$\times$|fréquence test$-$fréquence crédibilisé modélisée|. Nous déterminons le coefficient de crédibilité permettant de minimiser les écarts quadratiques :

Observons à présent les coefficients de crédibilité obtenus avec chaque simulation :

	Somme des écarts quadratiques		**Coefficient de crédibilité**		
Simulation	Combiné (2 variations)	Combiné (1 variations)	Combiné (2 variations)		Combine (1 variation)
1	45,7	46,4	-6,8	67,5	40,7
2	55,4	55,4	38,7	-6,3	24,9
3	47,9	48,3	-8,8	41,7	19,6
4	51,3	52,1	-11,9	94,2	50,6
5	43,1	43,2	27,3	10,9	46,9
6	45,5	45,6	17,9	25,9	54,8
7	39,3	39,9	-10,5	122,8	71,8
8	55,5	56,3	-9,8	38,9	14,0
9	48,3	48,4	0,6	23,0	23,6
10	54,9	55,3	-10,0	123,5	81,0
11	39,2	39,6	1,7	20,2	21,3
12	49,8	49,9	12,5	20,4	39,3
13	50,5	50,5	21,9	1,7	24,7
14	42,2	42,3	19,3	2,4	23,1
15	46,8	47,1	57,4	-9,7	29,8
16	42,1	42,2	5,3	11,1	17,9
17	48,9	48,9	12,9	11,0	27,7
18	45,5	45,6	1,6	16,5	18,3
19	44,6	45,3	-8,1	103,4	70,0
20	43,0	43,1	21,9	10,4	38,1
Moyenne	47,0	47,3	8,6	36,5	36,9
Variance	24,1	24,6	351,7	1789,6	392,5

FIGURE 9.1 – Comparaison des coefficients obtenus

Si l'on compare les coefficients de crédibilité des deux modélisations combinées, on constate

que :

- Certains coefficients de la modélisation combinée avec deux variations sont négatifs.
- La volatilité des coefficients de la modélisation avec deux variations est très élevée en comparaison de celle avec une seule variation.

Afin d'obtenir la modélisation Bonus-Malus la plus robuste, nous décidons de conserver la modélisation avec un seul coefficient de crédibilité. Nous obtenons ainsi une modélisation de la forme :

$$N^j = Z^j X^j + \left(1 - Z^j\right) D^j$$

Plus Z est grand, plus nous faisons confiance à la sinistralité historique. Plus il est faible, plus nous faisons confiance au score de conduite. Nous fixons le coefficient de crédibilité K à la moyenne des coefficients optimaux des 20 simulations soit : 36,9. Nous obtenons ainsi une fréquence crédibilisée qui dépend, à la fois, de la sinistralité obtenue et du score de conduite.

9.2 Tentative de construction d'une échelle Bonus-Malus combinée

Dès lors, nous cherchons à créer une modélisation Bonus-Malus grâce à cette fréquence crédibilisée.

Commençons par regarder le nombre de classes optimales pour une échelle Bonus-Malus :

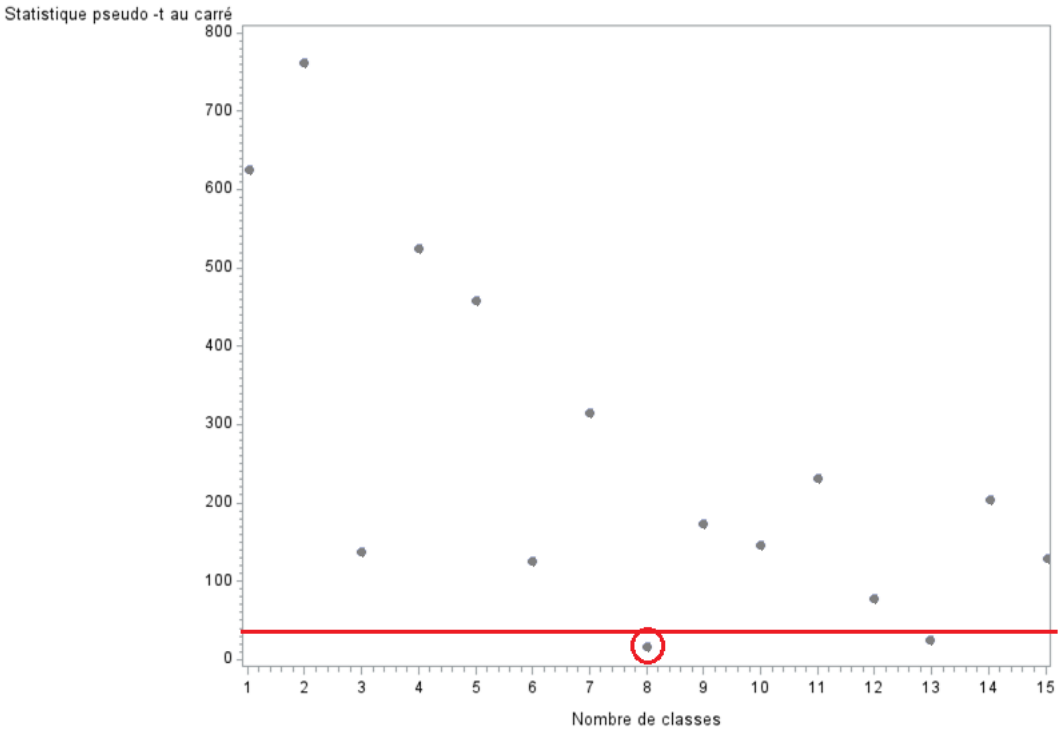

FIGURE 9.2 – Pseudo t au carré CRM composé

Nous choisissons de construire une échelle Bonus-Malus avec huit classes, pour minimiser l'erreur, et souhaitons ainsi délimiter chaque classe en fonction d'intervalles distincts de fréquence et de score de conduite.

Les classes ainsi obtenues sont ici représentées par des couleurs :

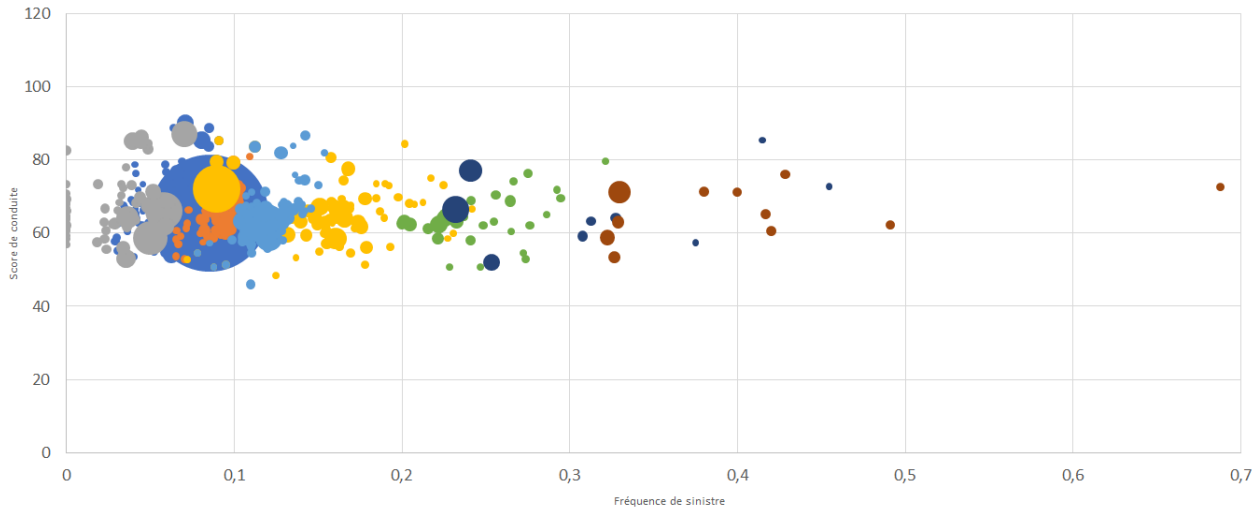

FIGURE 9.3 – Echelle combinée

Chaque couleur représentant une classe, on observe qu'elles sont principalement regroupées de façon verticale. On en déduit que les classes dépendent principalement de l'historique de la sinistralité et très peu du score de conduite.

Ce graphique nous montre que les différentes classes s'entremêlent. Il est donc compliqué de reconstituer les échelles en fonction du score de conduite et de la fréquence historique.

9.3 Echelle obtenue

La fréquence crédibilisée précédemment calculée n'est pas utilisable en pratique car les limites de score de conduite et de fréquence ne sont pas clairement établies.

Nous décidons donc de créer une nouvelle échelle à partir de l'échelle traditionnelle. Celle-ci sera composée de deux parties : la première partie reprend l'échelle traditionnelle et la seconde est fonction du score de conduite. Cette construction nous permet de garder une première partie, étage par étage, et une autre partie continue. Pour construire cette échelle, nous conservons la formule de crédibilité, mais nous l'adaptons de façon à ce que les étages de notre échelle soient calculés de la manière suivante :

$$\text{"Révision tarifaire composée"}^j = Z^j \times \text{"Révision tarifaire traditionnelle"}^j + (1 - Z^j) \times \frac{\text{Fréquence score}}{\mu_0} \times 100$$

avec $Z^j = \frac{W^j}{W^j + 36{,}9}$ le facteur de crédibilité de la police j, W^j son exposition et $\mu_0 = 10,90\%$ la fréquence moyenne du portefeuille sur Y6 et Y7.

Grâce à cette construction, nous obtenons l'échelle suivante pour chacune des échelles traditionnelles proposées précédemment :

Fréquence	Evolution tarifaire minimum	Evolution tarifaire moyenne	Evolution tarifaire maximum	Ecart d'évolution entre classes
0,0% - 7,0%	-28%	-16%	+ 1%	
7,0% - 8,9%	-17%	-8%	+ 7%	8%
8,9% - 11,5%	-6%	+ 2%	+ 13%	10%
11,5% - 15,9%	+ 5%	+ 13%	+ 20%	11%
15,9% - 24,7%	+ 15%	+ 31%	+ 49%	18%
24,7% - 32,1%	+ 38%	+ 53%	+ 80%	22%
32,1% - 45,5%	+ 43%	+ 68%	+ 118%	15%
45,5% - +	+ 103%	+ 117%	+ 132%	49%

FIGURE 9.4 – Echelle Bonus-Malus combinée issue de la première échelle traditionnelle

Si, pour construire l'échelle composée, nous utilisons plutôt l'échelle traditionnelle dont l'étendue a été diminuée, nous obtenons l'échelle composée suivante :

Fréquence	Evolution tarifaire minimum	Evolution tarifaire moyenne	Evolution tarifaire maximum	Ecart d'évolution entre classes
0,0% - 7,0%	-11%	-5%	+ 12%	
7,0% - 8,9%	-10%	-3%	+ 11%	2%
8,9% - 11,5%	-9%	-1%	+ 11%	2%
11,5% - 15,9%	-5%	+ 2%	+ 12%	3%
15,9% - 24,7%	-1%	+ 7%	+ 13%	5%
24,7% - 32,1%	+ 6%	+ 13%	+ 21%	6%
32,1% - 45,5%	+ 4%	+ 15%	+ 27%	2%
45,5% - +	+ 22%	+ 27%	+ 33%	12%

FIGURE 9.5 – Echelle Bonus-Malus combinée issue de la deuxième échelle traditionnelle

Nous obtenons ainsi deux échelles de Bonus-Malus qui combinent la sinistralité historique et le score de conduite.

Nous disposons à présent de plusieurs modélisations qui dépendent du score de conduite. Dans la dernière partie, nous nous attacherons à comparer ces différentes modélisations.

Chapitre 10

Comparaison des différentes modélisations

Nous avons réalisé précédemment plusieurs modélisations : deux modélisations GLM et cinq modélisations Bonus-Malus. Dans cette section, nous comparerons ces différentes modélisations.

Nous avons déjà comparé les modélisations GLM entre elles. La modélisation GLM avec le score de conduite est apparue plus précise que celle sans le score de conduite.

Nous allons désormais nous concentrer sur les modèles Bonus-Malus. Nous commencerons par illustrer l'application de ces modèles Bonus-Malus, puis nous comparerons la précision des différents modèles y compris les GLM. Nous observerons ensuite la convergence du score de conduite dans le temps. Enfin, nous ferons un récapitulatif des avantages et inconvénients des meilleures modélisations.

10.1 Application d'un modèle Bonus-Malus

Afin de comparer les différentes modélisations, nous simulons la charge sinistre sur Y7 pour la comparer avec la charge sinistre réelle. Pour les modélisations GLM, nous les avons appliquées sur les données contrats de Y7 en enlevant les impacts de la variable "année" qui ne servaient qu'à rendre la modélisation plus robuste. Pour les modélisations Bonus-Malus, nous avons appliqué les modèles sur les observations de Y6 afin d'obtenir Y7.

Nous avons déterminé la fréquence et le score moyen par client sur Y6. Grâce aux différentes échelles Bonus-malus, nous avons alors obtenu un coefficient de réduction-majoration. Nous avons appliqué cette majoration sur la prime pure GLM sans score de conduite en Y6, que nous avons pondéré par l'exposition en Y7. Cette approximation est réalisée dans le cadre de cette modélisation théorique car, en pratique, nous supposons que l'exposition est égale à 1 et que lors d'un départ de véhicule, la prime est ajustée.

Pour les flottes présentes en Y7 mais absentes en Y6, nous n'avons appliqué aucun changement de tarification, donc un coefficient de réduction majoration à 100.

Analysons désormais les impacts des différentes modélisations sur des cas concrets.

10.1.1 Illustration des modélisations sur une flotte de véhicules

Appliquons nos modélisations sur un exemple correspondant à une flotte de véhicules appelée flotte 0.

Score de conduite moyen	63
Fréquence de sinistres moyen	7,0%
Exposition Y6 $W_{Y6}^{\text{"flotte 0"}}$	14,31
Exposition Y7 $W_{Y7}^{\text{"flotte 0"}}$	10,10
Prime pure modélisée par le GLM sans score de conduite Y6	8 955

FIGURE 10.1 – Caractéristiques de la flotte 0

De ces différentes caractéristiques, nous déduisons les évolutions tarifaires selon le modèle :

Echelle traditionnelle de base	- 29 %
Echelle traditionnelle réduite	- 10 %
Echelle score	+ 6 %
Echelle combinée de base	- 10 %
Echelle combinée réduite	- 2 %

FIGURE 10.2 – Evolutions tarifaires des différentes modélisations sur la flotte 0

Pour chaque modélisation nous appliquons :

$$\text{valeur modélisée} = \frac{W_{Y7}^{\text{"flotte 0"}}}{W_{Y6}^{\text{"flotte 0"}}} \times (1 + \text{évolution tarifaire}) \times \text{modélisation GLM Y6}$$

Par exemple, pour la modélisation se basant sur l'échelle score, la flotte a un score de conduite moyen de 63, ce qui lui procure une évolution tarifaire de + 6 %. Sachant que sa prime pure GLM Y6 est de 8 955 €, la prime pure modélisée avec l'échelle Bonus-Malus "score de conduite" est de $\frac{10,10}{14,31} \times (1 + 6\%) \times 8955 = 6700\,€$.

Les modélisations GLM n'ont pas d'évolution tarifaire, les primes pures modélisées dépendent uniquement des caractéristiques de la flotte et du véhicule.

Nous obtenons donc pour cette flotte les primes pures Y7 modélisées suivantes :

GLM sans score	9 846 €
GLM avec score	8 982 €
Echelle traditionnelle de base	4 488 €
Echelle traditionnelle réduite	5 688 €
Echelle score	6 700 €
Echelle combinée de base	5 688 €
Echelle combinée réduite	6 194 €

FIGURE 10.3 – Prime pure modélisée de la flotte 0

10.1.2 Illustration des modélisations sur les flottes

Nous avons appliqué les différentes modélisations sur une flotte. Observons désormais leur application sur des flottes de véhicules.

Base de travail

Prenons une sous-base contenant 10 flottes pour lesquelles nous calculons les montants de primes pures modélisés pour l'année Y7 à partir de l'année Y6 :

Flotte cas	Exposition Y6	Exposition Y7	Fréquence Y6	score Y6	Prime pure GLM Y6
A	3 268	3 078	8,60%	68	1 148 995
B	482	580	8,51%	74	140 877
C	99	134	27,16%	73	80 204
D	37	38	19,01%	68	12 120
E	87	62	11,50%	82	33 083
F	15	19	12,99%	65	11 054
G	14	10	6,99%	63	8 955
H	366	269	5,19%	66	93 439
I	54	108	11,13%	64	12 555
K	31	49	9,79%	66	8 254

FIGURE 10.4 – Caractéristiques de la sous-base

Observons les différentes échelles associées :

Numéro cas	Echelle traditionnelle de base	Echelle traditionnelle réduite	Echelle score	Echelle combinée de base	Echelle combinée réduite
A	- 17%	- 7%	+ 1%	- 17%	- 7%
B	- 17%	- 7%	- 5%	- 17%	- 7%
C	+ 94%	+ 21%	- 2%	+ 80%	+ 17%
D	+ 56%	+ 12%	+ 1%	+ 37%	+ 7%
E	+ 3%	- 2%	- 5%	+ 0%	- 4%
F	+ 20%	+ 3%	+ 1%	+ 10%	+ 2%
G	- 29%	- 10%	+ 4%	- 10%	- 2%
H	- 29%	- 10%	+ 1%	- 28%	- 9%
I	+ 3%	- 2%	+ 4%	+ 3%	- 1%
K	+ 3%	- 2%	+ 1%	+ 2%	- 1%

FIGURE 10.5 – Echelle de la sous-base

Analyse en fréquence

Grâce à ces évolutions, nous pouvons modéliser les fréquences pour chaque échelle Bonus-Malus : les cellules qui se rapprochent le plus de la fréquence réelle Y7 apparaissent en vert dans le tableau :

Flotte	Fréquence réelle Y7	Modélisation traditionnelle de base	Modélisation traditionnelle réduite	Modélisation score	Modélisation combinée de base	Modélisation combinée réduite
A	8,48%	7,14%	8,00%	8,68%	7,14%	8,00%
B	9,31%	7,06%	7,91%	8,08%	7,06%	7,91%
C	37,28%	52,68%	32,86%	26,61%	48,88%	31,77%
D	13,18%	29,65%	21,29%	19,20%	26,04%	20,34%
E	3,23%	11,85%	11,27%	10,93%	11,50%	11,04%
F	15,89%	15,58%	13,38%	13,12%	14,29%	13,25%
G	9,90%	4,96%	6,29%	7,27%	6,29%	6,85%
H	6,69%	3,69%	4,67%	5,24%	3,74%	4,72%
I	10,16%	11,47%	10,91%	11,58%	11,47%	11,02%
K	6,16%	10,08%	9,59%	9,88%	9,98%	9,69%

FIGURE 10.6 – Fréquence modélisée sur les flottes de la sous-base

La modélisation qui se rapproche le plus souvent de la fréquence réelle est l'échelle Bonus-Malus "score de conduite". La version de base de l'échelle traditionnelle, ainsi que celle combinée, ont des variations très élevées comparé à la fréquence.

En revanche, la version réduite de la modélisation Bonus-Malus "combinée" a une modélisation qui est toujours comprise entre la modélisation traditionnelle et celle avec le score de

conduite. Elle n'apparaît pas comme la plus précise mais ses écarts sont toujours plus faibles que la moins bonne modélisation.

La flotte C est un mauvais risque car ses fréquences sont supérieures à 20% et son exposition supérieure à 95. En plus de sa fréquence élevée, cette flotte a un score de conduite élevé. C'est la raison pour laquelle la modélisation Bonus-Malus "score de conduite" indique une diminution tarifaire alors que la fréquence réelle indiquerait, elle, un effet inverse.

La flotte D a peut-être eu une "mauvaise" année. La modélisation a montré une évolution tarifaire de + 12%, alors qu'en passant à l'année Y7, sa fréquence a diminué de 6%. La modélisation "score de conduite" a, quant à elle, bien fonctionné.

La flotte E est comparable à la flotte D : sa fréquence est assez volatile d'une année sur l'autre. Or elle a un très bon score de conduite. Il paraît donc très probable que sa fréquence Y6 soit liée à une "mauvaise" année.

Ces cas appliqués nous permettent d'observer l'efficacité des modélisations ainsi que leurs limites :

- Le score de conduite peut ainsi être décorrélé de la fréquence de sinistres.
- La modélisation Bonus-Malus traditionnelle n'est pas très efficace lorsqu'il s'agit de modéliser des flottes qui ont une certaine volatilité dans leur fréquence de sinistres.

Analyse de la prime pure

A l'aide des révisions tarifaires et des primes pures GLM Y6, nous obtenons les primes pures Y7 des différentes flottes en fonction du type de modélisation :

Numéro cas	Coût réel Y7	GLM sans score	GLM avec score	Modélisation traditionnelle de base	Modélisation traditionnelle réduite	Modélisation score	Modélisation combinée de base	Modélisation combinée réduite
A	695	1 182	1 205	897	969	1 025	898	970
B	117	154	135	116	130	133	117	130
C	103	124	129	181	125	107	170	122
D	15	13	12	19	14	13	17	13
E	11	27	21	24	22	22	23	22
F	14	14	12	16	15	15	15	15
G	1	10	9	7	7	8	7	7
H	70	102	93	48	61	68	49	61
I	27	28	29	27	27	27	27	27
K	4	15	14	14	14	14	14	14

FIGURE 10.7 – Prime pure modélisée sur les flottes de la sous-base

D'un point de vue financier, la modélisation Bonus-Malus "score de conduite" est celle qui est le plus souvent la plus proche du coût réel.

Comparons désormais la précision de nos modèles sur tout le portefeuille.

10.2 Précision des modélisations

Nous observons à présent quelles sont les modélisations les plus précises sur tout le portefeuille composé de 430 flottes. Les modélisations ont été réalisées sur chacune des flottes puis sommées.

Pour information, les caractéristiques du portefeuille considéré sont les suivantes :

Score de conduite moyen Y6	67
Fréquence de sinistres moyen Y6	10,73%
Exposition Y6	15 693
Exposition Y7	16 461
GLM sans score de conduite Y6	6 666 413

FIGURE 10.8 – Caractéristiques du portefeuille

La prime pure sommée de toutes les flottes sur Y7 obtenue grâce aux différents types de modèles est la suivante :

Coût réel	6 520 243
GLM sans score	7 961 786
GLM avec score	7 782 900
Modélisation traditionnelle de base	7 561 219
Modélisation traditionnelle réduite	6 800 041
Modélisation score	6 969 962
Modélisation combinée de base	7 274 168
Modélisation combinée réduite	6 801 085

FIGURE 10.9 – Prime pure modélisée sur le portefeuille

Les modélisations qui se rapprochent le plus du coût réel sont : le Bonus-Malus combinée réduite, le Bonus-Malus traditionnelle réduite et le Bonus-Malus "score de conduite".

Afin d'affiner nos analyses, nous découpons le portefeuille en une sous-base contenant 20 % des données sélectionnées aléatoirement. Nous renouvelons ce procédé sur 20 sous-échantillons.

Pour chaque échantillon, des primes pures sont simulées sur Y7 par flotte puis sommées. La somme par échantillon des écarts observés (exposition par flotte $*$ |prime pure réelle $\times$ prime pure simulée|) pour chaque modélisation est la suivante :

	GLM sans score	GLM avec score	Modélisation traditionnelle de base	Modélisation traditionnelle réduite	Modélisation score	Modélisation combinée de base	Modélisation combinée réduite
Echantillon 1	1 916	1 909	1 376	1 300	1 316	1 337	1 299
Echantillon 2	2 390	2 395	1 879	1 804	1 820	1 842	1 803
Echantillon 3	1 946	1 939	1 408	1 310	1 317	1 365	1 306
Echantillon 4	2 131	2 134	1 616	1 545	1 564	1 584	1 545
Echantillon 5	2 096	2 086	1 401	1 331	1 347	1 372	1 331
Echantillon 6	1 947	1 936	1 475	1 394	1 407	1 439	1 392
Echantillon 7	2 000	2 011	1 501	1 429	1 443	1 470	1 428
Echantillon 8	1 988	1 996	1 451	1 381	1 396	1 423	1 381
Echantillon 9	2 166	2 173	1 592	1 518	1 531	1 561	1 517
Echantillon 10	2 441	2 428	1 971	1 911	1 929	1 946	1 911
Echantillon 11	2 167	2 164	1 557	1 480	1 494	1 526	1 481
Echantillon 12	2 588	2 584	2 049	1 972	1 981	2 014	1 969
Echantillon 13	2 461	2 461	1 955	1 886	1 903	1 924	1 886
Echantillon 14	2 017	2 012	1 508	1 447	1 466	1 484	1 448
Echantillon 15	2 674	2 674	2 042	1 966	1 976	2 008	1 963
Echantillon 16	2 528	2 523	1 943	1 878	1 895	1 914	1 879
Echantillon 17	2 180	2 187	1 620	1 557	1 575	1 591	1 557
Echantillon 18	2 735	2 757	2 271	2 192	2 205	2 239	2 191
Echantillon 19	2 493	2 508	2 028	1 946	1 959	1 990	1 944
Echantillon 20	2 115	2 097	1 552	1 483	1 500	1 525	1 484
Moyenne	**2 249**	**2 249**	**1 710**	**1 637**	**1 651**	**1 678**	**1 636**

FIGURE 10.10 – Ecart de chaque modélisation sur les vingt sous-échantillons

Nous attribuons un rang entre 1 et 7 à chaque modélisation pour chaque échantillon avec 1, le modèle ayant le plus petit écart et 7, le plus grand. Le graphique suivant regroupe les résultats :

	GLM sans score	GLM avec score	Modélisation traditionnelle de base	Modélisation traditionnelle réduite	Modélisation score	Modélisation combinée de base	Modélisation combinée réduite
Echantillon 1	7	6	5	2	3	4	1
Echantillon 2	6	7	5	2	3	4	1
Echantillon 3	7	6	5	2	3	4	1
Echantillon 4	6	7	5	1	3	4	2
Echantillon 5	7	6	5	2	3	4	1
Echantillon 6	7	6	5	2	3	4	1
Echantillon 7	6	7	5	2	3	4	1
Echantillon 8	6	7	5	2	3	4	1
Echantillon 9	6	7	5	2	3	4	1
Echantillon 10	7	6	5	1	3	4	2
Echantillon 11	7	6	5	1	3	4	2
Echantillon 12	7	6	5	2	3	4	1
Echantillon 13	6	7	5	1	3	4	2
Echantillon 14	7	6	5	1	3	4	2
Echantillon 15	6	7	5	2	3	4	1
Echantillon 16	7	6	5	1	3	4	2
Echantillon 17	6	7	5	2	3	4	1
Echantillon 18	6	7	5	2	3	4	1
Echantillon 19	6	7	5	2	3	4	1
Echantillon 20	7	6	5	1	3	4	2
Moyenne	**7**	**6**	**5**	**2**	**3**	**4**	**1**

FIGURE 10.11 – Rang de chaque modélisation sur les vingt sous-échantillons

La modélisation la plus précise est la modélisation Bonus-Malus combinée réduite, suivie par le Bonus-Malus traditionnelle réduite et enfin le Bonus-Malus score de conduite.

10.3 Avantages et inconvénients des différentes modélisations Bonus-Malus

La figure ci-dessous permet de récapituler les avantages et inconvénients de chaque type de modélisation Bonus-Malus :

	+ Points positifs	- Points négatifs
Modélisation Bonus Malus en fonction de la sinistralité	- Précision - Fonctionnement connu	- Biais possible sur les petites flottes
Modélisation Bonus Malus en fonction du score	- Indépendant de la sinistralité	- Effet boîte noire
Modélisation Bonus Malus combinée	- Précision - Fonctionnement connu en partie - Fonctionnement simple	- Effet boîte noire - Réaction client inconnue

FIGURE 10.12 – Avantages et inconvénients des types de modélisation Bonus-Malus

La modélisation Bonus-Malus en fonction de la sinistralité est globalement très efficace. Pourtant, sur les flottes avec peu d'exposition, la fréquence est très volatile, ce qui implique que la révision tarifaire le soit également. Cette échelle est cependant déjà existante sur la responsabilité civile des particuliers et, à ce titre, très connue.

La modélisation Bonus-Malus "score de conduite" est moins précise mais moins volatile. On observe cependant un effet "boîte noire" lors du calcul du score de conduite.

La modélisation Bonus-Malus "combinée" correspond à l'entre-deux des deux modélisations précédentes. La précision de ce modèle apparaît comme étant la meilleure des différents modèles observés.

A l'heure actuelle, ces modélisations et constructions sont inconnues du public : aucune offre de tarification utilise ce type de modèle combiné. Il est donc délicat d'anticiper la réaction des clients face à ce type de tarification.

Conclusion

Dans le secteur automobile, les nouvelles technologies permettent aux concepteurs de réfléchir à des modèles innovants : véhicules électriques, hybrides, autonomes. Les véhicules récents sont fréquemment équipés d'interfaces connectées : GPS intégré, Bluetooth, aides à la conduite. Dans ce contexte, les informations transmises par la télématique embarquée offrent un nouveau champ d'application pour l'assurance automobile. A ce jour, quelques contrats proposent d'intégrer ces nouvelles informations pour assurer les véhicules des particuliers, mais aucun pour les flottes automobiles.

Les boîtiers télématiques embarqués captent les informations relatives à la conduite : accélération, décélération, mouvement latéral, vitesse angulaire, et l'exploitation de ces données permet de déterminer un score de conduite. Il représente la dangerosité de la conduite d'un individu et correspond à une variable explicative de la sinistralité.

Le but de cette étude était d'optimiser l'utilisation de cette variable "score de conduite" dans les modélisations de la sinistralité afin de construire une offre tarifaire aux flottes automobiles dépendante de cette nouvelle variable. Pour répondre à cette problématique, nous avons observé l'apport de la variable "score de conduite" dans les modèles GLM.

Dans les modélisations GLM de la sinistralité matérielle, le score de conduite permet une nette amélioration de la modélisation, tant en fréquence qu'en coût moyen. Dans la modélisation GLM de la sinistralité corporelle, la variable "score de conduite" ne présente une plus-value que dans la modélisation en fréquence. Dans ces trois modèles GLM, l'évolution du score de conduite était à l'inverse de l'indicateur de la sinistralité. La comparaison des différentes modélisations sur une année déterminée prouve que la modélisation GLM avec le score de conduite est plus précise. Au travers de nos modélisations GLM, nous montrons que le score de conduite apporte une information sur le conducteur qui n'était jusqu'alors pas disponible dans l'offre d'assurance de flottes automobiles.

Afin d'affiner les tarifications proposées, nous avons modélisé plusieurs échelles Bonus-Malus. La première est une échelle traditionnelle prenant en compte l'historique de sinistralité. La seconde est dépendante du score de conduite. Enfin, la troisième échelle est une modélisation Bonus-Malus "combinée", qui dépend à la fois du score de conduite et de l'historique de la sinistralité.

Au final, la modélisation qui maximise le potentiel de la variable "score de conduite" est la modélisation Bonus-Malus basée simultanément sur l'historique et le score de conduite. Cette échelle attribue au score de conduite un impact plus ou moins important en fonction de la taille du client : plus la flotte est importante moins la variable "score de conduite" intervient.

L'utilisation judicieuse et raisonnée de ces informations pourrait permettre la création d'une offre de segmentation individualisée et une tarification optimisée de chaque flotte de véhi-

cules.

Dans une prochaine étude, il pourrait être intéressant d'utiliser d'autres applications de la télématique, comme le nombre de kilomètres parcourus ou le recensement des zones dangereuses. Ces variables pourraient permettre d'affiner les modélisations.

Table des figures

Bibliographie

[BP92] Par Jean-Luc Besson et et Christian Partrat. "Trend et systèmes de Bonus-Malus". In : *ASTIN Bulletin* 22.1 (1992), p. 11-31. DOI : 10.2143/AST.22.1.2005124.

[Pin03] Jean Pinquet. "Prise en compte de l'ancienneté des périodes dans les modèles de risque en fréquence : aspects théoriques et applications à la tarification des risques en assurance automobile". fr. In : *Journal de la société française de statistique* 144.3 (2003), p. 7-27. URL : http://www.numdam.org/item/JSFS_2003__144_3_7_0.

[ABD13] Boubakeur BENLAIB ABDELOUAHAB LATRECHE. "Systèmes de modification des primes a posteriori systèmes bonus-malus : Poisson-Gamma à effet aléatoire Etude empirique (Cas de la SAA)". Fr. In : 17.5 (2013), p. 236-250. ISSN : 1112-2382. URL : https://www.asjp.cerist.dz/en/article/54554.

[CHO] Christian CHOW. *Utilisation des données télématiques pour l'analyse de la sinistralité automobile*. URL : https://www.institutdesactuaires.com/se-documenter/memoire-d-actuariat-38?id=bb61369ed7b8629dd91a707f2422a157. (accessed : 10.08.2020).

[GRA] Lison GRAPPIN. *Space-time modelling of roads inherent risk using telematics data*. URL : https://www.institutdesactuaires.com/se-documenter/memoire-d-actuariat-38?id=ddf68d1a14bacab1d50167b7fdb2c41a. (accessed : 10.08.2020).

[LAN] Damien LANDON. *Construction d'une échelle bonus-malus pour la tarification des flottes de véhicules*. URL : http://www.ressources-actuarielles.net/C12574E200674F5B/0/F20E9C74DB455FA2C12580F30068829E. (accessed : 10.08.2020).

[QIU] Henri QIU. *Etude des données de conduite et du score pour un contrat d'assurance automobile télématique*. URL : http://www.ressources-actuarielles.net/C12574E200674F5B/0/3838DAF4ABF2BFA6C12581080033A664. (accessed : 10.08.2020).

[SAD] Kevin SADOUN. *Apport des télématiques dans la segmentation tarifaire en assurance automobile*. URL : http://www.ressources-actuarielles.net/C12574E200674F5B/0/2EF09BF7724C8EF3C125806500273FB1. (accessed : 10.08.2020).

[Son] Cristina SONTU. *Étude de l'évolution des règles de souscription en assurance automobile grâce à un produit télématique*. URL : http://www.ressources-actuarielles.net/C12574E200674F5B/0/DA2C629C81A31283C12582CE005C3059. (accessed : 10.08.2020).

[VIT] Léa VITRAC. *Analyse de la sinistralité observée sur le produit télématique par rapport à la sinistralité du produit traditionnel*. URL : http://www.ressources-actuarielles.net/C12574E200674F5B/0/4F7B3E48A46BCA00C125821900296C27. (accessed : 10.08.2020).

[To-] NGUYEN Thi TO-VONG. *Flottes automobiles : Un nouveau modèle de tarification. Impact de la conservation sur la distribution du ratio sinistres à primes*. URL : http://www.ressources-actuarielles.net/C12574E200674F5B/0/B7F2F5AF6084D90BC125762C002DFD03. (accessed : 10.08.2020).