IRM | institut du risk management BY Institut.

Mémoire présenté pour la validation de la Formation
« Certificat d'Expertise Actuarielle »
de l'Institut du Risk Management
et l'admission à l'Institut des actuaires

Par : Vivian Elena MOSQUERA PEREZ

Titre : Sophistication tarifaire de la garantie assistance routière

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

Membres présents du jury de l'Institut des actuaires :

_____

_____

_____

Membres présents du jury de l'Institut du Risk Management :

_____

_____

_____

_____

_____

_____

_____

_____

Secrétariat :

Bibliothèque :

Entreprise :

Nom : Europ Assistance Groupe

Signature et Cachet : Etienne BONNET

europ assistance
S.A. au capital de 48 123 637 €
Siège Social : 2, rue Pillet-Will
75009 PARIS
01 41 85 86 86
SIRET 451 366 405 00012

Directeur de mémoire en entreprise :

Nom : Olivier DANNEAUX

Signature :

Invité :

Nom : _____

Signature :

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

Signature du responsable entreprise

Signature(s) du candidat(s)

Remerciements

A mon directeur de thèse Olivier DANNEAUX, pour son soutien, ses précieux conseils et le temps qu'il m'a accordé tout au long de la réalisation de cette mémoire.

A Etienne BONNET, pour m'avoir soutenu dans ma démarche de devenir actuaire.

A mes collègues d'Europ Assistance, pour leurs encouragements constants, notamment dans les moments les plus stressants.

Aux intervenants du CEA, pour leur encadrement et leurs conseils durant les 2 années de formation.

A ma famille, pour leur soutien moral même à distance.

A mon copain Yoann, pour sa patience, son support technique et émotionnel durant l'élaboration de cette étude.

Mes remerciements vont aussi à toute personne qui, de près ou de loin, a contribué à la réalisation de ce mémoire.

Table des matières

Introduction

Le marché global de l'assistance routière est en forte croissance grâce aux nouveaux besoins qui émergent autour des véhicules hybrides et électriques, au vieillissement du parc automobile et au développement de la digitalisation dans le secteur automobile (véhicules connectés et autonomes). Les sociétés d'assistance ont le défi de s'adapter à un secteur automobile en constante transformation et donc à l'évolution des risques qui en découle.

Dans le cadre d'un modèle de distribution *Business to Business (BtoB)* souvent utilisé lors de la commercialisation de la garantie d'assistance routière, l'assureur n'est pas exposé à l'antisélection. En effet, le modèle BtoB est un modèle en inclusion. Le bénéficiaire acquiert directement la garantie d'assistance lors de la souscription d'une police d'assurance, de l'achat d'un véhicule ou d'un service proposé par l'entreprise partenaire. La prime est donc la même quel que soit le profil de risque de l'acquéreur. Par conséquent, la garantie d'assistance routière est souvent tarifiée à partir d'un modèle simple de tarification qui présente l'inconvénient majeur de ne pas capturer les facteurs de risque sous-jacents du portefeuille souscrit.

Ne pas prendre en compte les facteurs de risque ayant un impact sur la prime pure d'une garantie d'assistance routière empêche l'assureur d'avoir une connaissance précise du risque souscrit et implique, par conséquent, une gestion difficile de son positionnement lors de la participation aux appels d'offres pour remporter de nouveaux contrats ainsi que pour le maintien de la rentabilité des contrats existants.

Ce mémoire propose une méthode de tarification pour un contrat BtoB en France en utilisant des modèles linéaires généralisés (GLMs). L'objectif est de prendre en compte les facteurs de risque dans le calcul de la prime pure unique commune à tous les bénéficiaires afin de comparer les résultats obtenus avec la méthode de tarification simple actuellement utilisée. L'étude s'intéresse spécifiquement au segment des constructeurs automobiles. Ce segment particulièrement concurrentiel contribue de manière significative au chiffre d'affaires représenté par les contrats BtoB de la ligne produit automobile d'Europe Assistance Groupe.

Etant donné que les variables permettant de décrire le risque de fréquence et de sévérité pour le calcul de la prime pure ne sont pas souvent disponibles dans le cas des portefeuilles appartenant au segment client des constructeurs automobiles, une base de données a été mise à disposition par un assureur partenaire. La richesse et l'hétérogénéité de l'information disponible a été utilisée pour identifier les variables impactant les risques de fréquence et de sévérité dans le cas de deux fait générateurs : les pannes et des accidents.

L'étude s'articule en 6 parties :

- La *partie I* fournit des éléments de contexte pour mieux comprendre les problématiques abordées dans ce mémoire et présenter les principaux enjeux de l'étude.

- La *partie II* présente une analyse chiffrée permettant de qualifier le portefeuille en termes de fréquence et de coût moyen dans son ensemble. Après la définition du périmètre de l'étude, les ajustements effectués sur la base de données, la vérification de la stabilité du portefeuille dans le temps et les analyses exploratoires des variables explicatives seront présentés.

- La *partie III* expose les étapes de pré-sélection des variables réalisées préalablement à la modélisation afin de pouvoir traiter une base de données exploitable en termes de temps de calcul et de complexité.

- La *partie IV* intègre dans l'étude les disparités géographiques qui peuvent exister au niveau de la fréquence et le coût moyen du portefeuille étudié par la création d'un zonier.

- La *partie V* développe les modèles linéaires généralisés (GLM) pour estimer la fréquence et le coût moyen des deux faits générateurs (pannes et accidents). De même, elle présente l'estimation de la nouvelle prime pure qui sera comparée à celle estimée par un modèle de tarification simple.

1. Contexte général

L'objectif de ce chapitre est de fournir des éléments de contexte pour mieux comprendre les problématiques abordées dans ce mémoire et de présenter les principaux enjeux de l'étude.

1.1. Définition et fonctionnement

1.1.1. L'assistance

L'assistance est une branche de l'assurance non-vie plus particulièrement de l'assurance IARD (incendies, accidents et risques divers). Il s'agit d'une garantie de services destinés à porter assistance au bénéficiaire en cas de sinistre couvert par le contrat.

La garantie d'assistance est née en Suède dans les années 50 à la suite du développement du tourisme et des voyages de longue distance. Elle est principalement adressée aux voyageurs souhaitant se couvrir du risque de difficulté à l'étranger contre le paiement préalable d'une prime (Atlas Magazine, 2012). En 1963, Pierre Desnos fonde Europ Assistance soutenu par le Groupe Generali, la première société de services en France proposant une protection aux personnes voyageant à l'étranger (Europ Assistance, 2022). Depuis, l'assistance s'est développée dans d'autres pays européens et s'est étendue à d'autres segments tel que l'automobile, l'habitation et la santé.

Le premier article de la directive du Conseil CEE du 10 décembre 1984 défini cette garantie comme *« l'assistance fournie aux personnes en difficulté au cours de déplacements ou d'absences du domicile ou du lieu de résidence permanente. Elle consiste à prendre, moyennant le paiement préalable d'une prime, l'engagement de mettre immédiatement une aide à la disposition du bénéficiaire d'un contrat d'assistance lorsque celui-ci se trouve en difficulté par suite d'un événement fortuit, dans les cas et dans les conditions prévues par le contrat ».* Les opérations d'assistance sont régies par le code des assurances et sont classées dans la branche 18 d'après l'article R32, 1-1. De ce fait, la prestation doit être liée à un évènement aléatoire, le contrat doit être consenti par les deux parties et la compagnie d'assistance doit avoir un agrément de l'ACPR, être conforme à la Directive sur la Distribution des Assurances (DDA), et est assujettie règlementation Solvabilité II et comptabilité IFRS17/IFRS9.

Bien que l'assistance soit considérée comme une garantie d'assurance, elle diffère des garanties traditionnelles. La compagnie d'assistance s'engage à fournir au(x) bénéficiaire(s) une aide logistique plutôt qu'une indemnisation financière. C'est pourquoi la société d'assistance travaille en étroite collaboration avec un réseau de plusieurs partenaires qui fournissent différents types de services et qui interviennent au niveau international, y compris dans des endroits où la société n'est pas présente.

Des garanties d'assistance se trouvent principalement sur les segments suivants :

- Voyage : Rapatriement, assistance médicale, évacuation de sécurité, etc.
- Habitation : Assistance en cas de perte des clefs, problèmes de plomberie et/ou incendie ; protection du domicile et protection cyber ; réparation des appareils ménagers, garde d'enfants, etc.
- Santé : Soins et aides à domicile, téléconsultation, transport de patients malades, soins aux personnes âgées, envoi de médicaments, etc.
- Automobile : Dépannage, remorquage, véhicule de replacement, rapatriement du véhicule réparé ou non réparé, hébergement, etc.

Etant donné que ce mémoire se concentrera sur la garantie d'assistance routière automobile, une définition plus détaillée est présentée dans la section 1.1.2.

1.1.2. L'assistance routière automobile

L'objectif d'un produit d'assistance routière est de fournir un support au bénéficiaire lorsque son véhicule ne peut plus être conduit dans des conditions normales de sécurité. Les causes de ceci peuvent être diverses :

Figure 1.1 - Causes d'un sinistre d'assistance routière

Les principales prestations fournies par la société d'assistance sont :

Figure 1.2 - Principales prestations de l'assistance routière

En fonction des conditions spécifiques du contrat, d'autres avantages peuvent être proposés afin d'assurer la mobilité du bénéficiaire :

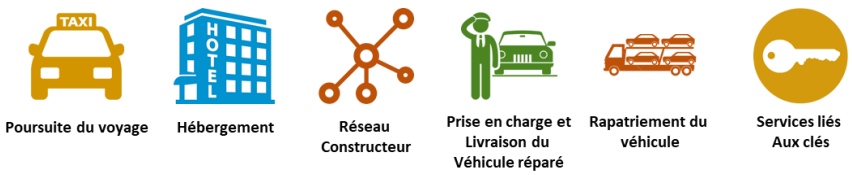


Figure 1.3 - Prestations secondaires de l'assistance routière

Des services supplémentaires peuvent être ajoutés sur demande pour répondre aux besoins spécifiques des clients (par exemple, une assistance médicale à l'étranger).

Des limites d'usage sont définies sur chacun des services inclus dans la garantie permettant de mitiger le risque de cout moyen et de fréquence. Les limites de service le plus courants sont :

- Rayon de distance maximal pour le dépannage et/ou le remarquage du véhicule.
- Distance maximale entre le lieu de la panne ou accident et la destination du bénéficiaire.
- Le montant maximal de frais d'assistance (e.g. taxi, billets d'avion/train, nuits d'hébergement, frais de rapatriement, etc.).
- Réseau de garagistes défini par la compagnie d'assistance.
- Nombre de jours de location du véhicule de remplacement.
- Durée de réparation minimale pour accéder à des services secondaires.

1.2. Modèles de distribution

La garantie d'assistance routière peut être commercialisée selon trois types de modèles de distribution : BtoB, BtoBtoC et BtoC. Dans les modèles BtoB et BtoBtoC, le client direct de la société d'assistance est une société partenaire qui est chargée de distribuer le produit d'assistance routière aux clients finaux. La garantie d'assistance routière est souvent associée à l'achat d'un véhicule, à la souscription d'une assurance automobile ou à un service fourni par la société partenaire. Dans le cas du modèle BtoC, la garantie d'assistance routière est vendue directement aux clients finaux par le biais de canaux de distribution directs.

Le type de modèle de distribution aura un impact direct sur la fréquence de sinistre. Les fréquences liées aux modèles facultatifs (produit optionnel pour le client final) seront plus élevées que celles des modèles en inclusion. Ceci est dû au risque d'antisélection intrinsèque aux modèles facultatifs ; ce sont surtout les profils ayant un risque de sinistre élevé qui choisiront de souscrire à la garantie d'assistance.

1.2.1. Modèle BtoB

Le modèle BtoB est un modèle en inclusion. Le bénéficiaire acquiert directement la garantie d'assistance lors de la souscription d'une police d'assurance, de l'achat d'un véhicule ou d'un service proposé par l'entreprise partenaire. La particularité de ce modèle est l'absence de risque d'antisélection. Par exemple, lorsqu'une personne achète un véhicule, la garantie d'assistance est incluse dans le prix de vente global du véhicule, qui est identique quel que soit le profil de risque de l'acquéreur. Elle n'est

pas optionnelle puisque celui-ci ne peut pas choisir d'acheter le véhicule sans la garantie. Ainsi, deux conducteurs ayant des profils de risque différents achetant le même véhicule vont payer le même prix, y compris pour la garantie d'assistance. Ce qui diffère de garanties d'assurance usuelles où la tarification est segmentée selon le profil de risque de l'assuré.

Ainsi, la difficulté de la tarification pour un modèle en inclusion réside dans la détermination du profil de risque "moyen" du portefeuille en amont de la mise en vente des véhicules, afin de définir un tarif non segmenté qui sera proposé à l'ensemble des acquéreurs. C'est un enjeu clé puisque si le profil de risque réel s'avère in fine moins bon que le profil de risque moyen évalué, il y a un risque d'avoir un ratio combiné supérieur à 100% et donc d'une perte globale sur le portefeuille.

1.2.2. Modèle BtoBtoC

Le modèle BtoBtoC est un modèle facultatif ou optionnel. Le client final a le choix d'inclure ou non la garantie d'assurance routière lors de la souscription d'une police d'assurance, de l'achat d'un véhicule ou d'un service proposé par la société partenaire sous condition du paiement d'une prime supplémentaire. Etant donné que ce type de modèle est soumis au risque d'antisélection, un tarif segmenté est proposé aux bénéficiaires.

1.2.3. Modèle BtoC

La garantie d'assistance routière est vendue directement aux clients finaux sur le site web de la compagnie d'assistance. Elle est souvent combinée à d'autres garanties telles que l'assistance médicale ou l'assurance voyage. Le risque d'antisélection observé dans le modèle optionnel BtoBtoC est également observé dans le modèle BtoC, par conséquent, les tarifs sont également segmentés dans ce modèle. Le modèle BtoC représente une part limitée du chiffre d'affaires du groupe Europ Assistance.

1.3. Segments client et produits d'assistance

Sur le marché de l'assistance routière, il existe plusieurs types de clients avec des caractéristiques différentes en termes de modèle de distribution, de produits et de niveau de risque. Parmi les principaux segments client se trouvent : constructeurs automobiles, assureurs, flottes d'entreprise/leasing, agences de location de véhicules, plateformes de covoiturage, marché BtoC, institutions financières, etc.

Le portefeuille d'assistance routière automobile du groupe Europ Assistance étant constitué à 95% d'assureurs et de constructeurs automobiles, ce mémoire se concentrera sur ces deux segments.

1.3.1. Constructeurs automobiles

Chaque constructeur automobile peut proposer une garantie d'assistance routière en plus de la garantie du constructeur[1]. La durée de la couverture est généralement la même que celle de la garantie du constructeur qui peut aller de 1 à 5 ans. Ce produit couvre principalement les pannes mais peut aussi inclure les accidents et les négligences. Europ Assistance a développé une forte expertise sur ce segment qui représente 15% du chiffre d'affaires global de la ligne métier automobile du groupe.

Les principaux produits de ce portefeuille sont :

- Assistance routière pour les nouveaux véhicules : Ce produit est distribué sous un modèle en inclusion BtoB. Cela signifie que c'est le constructeur automobile qui paie la prime et que ce sont ses clients directs, les acheteurs du véhicule, qui bénéficient de la garantie d'assistance pendant la durée de la garantie constructeur. Tous les clients finaux bénéficient d'une garantie d'assistance car elle est incluse dans la garantie constructeur qui est souscrite gratuitement à l'achat du véhicule.

- Extension de garantie – assistance routière : Dans certains cas, le constructeur automobile propose une extension de la garantie constructeur et donc de la garantie d'assistance routière, dont la durée peut aller de 1 à 3 ans à compter de la fin de la garantie du véhicule neuf. Cette extension de la garantie d'assistance routière peut être vendue au moment de l'achat du véhicule ou pendant la période de garantie avant la fin de la couverture standard. Ce produit est toujours optionnel pour l'acheteur du véhicule (modèle facultatif BtoBtoC). Comme mentionné dans la section précédente, cela a un impact direct sur la fréquence car, en raison de l'antisélection. En général, ce sont les personnes les plus susceptibles de tomber en panne ou d'avoir un accident qui décideront de prendre l'extension de garantie.

- Assistance routière activée par le service (SARA) : Afin de fidéliser les clients au sein de leur réseau d'ateliers de réparation, les constructeurs automobiles offrent une extension de l'assistance routière pendant un an, ou jusqu'au prochain entretien programmé du véhicule (selon ce qui se produit en premier) aux clients qui font entretenir leur voiture dans leur réseau. La couverture de ce produit implique généralement un âge maximal de la voiture à couvrir (c'est-à-dire un véhicule de 1 à N ans, généralement N<10). Ce produit contient les mêmes couvertures et avantages que l'assistance routière pour les nouveaux véhicules et le modèle de distribution est généralement en inclusion ce qui évite l'antisélection.

[1] La garantie constructeur couvre les coûts des pièces et des réparations effectuées en cas de défaut de construction du véhicule. Cette garantie est offerte par le constructeur automobile au propriétaire du véhicule au moment de l'achat.

1.3.2. Assureurs

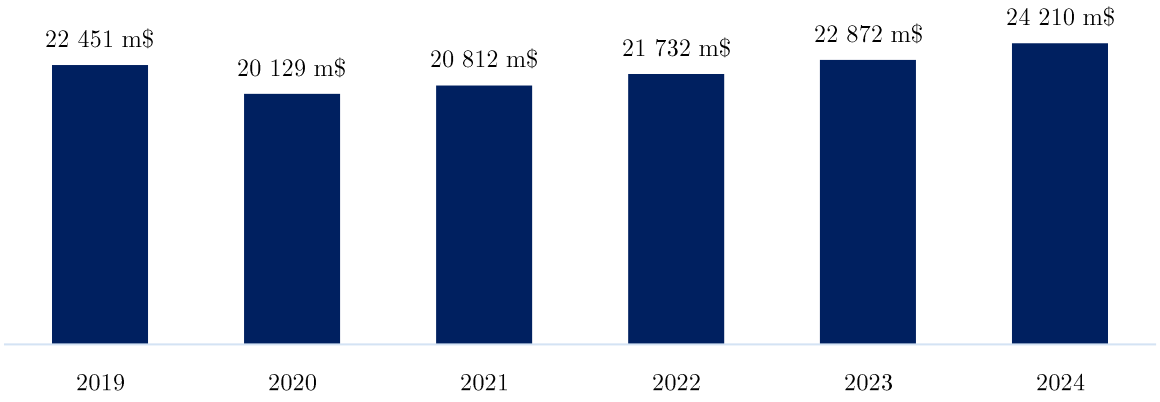
L'assistance routière est vendue aux assureurs et est incluse dans le contrat d'assurance automobile, dont la couverture est généralement d'un an. Le produit d'assistance peut être inclusif ou optionnel, cela dépend de la pratique du marché et de l'assureur concerné. Dans la plupart des cas, les *prestations de base* (dépannage sur place et remorquage) sont incluses et les autres prestations sont proposées en option.

L'impact sur la sinistralité dépend du marché, du niveau de communication et de l'âge du véhicule.

1.4. Contexte actuel du marché de l'assistance routière

Le marché global de l'assistance routière est en forte croissance grâce aux nouveaux besoins qui émergent autour des véhicules hybrides électriques, aux vieillissement du parc automobile, au développement de la digitalisation dans le secteur automobiles (véhicules connectés et autonomes). D'après l'étude de marché réalisée par l'entreprise de recherche Technavio, le chiffre d'affaires global[2] de l'assistance routière s'élevait à 22,4 milliards de dollars 2019 et il augmentera à 24,2 milliards en 2024 ce qui représente un taux composé de croissance annuel de 1,52% (CAGR).

Chiffre d'affaires global de l'assistance routière 2019-2024 (m$)



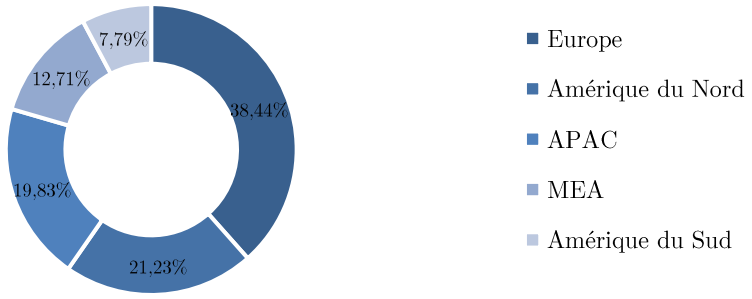
Graphique 1.1 - Chiffre d'affaires global de l'assistance routière 2019-2024 (Technavio Research Firm, 2020)

En termes de répartition géographique, l'Europe et l'Amérique du Nord ont les plus grandes parts de marché avec une contribution de 38,44% et 21,23% respectivement en 2019. En Europe les pays leaders sont le Royaume Uni, l'Allemagne, la France, l'Italie et l'Espagne alors qu'en Amérique du Nord les Etas Unis correspond au plus gros contributeur de la région. La répartition des parts de marché ne devrait

[2] Régions considérées : Europe, Amérique du Nord, Asie-Pacifique (APAC), Moyen Orient et Afrique (MOA) et Amérique du Sud

pas changer dans les années à venir, cependant, l'Europe et l'Amérique du Nord auront une croissance plus faible, respectivement 1,03% et 1,18% par an, que les marchés émergents.

Part de marché par zone géographique 2019 (%)

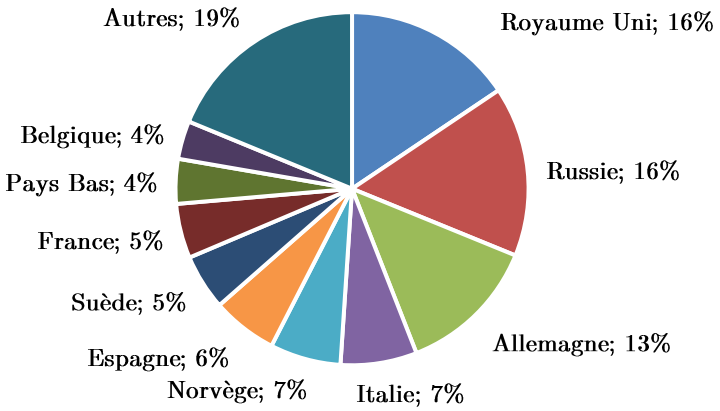


Graphique 1.2 - Parts de marché par zone géographique 2019 (Technavio Research Firm, 2020)

1.4.1. Le marché actuel de l'assistance routière en Europe

L'Europe a contribué à 38,44% du marché global en 2019 et diminuerait sa contribution à 37,52% du marché mondial de l'assistance routière en 2024 selon l'étude de marché réalisée par *Technavio*. Bien que l'Europe devrait être la région à la croissance la plus lente, ses revenus passeront de 8 630 millions de dollars en 2019 à 9 082 millions de dollars en 2024, soit une augmentation globale de 5,2 % et de 1,03 % par an. Le marché est distribué sur 20 pays avec le Royaume Uni, la Russie, l'Allemagne et l'Italie comme plus grosses régions avec un participation respective de 16%, 16%, 13% et 7% respectivement en 2018 (Finaccord, 2019). Ces quatre pays représentent plus de 50% du marché européen de l'assistance routière.

Parts de marché de l'assistance routière par pays en Europe (2018)



Graphique 1.3 - Parts de marché de l'assistance routière par pays en Europe (2018) (Finaccord, 2019)

L'assistance routière est un marché très fragmenté avec des nombreux acteurs qui ont chacun des parts de marché réduites. Les trois filiales d'assistance des trois plus gros groupes d'assureurs européens : Allianz Partners (Groupe Allianz), Axa Assistance (Groupe Axa) et Europe Assistance (Groupe Generali) ne contribuaient qu'à hauteur de 28,7% en 2018. Néanmoins, lorsque des partenariats tels qu'ARC Europe ou ERA Automotive (dont Europ Assistance fait partie) sont considérés comme une seule entité, le marché est plus consolidé et ARC Europe[3] ressort comme la compagnie d'assistance disposant de la plus grande part de marché (32,5%).

Allianz Partners	11,0%
Axa Assistance	10,3%
Europ Assistance	7,4%
RAF	6,5%
Viking	6,3%
ADAC	5,6%
VW Group	5,6%
Falck	3,5%
RAC	3,5%
Inter Mutuelles Assistance	3,2%
Others	37,2%

ARC Europe	32,5%
ERA Automotive	14,4%
Allianz Partners	11,0%
RAF	6,5%
Viking	6,3%
VW Group	5,6%
Inter Mutuelles Assistance	3,2%
Axa Assistance	2,6%
Sampo Group	2,5%
BMW Group	2,5%
Others	13,1%

Tableau 1.1 - Parts de marché de l'assistance routière par compagnie d'assistance en Europe (2018) (Finaccord, 2019)

Le vieillissement du parc automobile et l'augmentation de la part des véhicules électriques et hybrides sont les principaux facteurs qui stimulent la croissance du marché de l'assistance routière dans la région.

D'après l'ACEA[4], l'âge moyen des voitures de l'Union Européenne a augmenté de 6,3% entre 2017 et 2021 passant de 11,1 ans à 11,8 ans. La Lituanie et la Roumanie possèdent les parcs automobiles les plus anciens, avec des véhicules de près de 17 ans d'ancienneté moyenne. Les voitures les plus récentes se trouvent au Luxembourg (6,7 ans). Le tableau ci-dessous montre la moyenne d'âge les pays contribuant le plus au marché européen de l'assistance routière.

Pays	Moyenne d'âge (ans)
Royaume Uni	8,6
Allemagne	9,8
France	10,3
Italie	11,8
Espagne	13,1

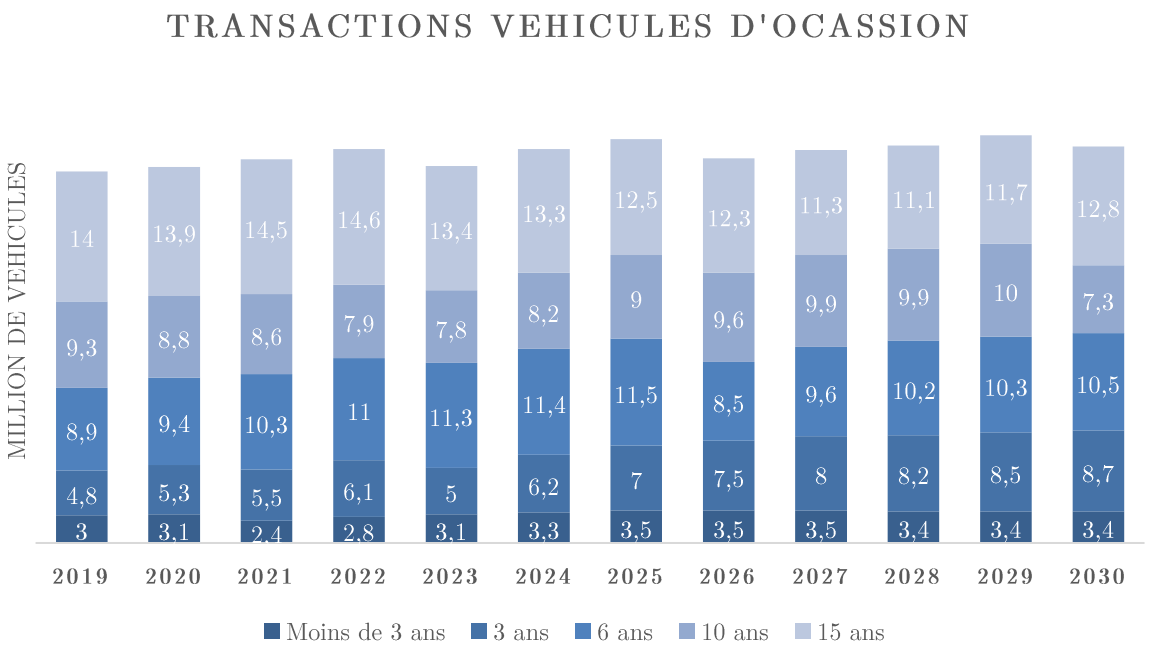
Tableau 1.2 - Âge moyen du parc automobile par pays (ACEA, 2022)

[3] ARC Europe est intégré par 8 clubs automobiles en Europe : AA (Royaume Uni), ACI(Italie), ADAC (Allemagne), ANWB (Pays Bas), ÖAMTC (Autriche), RACE (Spain), TCB (Belgique), TCS (Suisse).

[4] ACEA : Association des constructeurs européens d'automobiles

L'augmentation de l'âge moyen des véhicules en Europe s'explique par une progression faible des ventes de voitures neuves. Le nombre de voitures neuves vendues entre 2018 et 2019 a augmenté de seulement 1,28% d'après l'étude réalisée par Observatoire Cetelem (PWC, 2020). De plus, le marché a été fortement impacté par la crise Covid-19 qui a entrainé une diminution des ventes de 27% et dont la reprise était attendue pour 2023 (Brian Collie, 2020). Cependant, une série des nouvelles crises a empêché la reprise. Les pénuries d'approvisionnement, la guerre en Ukraine et le contexte inflationnaire croissant ont maintenu le marché automobile sous pression. Dans un premier temps, c'est l'offre qui a été la plus touchée, mais aujourd'hui, les inquiétudes concernant la baisse de la demande se font plus pressantes. Ainsi, l'ACEA prévoit que le marché européen des voitures neuves se contractera de 26 % cette année, par rapport au niveau de 2019, avant la pandémie.

Par conséquent, le marché de véhicules d'occasion grandit plus rapidement que celui des véhicules neufs et passera de 40,1 millions de véhicules vendus en 2019 à 42,8 millions en 2030 (+6,73%) surtout dans le segment de moins de 3 ans d'âge moyen (IHS, ACEA, Autovista Group simulation, 2020).



Graphique 1.4 - Transactions de voitures occasion sur la période 2019 - 2030 (IHS, ACEA, Autovista Group simulation, 2020)

En ce qui concerne les véhicules électriques (BEV) et hybrides (HEV[5] et PHEV[6]), une forte croissance de ventes a été observée ces dernières années. Au premier trimestre 2022, les immatriculations de véhicules hybrides (HEV) ont augmenté de 20 % atteignant 25,1 % des ventes totales dans l'UE. Les véhicules électriques (BEV) ont presque doublé leur part de marché par rapport à la même période en 2021, ils représentent désormais 10,0 % de l'ensemble des ventes totales dépassant la part de

[5] *Hybrid Electric*

[6] *Plug-in Hybrid*

marché des véhicules hybrides rechargeables (PHEV), qui représentent 8,9 % du marché de l'UE. En revanche, les véhicules à moteur thermique ont continué à perdre des parts de marché. Néanmoins, les immatriculations de voitures traditionnelles à carburant dominent toujours le marché, avec une part combinée de 52,8% (ACEA, 2022).

L'augmentation de la demande de véhicules électriques et hybrides est due aux incitations réglementaires dans les pays européens pour promouvoir les véhicules à faible émission de CO_2. La plupart des Etats membres de l'UE offrent désormais une forme d'aide fiscale ou financière pour stimuler l'adoption des véhicules électriques sur le marché. Par conséquent, le soutien gouvernemental favorise l'utilisation de véhicules électriques et hybrides, ce qui entraîne le besoin de progrès dans les services d'assistance routière.

Les tableaux ci-dessous montrent l'évolution des immatriculations des véhicules électriques et hybrides dans les pays contribuant le plus au marché européen de l'assistance routière.

Pays	Q1 2022	Q1 2021	Var.
Royaume Uni	64 165	31 779	102%
Allemagne	83 774	64 809	29%
France	43 510	30 491	43%
Italie	11 289	13 272	-15%
Espagne	7 253	3 449	110%

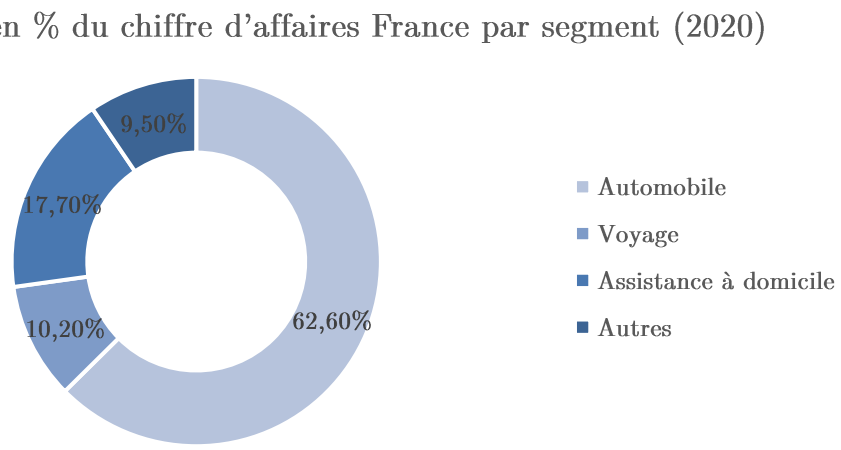
Tableau 1.3 - Evolution des immatriculations de véhicules électriques Q1 2022 - Q1 2021 (ACEA, 2022)

Pays	Q1 2022	Q1 2021	Var.
Royaume Uni	151 940	135 960	12%
Allemagne	189 312	179 373	6%
France	105 119	101 994	3%
Italie	132 216	136 698	-3%
Espagne	57 122	48 171	19%

Tableau 1.4 - Evolution des immatriculations de véhicules hybrides Q1 2022 - Q1 2021 (ACEA, 2022)

1.4.2. Le marché actuel de l'assistance routière en France

L'automobile représente plus de la moitié des interventions des sociétés d'assistance en France. Elle contribue à 63% du chiffre d'affaires global de l'assistance et 51,5% des dossiers traités. D'après le Syndicat National des Sociétés d'Assistance (SNSA), le chiffre d'affaires de l'assistance routière en 2020 s'est élevait à 2,157 milliards d'euros, ce qui représente une légère augmentation de 0,5% par rapport à 2019, même dans une période de crise économique pendant laquelle le chiffre d'affaires global sur tous les segments d'assistance a reculé de 2,6%.



Graphique 1.5- Part en % du chiffre d'affaires en France par segment (2020) (Vernet & Grafe, 2021)

Le marché français des véhicules électriques/hybrides est également en pleine croissance. Le nombre d'immatriculations de ce type de véhicules a augmenté de 206% en 2020 par rapport à l'année 2019. Ce phénomène a également entrainé une tendance à la hausse du volume de pannes. Les batteries de ces véhicules sont plus sensibles aux températures extrêmes et au manque d'utilisation comparativement aux véhicules thermiques. Par conséquence, des augmentations significatives du nombre de dossiers ont été observés en période de grand froid et post-confinement. Les sociétés d'assistance ont beaucoup d'expérience avec les véhicules thermiques mais moins avec les véhicules rechargeables. Selon le SNSA : « *plusieurs initiatives ont été déclenchées dans le secteur tels que la formation des dépanneurs, le déploiement de plateformes dédiées pour assister les clients sur les questions les plus fréquentes ou sur les terminaux disponibles les plus proches.* » pour faire face aux nouveaux besoins des bénéficiaires et réussir à transformer le secteur de l'assistance automobile.

1.4.3. Focus sur l'assistance routière chez le groupe Europ Assistance

Pionnière et leader mondial de l'assistance, Europ Assistance a étendu sa présence à l'international au cours des années comptant actuellement 41 centres d'assistance au niveau mondial, 8000 employés et un réseau d'environ 750 000 prestataires d'assistance médicale, voyage et routière. Depuis le début de son activité en 1963, la compagnie d'assurance d'origine italienne Generali faisait partie des actionnaires. Depuis 2001, le Groupe Generali a acquis 100% des parts du Groupe Europ Assistance.

Sur le marché européen, Europ Assistance est aujourd'hui l'une des plus grandes sociétés d'assistance opérant à travers des modèles de distribution BtoB, BtoBtoC et BtoC. Elle est présente actuellement sur 5 différentes lignes métiers : Automobile, Voyage, Domicile, Santé et Conciergerie. Son chiffre d'affaires s'élevait à 1,9 millions d'euros en 2021 avec 37 millions d'appels traités et 10 millions d'interventions réalisées dont 5,3 millions liées au secteur automobile sur plus de 200 pays et territoires.

En 2021, le secteur automobile représentait 46% du chiffre d'affaires dont l'assistance routière automobile contribuait à hauteur de 76%. Les principaux pays européens contribuant le plus à l'activité de la ligne métier automobile sont l'Italie, la France et l'Espagne. Ces trois pays participent à plus de 50% du chiffre d'affaires global de l'assistance routière en Europe.

Le plan stratégique de la ligne métier automobile 2021-2025 chez Europ Assistance vise principalement à faire face aux 6 tendances techniques et environnementales sur le marché de l'automobile, de la mobilité et de l'assistance routière :

1. La croissance rapide du déploiement des voitures électriques et hybrides encouragée non seulement par les réglementations et initiatives gouvernementales (interdiction du diesel dans les villes, objectifs CAFE de réduction des émission CO2 pour les voitures neuves…) mais également par l'ambition des constructeurs automobiles de rentrer sur ce marché : 143 modèles ont été lancés en 2019 (105 BEV et 38 PHEV) au niveau mondial et l'Europe est considérée comme la région avec le plus fort taux de pénétration de marché (Thomas Gersdorf, 2020).

2. Incursion du marché des véhicules autonomes en Europe grâces aux systèmes ADAS[7].

3. La croissance des véhicules connectés dont la barre dépassera le 50% du parc automobile dans l'Union Européenne d'ici en 2025 (PWC, 2020).

4. La consolidation du segment de constructeurs automobiles. Le groupe PSA et FCA créent en 2021 Stellantis, un leader de la mobilité durable. Le constructeur automobile chinois Geely discute d'une fusion complète avec Volvo, afin de créer le premier constructeur automobile chinois d'envergure mondiale.

5. Evolution de l'assistance routière vers une vision moins centrée sur la voiture mais plus centrée sur l'expérience client.

6. Diminution prononcée du nombre de propriétaires de véhicules, en particulier dans les zones urbaines : les consommateurs et les entreprises délaissent la possession d'un véhicule au profit de services (leasing, covoiturage, location de courte durée…).

[7] Advanced Driver Assistance Systems

1.5. Les enjeux du mémoire

1.5.1. Processus de souscription pour des contrats internationaux

Le groupe Europ Assistance possède au sein de la ligne métier Automobile un portefeuille de contrats dits internationaux fournissant des services d'assistance automobile à des entreprises partenaires (schéma BtoB) présentes dans plusieurs pays d'Europe et/ou du monde. Au sein de ce portefeuille, le segment client le plus représenté correspond aux constructeurs automobiles présents en Europe. Les constructeurs automobiles proposent à leur clients (acquéreur d'un véhicule) une garantie d'assistance routière automobile avec une période de couverture qui peut aller de 1 à 5 ans selon le contrat.

Etant donné qu'Europ Assistance a une connaissance approfondie de son réseau de partenaires fournissant chacun des services d'assistance et que les contrats internationaux actuels font partie du portefeuille depuis plusieurs années, l'entreprise est en mesure d'estimer la prime pure par véhicule de chacun des contrats historiques se basant uniquement sur certains critères tels que la marque du véhicule, le pays, le type de couverture et les caractéristiques de services proposés. Le portefeuille actuel est composé de trois marques principales, ayant des véhicules présentant des caractéristiques relativement similaires. L'estimation de la prime pure devient plus compliquée lorsqu'il s'agit d'un nouveau contrat pour lequel l'historique des sinistres n'est pas disponible ou lorsque les caractéristiques des véhicules commercialisés par les constructeurs déjà présents dans le portefeuille évoluent. C'est notamment le cas ces dernières années avec la croissance des véhicules rechargeables. Dans ces deux cas, la sinistralité moyenne du portefeuille actuel n'est pas adaptée pour prévoir la sinistralité future.

Lorsqu'un constructeur automobile cherche à souscrire une garantie d'assistance, il démarre un processus d'appel d'offres où il décrit de manière détaillée les spécifications et les conditions de la garantie d'assistance qu'il souhaite souscrire. De même, il peut fournir aux compagnies d'assistance participantes certaines informations leur permettant de partiellement caractériser le futur portefeuille :

- Catégorisation des produits (garantie standard, extension de garantie et/ou SARA) ;
- Fait générateur de sinistre (panne et accidents majoritaires) ;
- Services inclus et limites ;
- Chiffres clés (prévision de ventes, nombre de véhicules actuellement couverts, fréquence globale par produit et marché)

Le choix du fournisseur de la garantie d'assistance par les constructeurs automobiles se base principalement sur le prix le plus avantageux à la suite de

plusieurs tours de négociation. Pour cela, il est essentiel pour la société d'assistance d'estimer au mieux la prime pure avec l'information disponible afin de piloter au mieux la marge cible tout en restant compétitif vis-vis des autres concurrents. Cependant, des données fiables et de bonne qualité permettant de caractériser d'une manière complète le nouveau portefeuille ne sont pas toujours disponibles, principalement en raison des deux points suivants :

- L'asymétrie d'information entre la compagnie d'assistance et son futur client. Le constructeur automobile ne fournit pas toujours des informations précises et essentielles pour définir le prix de la garantie assistance (par exemple la fréquence empirique par marché et par produit, le pourcentage de véhicules hybrides électriques, la prévision de ventes).

- Le portefeuille actuel très homogène ne disposant qu'un certain type de véhicule, il n'est donc pas adapté pour estimer la prime pure du nouveau portefeuille.

De plus, le modèle collectif de tarification sans segmentation appliqué actuellement au sein de la ligne métier automobile, ne permet pas d'identifier avec précision les variables qui ont réellement un impact sur la prime de risque.

L'enjeu de ce mémoire est d'utiliser un portefeuille présentant des caractéristiques plus hétérogènes afin d'améliorer l'estimation de prime pure pour des nouveaux contrats internationaux et/ou des changements de typologie de véhicule au sein du portefeuille actuel. Néanmoins, les contrats internationaux sont souvent distribués dans un modèle BtoB, qui est un modèle en inclusion. Dans ce cas, comme expliqué en section 1.2.1, l'assureur n'est pas exposé au risque d'antisélection. L'objectif n'est donc pas de calculer une prime pure segmentée mais de mieux identifier les caractéristiques du nouveau portefeuille et de calculer une prime de risque moyenne adaptée au risque souscrit.

1.5.2. Les assureurs : un portefeuille riche et hétérogène

Les portefeuilles appartenant au segment client des assureurs sont caractérisés comme étant très hétérogènes et riches en termes de variables. Comme la couverture d'assistance est intégrée à une assurance automobile vendue à tout type de véhicule et de conducteur, l'assureur est en mesure de collecter plusieurs informations au moment de la souscription, ce qui lui permet de tarifer au mieux l'assurance automobile. Dans certains cas, les données collectées par l'assureur peuvent également être récupérées par la compagnie d'assistance, ce qui facilite la mise en place un modèle de tarification plus sophistiqué.

Le segment de clientèle des assureurs n'est actuellement pas présent dans le portefeuille de contrats internationaux du groupe Europ Assistance. En revanche, ils sont présents au sein des filiales parmi le portefeuille de contrats locaux, c'est-à-dire, les contrats pour lesquels l'assistance routière est fournie dans un seul pays et qui sont

souscrits et gérés directement par la filiale avec la société partenaire présente dans le pays de référence.

Une base des données assureur contenant les variables déclaratives à la souscription ainsi que l'historique des sinistres sur trois années a été mise à disposition par la filiale française d'Europ Assistance. Elle comporte dans son état original 21 021 259 lignes et 82 variables.

Les principaux objectifs de ce mémoire seront donc de :

- Qualifier le portefeuille d'assurance de l'entité Europ Assistance France à l'aide d'analyses exploratoires.
- Déterminer quelles variables explicatives et quelles interactions sont les plus pertinentes pour estimer la fréquence et la sévérité des prestations d'assistance de base.
- Modéliser la fréquence et la sévérité des deux faits générateurs (pannes et accidents) en utilisant à l'aide des modèles linaires généralisés (GLM).
- Choisir les modèles prédictifs le plus adaptés en termes d'interprétabilité et de complexité.
- Estimer une nouvelle prime pure par marque de véhicule sur la base des résultats de la modélisation et la comparer à celle estimée par un modèle de tarification simple.

Cette étude permettra d'identifier les facteurs de risque qui ont un réel impact sur la fréquence et la sévérité, de capter l'impact des nouvelles tendances potentielles du marché automobile sur la sinistralité future, notamment celui des véhicules rechargeables (hybrides et électriques), et de mieux gérer la rentabilité des portefeuilles « BtoB ».

2. Présentation de la base de données

L'objectif de ce chapitre est de caractériser la base de données disponible dans le cadre du mémoire. Une analyse chiffrée permettra de qualifier le portefeuille en termes de fréquence et de coût moyen dans son ensemble. Après la définition du périmètre de l'étude, les ajustements effectués sur la base de données, la vérification de la stabilité du portefeuille dans le temps et les analyses exploratoires des variables explicatives seront présentés.

2.1. Description

La base de données principale utilisée pour l'étude correspond à un portefeuille d'assurance automobile avec une couverture d'assistance routière fournie par la filiale Europ Assistance France sur la période de survenance de 2017 à 2019. Elle contient à la fois des caractéristiques de l'assuré et du véhicule collectées par l'assureur partenaire ainsi que des informations sur la sinistralité collectées par Europ Assistance France. Cette base est produite à partir d'une jonction de deux bases : la base ventes et la base sinistres, la première contenant les variables déclaratives à la souscription et la deuxième contenant le nombre et le coût des sinistres.

Pour fusionner ces deux bases, il a d'abord fallu calculer l'exposition au risque pour chaque véhicule par année de survenance. Il s'agit du rapport entre l'intervalle de temps pendant lequel le véhicule a été exposé au risque durant l'année de survenance et la période totale de couverture du véhicule.

Équation 2.1: L'exposition du véhicule A sur l'année de survenance N.

$$Exp_A^N = \frac{min(31/12/N;\ Date\ fin\ de\ couverture) - max(date\ début\ couverture;\ 01/01/N)}{365}$$

En ce qui concerne les sinistres, ils peuvent être regroupés par fait générateur (accident ou panne) et par service d'assistance fourni (services de base[8], véhicule de remplacement et poursuite du voyage[9]). Un compteur des sinistres survenus et la somme du coût des sinistres par fait générateur et service d'assistance fourni a été défini et ensuite rattachés à chaque véhicule par année de survenance.

Le tableau 2.1 montre les variables explicatives liées à l'assuré ainsi qu'au véhicule disponibles dans la base de données initiale.

[8] Remorquage et dépannage sur place

[9] Taxi ou train

Assuré	Véhicule
Ancienneté de permis du conducteur principal	Identifiant SRA du véhicule
Ancienneté de permis du conducteur secondaire le plus aggravant	Code de la commune du lieu de stationnement du véhicule
Profession du conducteur principal	Formule de garanties fournie Europ Assistance France
Tranches de coefficient bonus-malus	Groupe SRA du véhicule
Formule de garanties fournie par l'assureur	Ancienneté de commercialisation du véhicule
Offre 8000 km (le conducteur s'engage à faire moins de 8000 km par an)	Ancienneté d'acquisition du véhicule
	Carrosserie / Silhouette
	Alimentation / Carburant
	Marque constructeur
	Usage fait du véhicule (privé ou professionnel)
	Type de garage pour le véhicule

Tableau 2.1 - Variables de segmentation initiales liées à l'assuré et au véhicule

2.1.1. Identifiant SRA

L'association française SRA (Sécurité et Réparation Automobiles) a été fondée en 1977 dans le but de fournir aux membres des informations détaillées sur les véhicules immatriculés ainsi que de mettre en œuvre plusieurs études et statistiques pouvant contribuer à limiter la fréquence et le coût des sinistres automobiles (Sécurité et Réparation Automobiles, 2022).

Cette association a mis en place un fichier qui recense un grand nombre de caractéristiques techniques et commerciales des véhicules (4, 3 et 2 roues de moins de 3,5 tonnes) et qui est régulièrement mis à jour en fonction des nouveaux véhicules apparaissant sur le marché automobile français. Un véhicule peut être retrouvé dans ce fichier grâce à l'identifiant SRA qui correspond à un code d'identification unique associé à un seul modèle de véhicule. Il est composé de lettres et de chiffres représentant la marque, le modèle et la version du véhicule. Par exemple l'identifiant *SRA FI49037* correspond à une *Fiat 500L 1.6 MJT 120 Trekking*. Le système d'identification des véhicules mis en place par le SRA permet aux assureurs de connaître de manière précise les caractéristiques d'un véhicule et de proposer des garanties et des tarifs adaptés.

L'identifiant et le fichier SRA ont été utilisés pour enrichir la base de données initiale avec des variables explicatives supplémentaires liées aux caractéristiques du véhicule. Ainsi, **la base de données finale utilisée pour cette étude comporte 21 021 259 lignes et 82 variables**, dont 37 sont des variables catégorielles, 31 sont des variables numériques (discrètes et continues), 6 correspondent au nombre de sinistres par fait générateur et par prestation fournie, 6 correspondent au montant des sinistres par fait générateur et par prestation fournie, et 2 sont des variables d'identification.

L'annexe A.1 présente la description des 82 variables de la base de données utilisée pour l'étude.

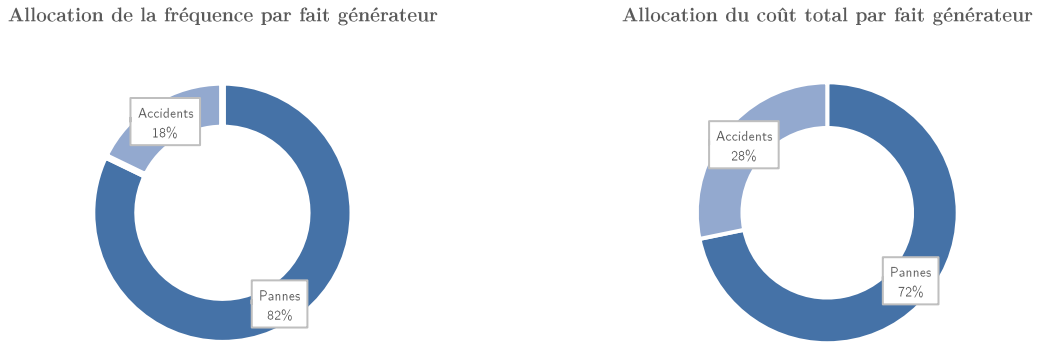
2.1.2. Chiffres clés

Le tableau 2.2 montre quelques chiffres clés apportant une vision globale du portefeuille en termes d'exposition et de sinistralité :

Année	Exposition	Fréquence[10]	Coût Global	Coût moyen par sinistre
2017	599 186	9,25%	9 121 553 €	164,49 €
2018	621 796	8,89%	7 021 970 €	126,99 €
2019	655 828	7,76%	6 425 797 €	126,30 €
Total	**1 876 810**	**8,61%**	**22 569 320€**	**139,64 €**

Tableau 2.2 - Chiffres clés concernant l'exposition et la sinistralité du portefeuille

Les pannes sont le fait générateur des sinistres le plus représenté en termes de coût total et de fréquence dans le portefeuille. Elles représentent 82% de la fréquence globale et 72% du coût total sur la période étudiée, tandis que les accidents représentent respectivement les 18% et 28% restants (voir graphique 2.1).



Graphique 2.1 - Allocation du coût total par fait générateur

Une analyse par fait générateur ainsi que par prestation fournie est présentée ci-dessous apportant une vision plus détaillée sur la sinistralité du portefeuille :

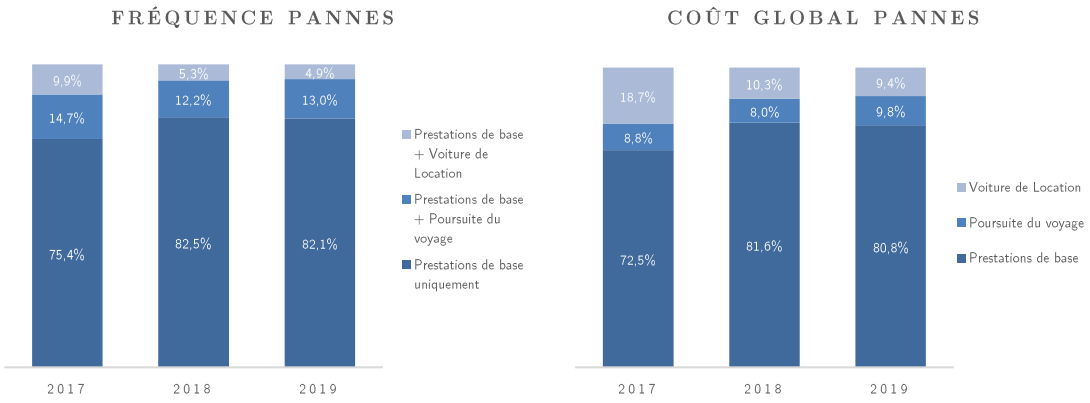
Pannes

Le tableau 2.3 montre les chiffres clés annuels en termes de fréquence et de coût pour les sinistres liés aux pannes. De même, le graphique 2.4 comporte la répartition du nombre de sinistres et du cout total par type de prestation :

[10] La fréquence est calculée comme le rapport entre le nombre de sinistres ayant eu une *prestation de base* et l'exposition. L'exposition a été définie dans l'équation 2.1.

Année	Fréquence	Coût Global	Coût moyen par sinistre
2017	7,55%	6 241 574 €	138,06 €
2018	7,34%	5 234 855 €	114,65 €
2019	6,40%	4 724 346 €	112,63 €
Total	**7,08%**	**16 200 776 €**	**121,98 €**

Tableau 2.3 - Chiffres clés concernant la sinistralité liée aux pannes



Graphique 2.2- Répartition de la fréquence et du coût des pannes par type de prestation

Sur un portefeuille d'assistance routière automobile, le dépannage sur place et le remorquage sont des prestations dites *de base* car ce sont les prestations minimales fournies à tous les bénéficiaires de la garantie d'assistance en cas de sinistre. Toutes les garanties d'assistance s'engagent à fournir à minima ces deux services. Cela veut dire que tous les sinistres sont concernés par ce type de prestation. Comme mentionnée dans la section 1.1.2 de ce document, des prestations secondaires peuvent être proposées, en fonction des conditions spécifiques du contrat, afin d'assurer la mobilité du bénéficiaire. Les prestations secondaires les plus représentées de ce portefeuille sont la poursuite du voyage pouvant correspondre à un taxi ou un train, et la voiture de location proposée en complément des prestations de base.

Dans le cas des pannes, les *prestations de base* seules sont les prestations les plus représentées que ce soit en termes de fréquence ou de coût avec une contribution de plus de 70% (cf. graphique 2.2). Cela s'explique par deux raisons principalement :

- Les prestations secondaires ne sont pas incluses dans toutes les formules de vente. Ce portefeuille contient 4 formules d'assistance :
    - Formule Socle (67% de l'exposition du portefeuille étudié) : Formule d'assistance proposant uniquement les *prestations de base* et d'autres prestations secondaires qui ne sont pas abordées dans cette étude.
    - Formule Economique, Particulier ou Professionnel (33% de l'exposition du portefeuille étudié) : Formules d'assistance offrant les mêmes prestations que la formule Socle, plus les prestations

secondaires de poursuite du voyage et de véhicule de remplacement. Chacune de ces formules diffère par le type et le nombre de jours de location du véhicule de remplacement.

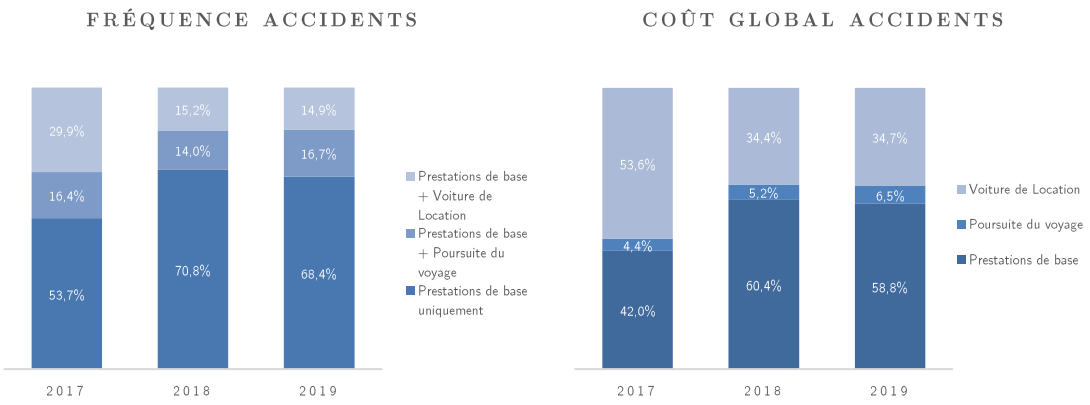
- Lorsque la voiture est réparée sur le lieu de la panne, une prestation secondaire n'est pas nécessaire.

Accidents

Le tableau 2.4 montre les chiffres clés annuels en termes de fréquence et de coût pour les sinistres liés aux accidents. De même, le graphique 2.4 comporte la répartition du nombre de sinistres et du cout total par type de prestation :

Année	Fréquence	Coût Global	Coût moyen par sinistre
2017	1,71%	2 879 979 €	281,14 €
2018	1,55%	1 787 115 €	185,48 €
2019	1,36%	1 701 450 €	190,47 €
Total	**1,54%**	**6 368 544 €**	221,04 €

Tableau 2.4 - Chiffres clés concernant la sinistralité liée aux accidents



Graphique 2.3- Répartition de la fréquence et du coût des accidents par type de prestation

D'après le tableau et les graphiques ci-dessus, les accidents sont des sinistres à faible fréquence mais à forte gravité. La fréquence d'un accident est inférieure de 78% à celle d'une panne en moyenne ; néanmoins, le coût moyen par sinistre est supérieur de 81 %. En ce qui concerne la répartition du nombre et du coût des sinistres par service, on constate qu'elle est différente de celle des pannes. Le poids des *prestations de base* est plus faible, ce qui représente un poids plus élevé des services secondaires, notamment le véhicule de remplacement. En effet, en cas d'accident, les véhicules sont difficilement réparables sur place et, étant donné que le temps de réparation peut prendre plusieurs jours, un véhicule de remplacement doit être fourni au bénéficiaire pour assurer sa mobilité. Cela a le plus grand impact sur le coût moyen du sinistre, puisque le véhicule de remplacement représente le service le plus onéreux parmi les trois types de services étudiés. Néanmoins, les *prestations de base* restent le principal

facteur contribuant au coût total des accidents, en particulier pour les années 2018 et 2019.

Cette étude se concentrera uniquement dans l'analyse et la modélisation des *prestations de base* liées aux pannes et aux accidents.

2.2. Retraitements

Comme la base de données est produite à partir de la jonction de plusieurs bases de données sous-jacentes, certains problèmes de qualité des données ont été identifiés. En outre, la taille importante de la base de données a entraîné un temps d'exécution assez long pour l'analyse et la modélisation. Certains ajustements ont été effectués afin d'avoir une base de données plus facile à gérer et des données plus précises.

2.2.1. Voluminosité

La base de données initiale fournie par l'assureur partenaire étant très volumineuse avec plus de 21 millions de lignes, il était difficile d'effectuer l'analyse et la modélisation en raison des restrictions de mémoire et de temps d'exécution. L'un des moyens trouvés pour réduire la taille de la base de données et faciliter son exploitation a été de regrouper les couvertures mensuelles par client et par année de survenance sur une seule couverture annuelle en une seule ligne. Cela a permis de réduire considérablement la taille de la base de données, de 21 021 259 à 2 628 356 lignes.

De même, réduire le périmètre de l'étude a également permis de réduire le nombre des lignes :

- Les polices appartenant à la catégorie « autres » de la variable *marque* ont été retirées afin de centrer l'étude sur les marques de véhicules les plus représentées dans le portefeuille.
- Seuls les véhicules à usage privé et professionnel ont été pris en compte. Les véhicules à usage tournée régulière ont été exclus, il ne représentait que 0,95% de l'exposition.
- Comme cette étude se concentre sur le segment client des constructeurs automobiles dont les polices ne couvrent que les véhicules de moins de 10 ans, les polices pour les véhicules de plus de 10 ans ont également été retirées.

Suite à ces ajustements, le jeu de données de l'étude comporte 1 192 195 lignes. Néanmoins, la base de données reste volumineuse, notamment en raison des 83 variables explicatives.

2.2.2. Valeurs Manquantes

La complétude des données est un aspect essentiel dans la construction d'une base de données fiable. Cependant, la gestion des valeurs manquantes est souvent un problème incontournable dans l'analyse et modélisation des données. Dans ce cas, il est apparu que certaines variables fournies par la base de données SRA étaient incomplètes.

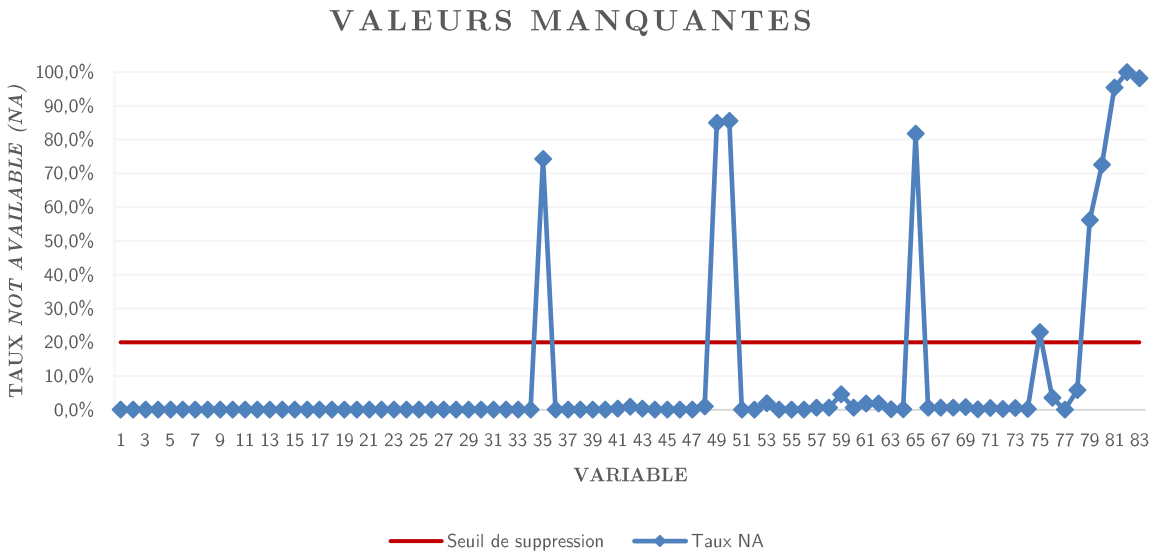
Deux méthodes sont possibles dans le processus d'analyse d'une base de données incomplète : supprimer les individus pour lesquels il y a des valeurs manquantes ou trouver un moyen d'affecter ces valeurs. Cette dernière méthode sera toujours privilégiée car elle préserve les informations disponibles.

L'approche de gestion et d'affectation des valeurs manquantes du jeu des données de cette étude a été appliquée en trois étapes :

1. Suppression des variables avec un taux *Not Available (NA)* supérieur à 20% pour éviter d'intégrer un biais important aux analyses ainsi qu'à la modélisation (9 variables supprimées).

2. Affectation des valeurs manquantes dans variables quantitatives par l'analyse de composantes principales.

3. Affectation des variables manquantes dans les variables qualitatives (catégorielles) en créant une nouvelle catégorie indépendante des autres.



Graphique 2.4 - Taux de Not avalaible (NA) par variable

Dans les paragraphes suivants, la deuxième étape du traitement des valeurs manquantes sera détaillée :

Méthodes d'analyse factorielle

L'analyse factorielle est une technique statistique de réduction des données utilisée pour expliquer les corrélations entre les variables observées en termes d'un plus petit nombre de variables non observées appelées facteurs. Les variables observées sont modélisées comme des combinaisons linéaires de facteurs plus des expressions d'erreur. Cette méthode est apparue en 1904 avec la théorie du facteur commun de Charles Spearman postulant que l'intelligence des individus pouvait être ordonnée selon une seule dimension. En France, elle ne s'est pas développée que dans les années 1930 avec les travaux de Jean-Paul Benzécri, dont l'apport concerne notamment les aspects géométriques et les représentations graphiques.

Silvia Bucci s'appuie sur ces techniques et présente dans son mémoire d'actuariat deux approches d'utilisation de l'analyse factorielle :

- L'**approche descriptive non supervisée** basée sur un modèle géométrique. Elle a comme but de proposer un nouveau système de représentation des variables latentes formées de combinaisons linéaires des variables explicatives, qui permettent de distinguer au mieux les groupes d'individus. Dans cette optique, une analyse graphique peut être réalisée à partir des composantes principales étant calculées sur les centres de gravité conditionnels des nuages de points. Cette approche sera utilisée dans le processus de pré-sélection des variables pour la modélisation de la fréquence et la sévérité.

- L'**approche décisionnelle supervisée** qui est plus récente et qui cherche à prédire le groupe d'appartenance d'un individu à partir des valeurs prises par le variables explicatives. En ce sens, cette technique est plus proche des techniques supervisées de l'apprentissage automatique. En pratique, la fonction de classement est représentée par une combinaison linéaire de variables explicatives, ce qui la rend facilement interprétable. Cette approche sera utilisée comme méthode d'affectation des valeurs manquantes des variables quantitatives.

Analyse des composantes principale (ACP)

L'ACP est une technique de réduction de dimension qui permet l'exploration et la visualisation d'un tableau d'individus (une police dans cette étude) par des variables quantitatives. Classiquement, l'ACP se présente comme la projection des individus sur un sous-espace de dimension inférieur. L'objectif est de trouver un sous-espace qui maximise la variance, également appelée inertie, des points projetés, autrement dit le sous-espace qui représente le mieux la diversité des individus. De manière équivalente, l'ACP peut être présentée comme la recherche du sous-espace qui minimise l'erreur de

reconstruction, c'est-à-dire la distance entre les individus et leur projection (Husson, Josse, & Pagès, 2009).

Soit X la matrice centrée à étudier de dimension $I\,x\,K$, $x_{i.}$ la ligne i, $x_{.k}$ la colonne k et $\|A\| = \sqrt{tr(AA')}$ la norme de Frobenius. Trouver une matrice ayant un rang inférieur $S\,(S < K)$ qui approche mieux la matrice X au sens des moindres carrés permet de minimiser l'erreur de reconstitution. Cela revient à chercher deux matrices F_{IxS} et u_{KxS} minimisant le critère de minimisation défini ci-dessous :

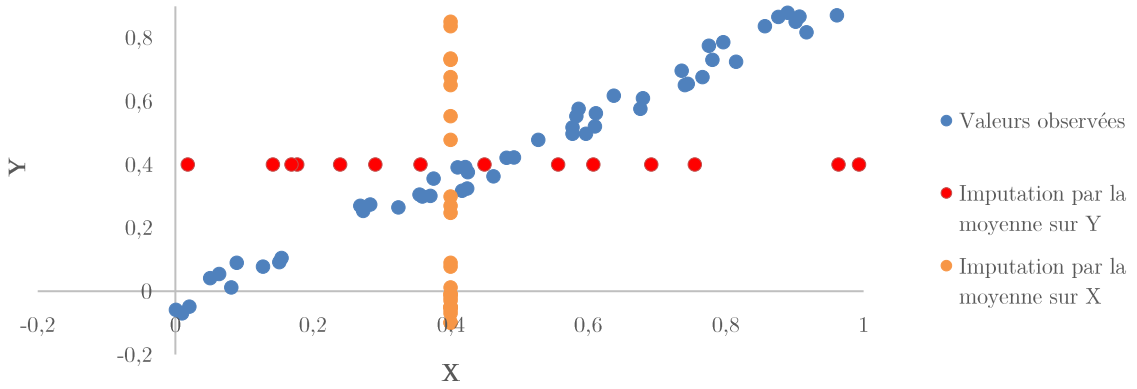
Équation 2.2: Critère de minimisation de l'erreur de reconstruction.

$$C = \|X - Fu'\|^2 = \sum_{i=1}^{I}\sum_{k=1}^{K}(x_{ik} - \sum_{s=1}^{S} F_{is}u_{ks})^2$$

En appliquant la contrainte d'orthogonalité de la base, la solution unique est fournie par les composantes principales notées $\hat{F}$ (normées à la valeur propre λ) et les axes principaux notés $\hat{u}$ de l'ACP correspondant respectivement aux vecteurs propres de la matrice de produit scalaire et de variance-covariance.

Affectation des valeurs manquantes des variable par l'ACP

Compte tenu de la taille de la base de données, qui rendait le temps d'exécution beaucoup trop important, le traitement des valeurs manquantes a été effectué sur deux bases de données sous-jacentes de la base globale de l'étude : l'une qui ne comporte que des variables quantitatives et l'autre qui ne comporte que des variables qualitatives. En ce qui concerne la base de données avec des variables quantitatives, la méthode d'affectation par l'analyse des composantes principales (ACP) itérative développée par le professeurs François Husson, Julie Josse et Jerome Pagès du Laboratoire de mathématiques Agrocampus a été utilisée. A la différence d'une affectation par la moyenne qui est souvent appliquée dans ce type d'étude, cette méthode ne déforme pas la distribution des variables explicatives (cf. graphique 2.5).



Graphique 2.5 - Exemple des deux variables dont les valeurs manquantes ont été affectées par la moyenne

Cette méthode reprend les travaux de Kiers en 1997 qui s'est intéressé à la minimisation générale des moindres carrées pondérés $\|W * (X - \mathcal{M})\|^2$ où W correspond à une matrice de poids et $\mathcal{M}$ un modèle général pour les données. Kiers s'est inspiré des travaux de Heiser en 1995, montrant qu'il possible de minimiser ce critère en utilisant de manière itérative un algorithme qui minimise le critère de moindres carrées ordinaire $\|(X - \mathcal{M})\|^2$. Dans le cas particulier de l'ACP où $M = Fu'$ et W une matrice de poids constituée uniquement de 0 et de 1 dont $w_{ik} = 0$ si x_{ik} est manquante et $w_{ik} = 1$ sinon, les étapes itératives sont les suivantes :

1. Initialisation de l'itération $\ell = 0$ où X^0 est obtenu en remplaçant les valeurs manquantes par une valeur initiale, par exemple, la valeur moyenne.

2. Lancement de l'itération ℓ :

   a. **Construire l'ACP sur la base de données complétée lors de l'itération $\ell - 1$:** Recherche de $(\hat{F}^\ell, \hat{u}^\ell)$ comme les paramètres de (F, u) minimisant le critère $\left\|(X^{\ell-1} - Fu')\right\|^2$, S dimensions sont retenues. S est défini par jugement d'expert.

   b. **Affecter les données manquantes par l'ACP en utilisant les coordonnées des individus et de variables :** X^ℓ est obtenu en remplaçant les valeurs manquantes de X par les valeurs reconstituées $\hat{X}^\ell = \hat{F}^\ell \, \hat{u}'^\ell$. Le nouveau tableau complété peut s'écrire comme $X^\ell = W * X + (1 - W) * \hat{X}^\ell$.

   c. **Mise à jour des moyennes et écart-types des variables après l'affectation.**

   d. Les étapes (a), (b) et (c) sont répétées jusqu'à la convergence.

La procédure décrite ci-dessus consiste à faire des ACP de façon itérative sur des bases de données complètes. L'idée est d'affecter les valeurs manquantes de la base initiale X^0 en prenant en compte la ressemblance globales entre individus (polices) et les liaisons entre les variables quantitatives. La figure 2.1 présente un exemple des étapes (a) et (b) d'une itération de l'algorithme.

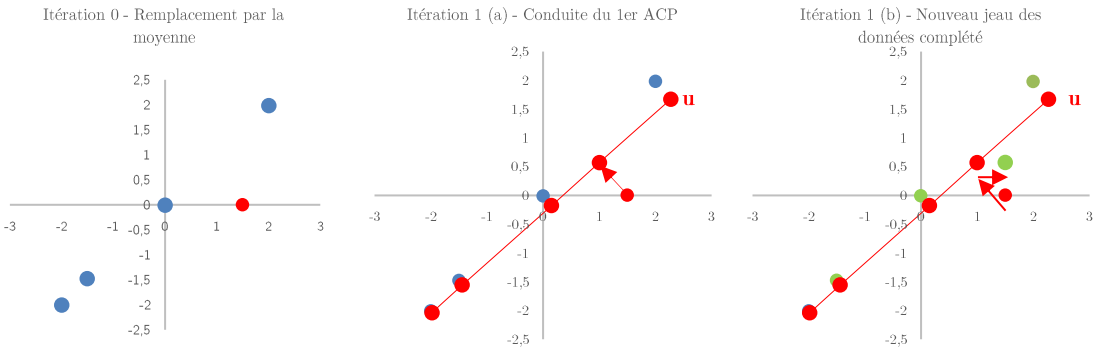


Figure 2.1 - Visualisation des étapes (a) et (b) d'une itération de l'algorithme ACP itérative

Cet algorithme conduit souvent au surajustement à cause de la croyance trop élevée dans le lien entre les variables dont le lien en réalité n'est pas si fort, notamment à cause des données manquantes. Pour résoudre ce problème, la fonction *imputePCA* des pacakges *missMDA* et *FactomineR* de R, met en œuvre une version régularisée de l'algorithme qui ne sera pas détaillée dans cette étude mais qui, de manière générale, croit moins aux liens entre les variables et affecte la valeur manquante par une valeur plus proche de la moyenne, ce qui atténue le risque lors de l'affectation. Enfin, l'algorithme complète la base de données avec des données qui n'influencent pas le résultat de l'ACP, c'est-à-dire qui n'influencent pas les coordonnées des individus et des variables.

Cet algorithme aurait pu être utilisé sur la base sous-jacente avec les variables qualitatives grâce à l'analyse de composantes multiples itérative ou même sur la globalité de la base grâce à l'analyse factorielle des données mixtes itérative. Cependant, étant donné la taille de la base de données globale et le grand nombre de catégories pour certaines variables catégorielles, notamment pour la variable modèle, les limitations de la mémoire de calcul ont empêché la réalisation de ces analyses.

2.2.3. Valeurs Aberrantes

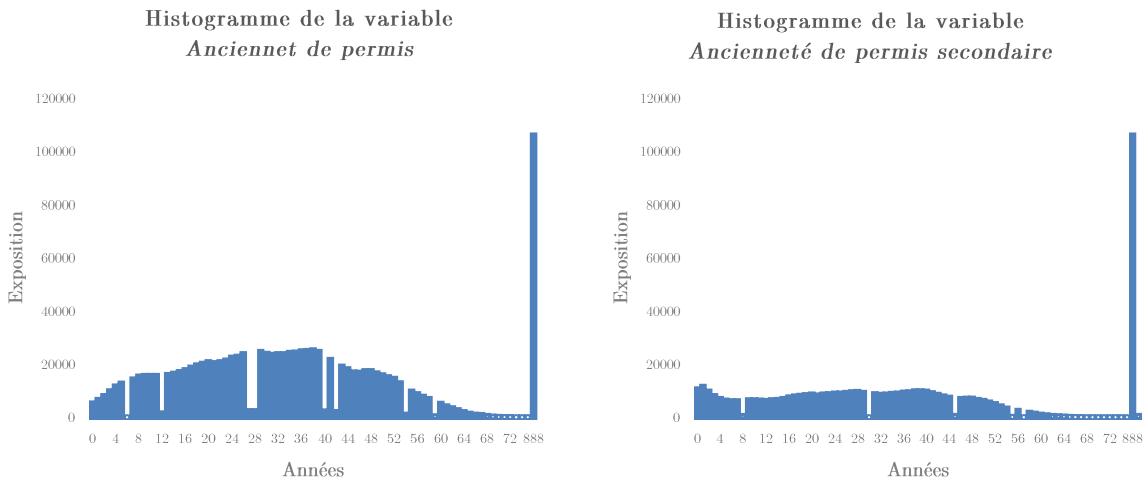
Une détection et un traitement approprié des valeurs aberrantes évitent d'introduire un biais considérable, notamment dans les modèles de régression. Il n'existe pas de procédure standard pour traiter les valeurs aberrantes, mais il est essentiel de les identifier et de comprendre leur impact dans les modèles prédictifs.

Dans le cadre d'une approche multivariée, pour déterminer si une police (représentée par une observation dans le jeu de données) est aberrante ou non, il est nécessaire de considérer collectivement les variables explicatives et la variable à expliquer (fréquence et coût moyen).

Valeurs aberrantes dans les variables explicatives

L'identification des valeurs aberrantes dans les variables explicatives a été réalisée par l'analyse d'histogrammes. Des anomalies ont été identifiées uniquement dans les variables *ancienneté de permis* et *ancienneté de permis du conducteur secondaire* exprimées en années. Certaines observations présentaient des anciennetés de 888 et 889 ans, ce qui n'est pas cohérent. Il a été décidé de supprimer la variable *ancienneté de permis du conducteur secondaire* car 62% des observations contenaient des valeurs aberrantes. La variable *ancienneté du permis* a été conservée dans l'analyse car elle est corrélée à l'âge du conducteur, variable qui n'est pas disponible dans la base de données de l'étude, et constitue probablement une variable discriminante pour expliquer la fréquence. De plus, seulement 9% des observations contenaient des valeurs à hauteur de 888 ans. Ces observations ont été conservées dans la base de données et traitées comme une modalité indépendante dans la modélisation. On verra plus loin dans le document que toutes les variables étudiées ont été considérées comme des

variables qualitatives, et en particulier les variables numériques ont été considérées comme des variables ordinales.



Graphique 2.6 - Histogrammes de l'exposition des variables Ancienneté de permis et Ancienneté de permis secondaire

Valeurs aberrantes dans les variables à expliquer

Le coût moyen étant une variable continue, l'analyse des valeurs aberrantes par l'analyse des histogrammes a été complétée par l'analyse du diagramme de type « Q-Q plot » ainsi que par l'analyse de la densité et de la fonction de distribution cumulée empiriques et théoriques. La distribution du coût moyen par sinistre a été ajustée à la loi Gamma en utilisant la méthode du maximum de vraisemblance.

Pour rappel, l'*estimateur de maximum de vraisemblance* est un estimateur statistique utilisé pour déduire les paramètres de la distribution de probabilité d'un jeu des données en recherchant les valeurs des paramètres qui maximisent la fonction de vraisemblance.

Équation 2.3: La densité de la loi Gamma $\Gamma(a,b)$ est définie par :

$$f_{(a,b)}(x) = \frac{b^a}{\Gamma(a,b)} - x^{a-1}\, e^{-bx}, \qquad x > 0$$

Avec des paramètres $a > 0$ et $b > 0$. Soit $x = (x_1, \ldots, x_n)$ une observation de $X = (X_1, \ldots, X_n)$ où les X_i sont i.i.d et suivent la loi Gamma.

Équation 2.4: La fonction de log-vraisemblance dans le cas de la loi Gamma est définie comme suit :

$$\ell(a,b) = \sum_{i=1}^{n} log\, f_{(a,b)}(x_i) = n \times a \times log(b) - n \times log(\Gamma(a)) + (a-1)\sum_{i=1}^{n} log(x_i) - b\sum_{i=1}^{n} log(x_i)$$

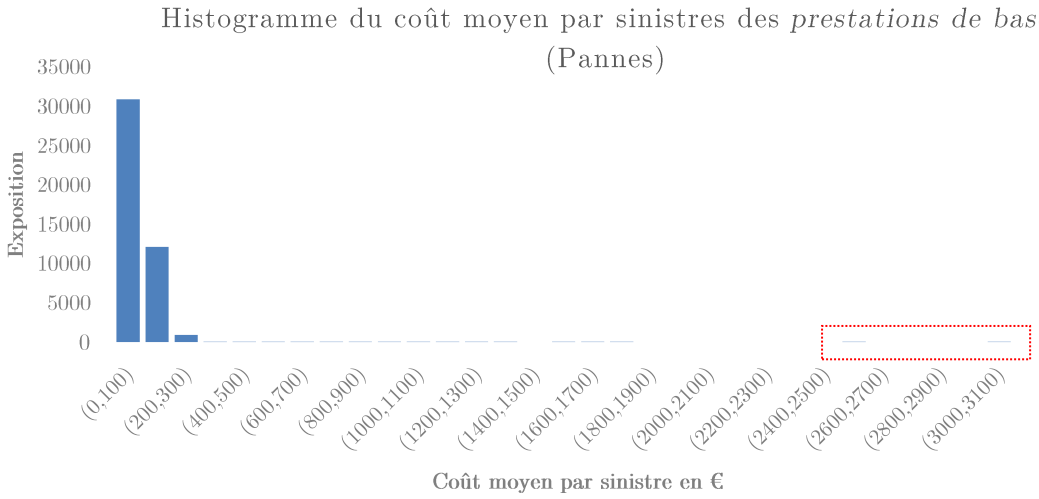
Équation 2.5 : L'estimateur du maximum de vraisemblance de a et b est donné par :

$$(\hat{a}, \hat{b}) = arg\, max_{a>0,b>0} \ell(a, b)$$

Comme mentionné dans l'équation 2.3, la densité de la distribution Gamma n'accepte que des valeurs positives. Pour respecter cette condition, l'identification des valeurs aberrantes a été effectuée sur l'échantillon de données ayant un coût des sinistres strictement supérieur à 0, c'est-à-dire les polices pour lesquelles au moins 1 sinistre est survenu. Cette base de données sous-jacente de la base de données agrégée compte 45 675 observations.

Traitement des valeurs aberrantes du coût moyen des sinistres liés aux pannes

Le graphique 2.7 représente l'histogramme du coût moyen par sinistre lié aux pannes. Les valeurs présentant un écart important par rapport au reste des observations ont été considérées comme aberrantes. Cela correspond aux valeurs observées au-delà de 2500€ dans ce cas.



Graphique 2.7- Histogramme du coût moyen des prestations de base par sinistre (Pannes)

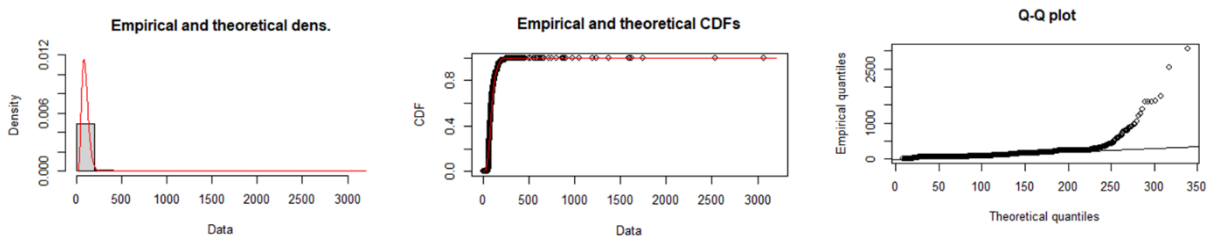
La figure 2.2 montre les résultats de l'ajustement à la loi Gamma du coût moyen des sinistres liés aux pannes. Ils sont représentés à travers de 3 graphiques différents détaillés ci-dessous :

- Densité empirique et théorique (1) : L'histogramme des données est à nouveau présenté, mais contrairement au graphique 2.6, la densité théorique est tracée en rouge. La ressemblance entre la densité empirique (histogramme) et la densité théorique permet de conclure que le coût moyen par sinistre suit une loi Gamma. Néanmoins, la queue de distribution est assez longue et montre une déviation importante entre la moyenne et les derniers quantiles. Ceci s'explique par les valeurs extrêmes identifiées dans le graphique 2.6.

- Fonction de distribution cumulée empirique et théorique (2) : La fonction de distribution cumulative empirique est représentée par les points noirs et la

fonction théorique par la courbe rouge. Ce graphique confirme également l'hypothèse que le coût moyen suit une loi Gamma, étant donné la similitude des deux courbes. En ce qui concerne les valeurs extrêmes, ce graphique montre que les valeurs supérieures à 500€ euros s'éloignent de la courbe de distribution cumulée et donc peuvent être considérées comme des valeurs aberrantes.

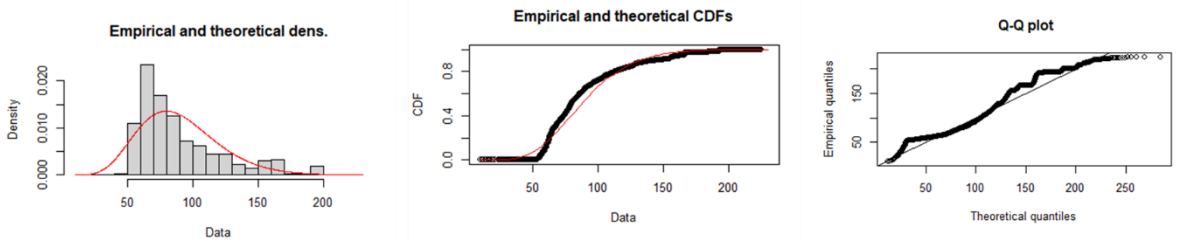
- Diagramme « Q-Q plot » (3) : Ce diagramme permet de comparer la distribution théorique avec la distribution empirique en traçant leurs quantiles l'un par rapport à l'autre. Si les deux distributions comparées sont similaires, les points du diagramme se situeront approximativement sur la ligne d'identité $y = x$. Le « Q-Q plot » peut également aider à visualiser les valeurs aberrantes. Les observations qui s'écartent significativement de la ligne droite de référence ne suivent pas une distribution Gamma et peuvent donc être considérées comme des valeurs aberrantes. Contrairement à ce qui est observé sur le graphique (2), le « Q-Q plot » montre que les valeurs supérieures à environ 225€ peuvent également être considérées comme des valeurs extrêmes en raison de la forte déviation de ces points par rapport à la ligne droite de référence.



(1) (2) (3)

Figure 2.2 - Graphiques d'ajustement de distribution des coûts moyens de sinistres à la loi Gamma (Pannes)

Ainsi, les observations ayant un coût moyen supérieur à 225€ ont été retirées de la base. Elles représentent 1,85% de la base sous-jacente ayant seulement les coûts des sinistres strictement supérieurs à 0 et 0,07% de la base globale agrégée. Lorsque les graphiques présentés dans la figure 2.1 ont été tracés sur la base de données sans les valeurs extrêmes détaillées auparavant, des coûts moyens inférieurs à 10€ ont été identifiés. Ces valeurs s'écartent considérablement des autres observations et du coût moyen observé des *prestations de base*. Par conséquent, elles ont également été considérées comme aberrantes et retirées de la base de données. Seules 12 observations étaient concernées. La Figure 2.3 présente l'ajustement des données sans valeurs aberrantes liées aux pannes

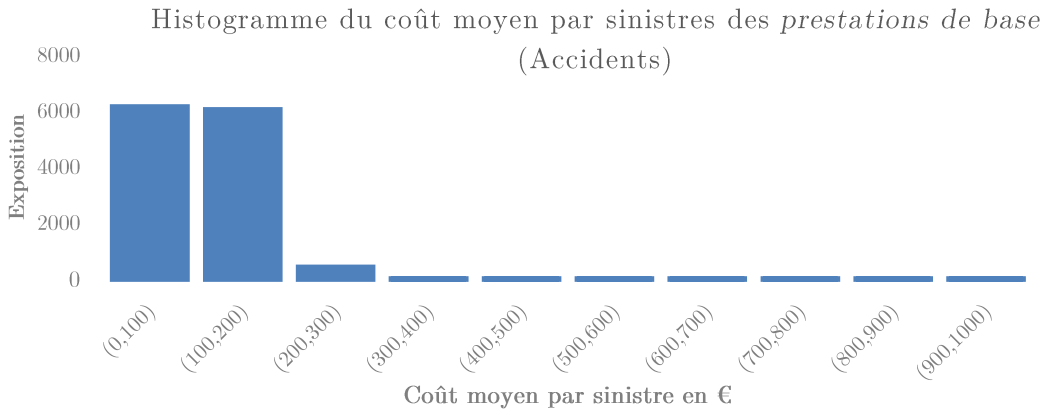


(4) (5) (6)

Figure 2.3 – Graphiques d'ajustements de distribution des coûts moyens de sinistres (sans valeurs aberrantes) à la Loi Gamma (Pannes)

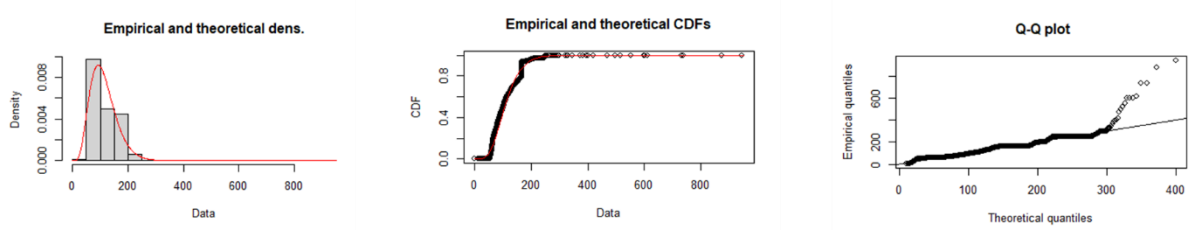
Traitement des valeurs aberrantes du coût moyen des sinistres liés aux accidents

L'histogramme du coût moyen des sinistres liés aux accidents est présenté dans le graphique 2.8. A première vue, il n'y a pas des valeurs aberrantes dans le cas des accidents.



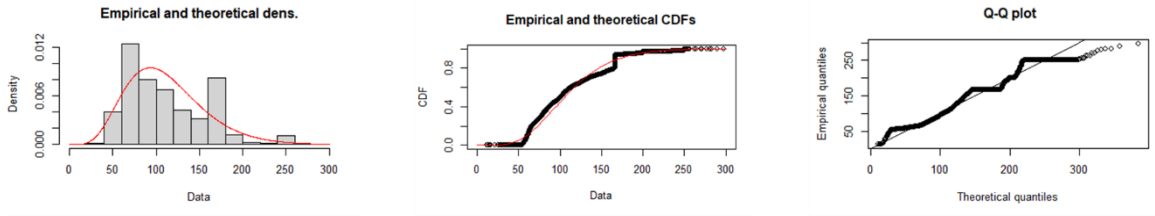
Graphique 2.1 - Histogramme du coût moyen par sinistres des services de base (Accidents)

L'ajustement à la loi Gamma par la méthode de maximum de vraisemblance réalisée pour les pannes a été également fait pour les accidents. Contrairement à ce qui est montré sur l'histogramme, la distribution cumulée et le diagramme « Q-Q plot » montrent que les coûts moyens supérieurs à 300€ et inférieur à 10€ peuvent être considérées comme aberrantes (cf. Figure 2.4). De ce fait, 32 observations ont été retirées de la base de données globale. La figure 2.5 montre le résultat de l'ajustement suite à l'écrêtement.



(1) (2) (3)

Figure 2.4 - Graphiques d'ajustement de distribution des coûts moyens de sinistres à la loi Gamma



(4) (5) (6)

(Accidents)Figure 2.5 - Graphiques ajustements de distribution des coûts moyens de sinistres (sans valeurs aberrantes) à la Loi Gamma (Accidents)

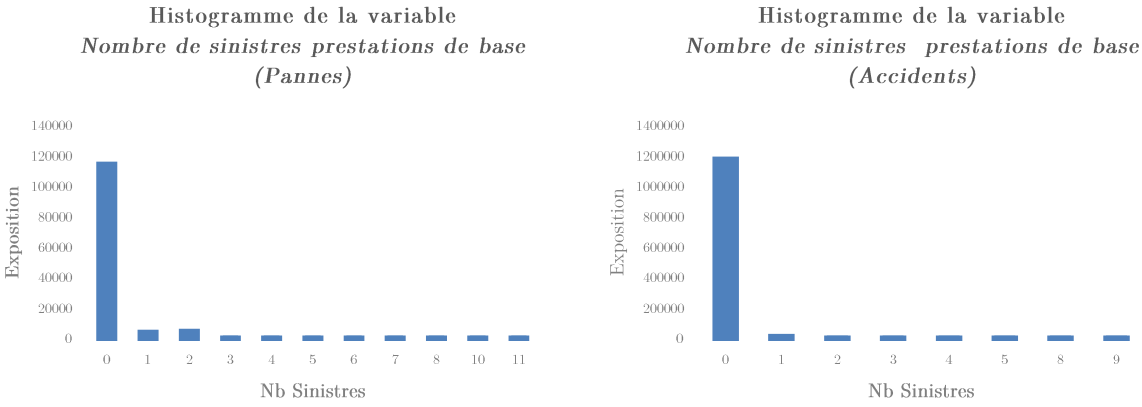
Les coûts moyens inférieurs à 10€ peuvent être considérés comme des erreurs d'enregistrement car il est très peu probable que les services de base aient un coût aussi bas. En revanche, les coûts supérieurs à 225€ dans le cas de pannes et à 300€ dans le cas d'accidents sont considérés comme des sinistres graves. En principe, ils devraient être traités séparément en utilisant la théorie des valeurs extrêmes par exemple, mais comme seulement 0,07% des observations totales ont été retirés de la base de données, cette étude se concentrera uniquement sur la modélisation des sinistres attritionnels.

L'écrêtement des données liées au coût moyen permet de réduire la déviation entre la distribution théorique et la distribution empirique des pannes ainsi que des accidents. Cependant, des déviations sont encore observées et peuvent être expliquées par le fait que le coût des *services de base* est composé du coût des deux services différents : le *dépannage sur place* et le *remorquage*, ce qui ajoute du bruit à la distribution du coût total. Sur la période étudiée, un *dépannage sur place* a coûté en moyenne 85€ et un *remorquage* a coûté en moyenne 115€, soit une différence de 35% entre le coût des deux services. De plus, la contribution des deux services au coût total dépend du fait générateur ; dans le cas des pannes, le poids du dépannage sur place sera plus important, tandis que dans le cas des accidents, le poids du remorquage sera plus important. Par conséquent, chacun des services peuvent suivre une loi Gamma mais avec des paramètres différents.

Il serait plus approprié de traiter séparément le *dépannage sur place* et le *remorquage*, mais l'assureur partenaire les enregistre comme un seul type de service, justifiant que le service de *remorquage* est largement majoritaire. Ceci sera considéré comme une limitation dans la modélisation du coût moyen.

En ce qui concerne le nombre de sinistres, l'identification des valeurs aberrantes s'est basée sur l'analyse des histogrammes, de la densité et la fonction de distribution

cumulée empiriques et théoriques. La distribution empirique du nombre de sinistres a été ajustée à la loi Poisson en utilisant également la méthode du maximum de vraisemblance expliquée auparavant pour la distribution du coût moyen. Le graphique 2.8 et la figure 2.5 montrent qu'il n'y a pas de valeurs atypiques dans les données, donc aucun ajustement n'a été fait sur le nombre de sinistres.



Graphique 2.8 - Histogrammes du nombre des sinistres lié aux services de base

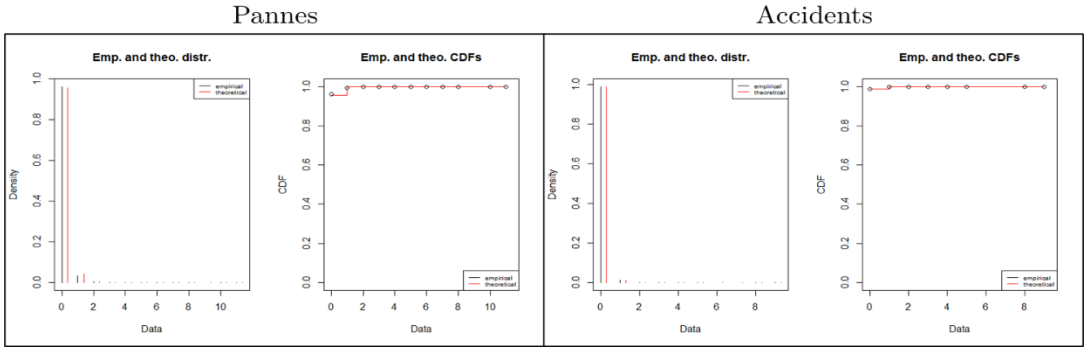


Figure 2.6 - Graphiques d'ajustement de distribution du nombre des sinistres

Enfin, le tableau 2.5 recapitule le détail des lignes retirées à la suite du traitement des valeurs aberrantes.

	Nb. Observations
Base de données avec des valeurs aberrantes	1 192 195
Valeurs aberrantes liées aux pannes	850
Dont coûts moyen inférieurs à 10€	*12*
Dont coûts moyen supérieurs à 225€	*838*
Valeurs aberrantes liées aux accidents	37
Dont coûts moyens inférieurs à 10€	*5*
Dont coûts moyens supérieurs à 300€	*32*
Base de données sans valeurs aberrantes	1 191 308
% des observations retirées	**0,07%**

Tableau 2.5 - Détail des observations retirées suite au traitement des valeurs aberrantes

2.2.4. Caractère temporel du coût moyen

Comme mentionné dans le chapitre 1, Europ Assistance dispose d'un réseau large de partenaires en charge de fournir les différents services d'assistance automobile. Par conséquent, le coût d'un sinistre dépend fortement de son année de survenance, du fait de la renégociation des prix entre Europ Assistance et ses partenaires et l'inflation annuelle du coût du transport notamment dans le cas des *services de base*. Pour considérer ces deux effets dans la modélisation, l'année de survenance sera intégrée dans la modélisation du coût moyen en tant que variable qualitative.

Une fois les ajustements effectués, la base de données de l'étude est prête à être exploitée. Il est maintenant temps de vérifier la stabilité du portefeuille dans le temps et d'effectuer des analyses exploratoires pour mieux comprendre les caractéristiques du risque à assurer.

2.3. Stabilité du portefeuille dans le temps

La vérification de la stabilité du portefeuille dans le temps assure la cohérence de la modélisation tout au long de la période d'étude. A cette fin, la stabilité de l'exposition des variables explicatives a été vérifiée en analysant la répartition de l'exposition des modalités au cours de la période étudiée. Les résultats de cette analyse pour l'ensemble des variables ne sont pas présentés dans ce document, mais la figure 2.6 montre les histogrammes empilés pour certaines variables à titre d'exemple.

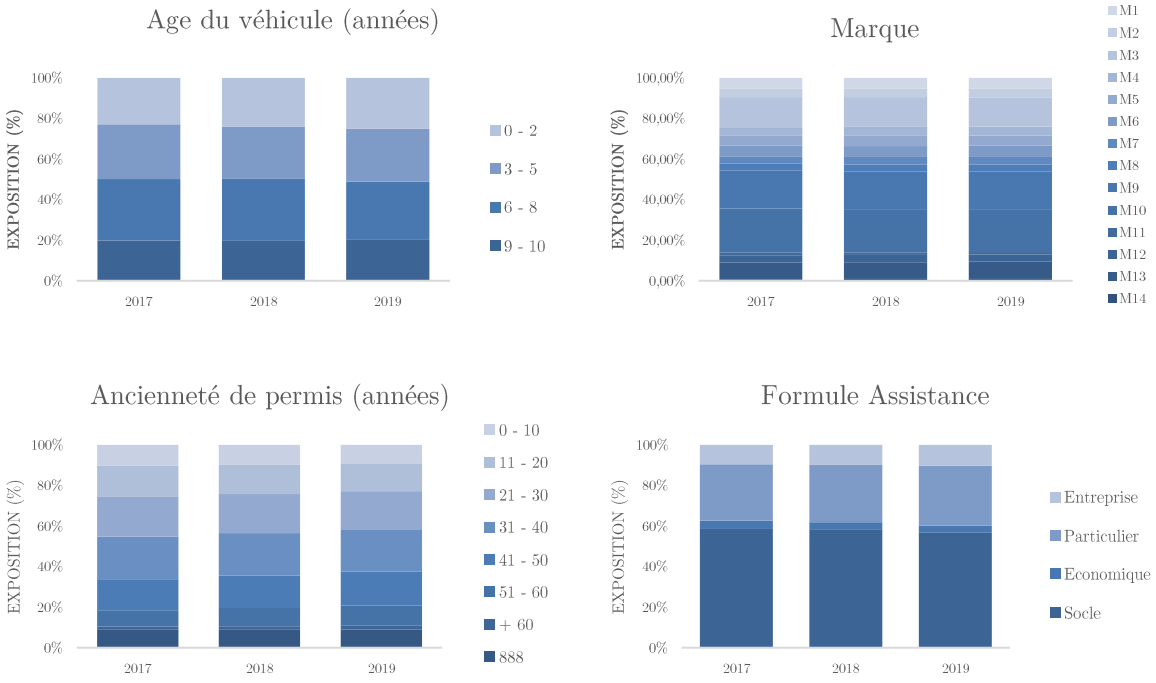


Figure 2.7 - Stabilité dans le temps des variables explicatives

Les graphiques ci-dessus montrent que les modalités des variables sont stables tout au long de la période étudiée. Après avoir effectué cet exercice pour l'ensemble des variables explicatives, il apparait que le portefeuille est globalement stable dans le temps.

2.4. Analyses exploratoires

Après la description et l'ajustement de la base de données à étudier, des analyses exploratoires de la fréquence et du coût moyen ont été réalisées. Ceci permettra de mieux caractériser le portefeuille et d'identifier les variables explicatives ayant un impact sur la fréquence et le coût moyen et donc étant potentiellement discriminantes dans la modélisation de la prime de risque des *prestations d'assistance de base*. Compte tenu du grand nombre de variables explicatives, ce type d'analyse n'a été réalisé que pour les variables habituellement utilisées dans le domaine de l'assurance et l'assistance automobile :

- Variables liées à l'assuré : ancienneté de permis, bonus-malus, profession, et formule assistance.
- Variables liées au véhicule : âge du véhicule, marque, carrosserie et type de moteur.

2.4.1. L'analyse univariée des variables liées à l'assuré

Analyses univariées sur la fréquence

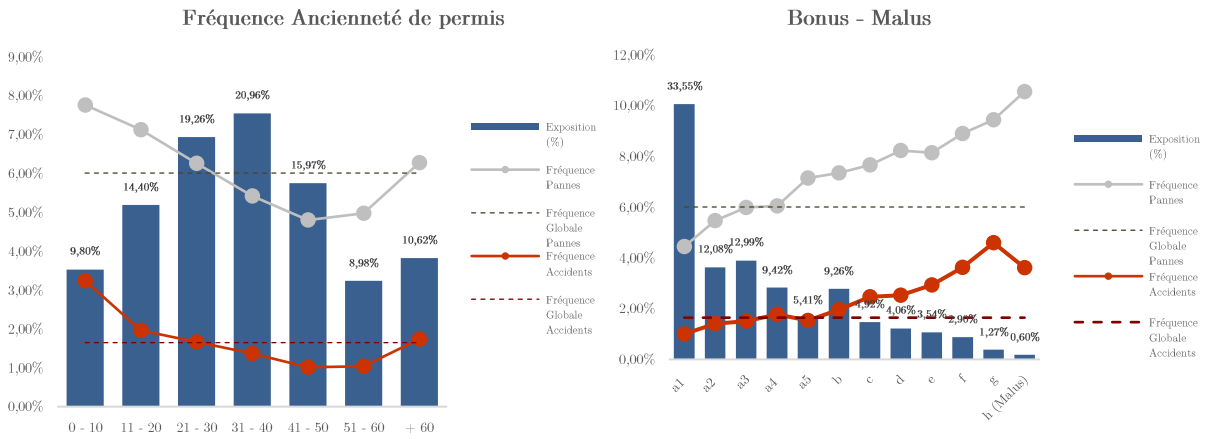


Figure 2.8 - Fréquence par ancienneté de permis (années) et niveau de bonus-malus

D'après la figure 2.8, les variables d'ancienneté du permis et de bonus-malus ont un impact non négligeable sur la fréquence des pannes et des accidents. Dans le cas des accidents, la forme convexe des courbes de fréquence de la variable ancienneté du permis de conduire est cohérente avec le phénomène lié à l'âge du conducteur usuellement observé en assurance automobile où les conducteurs les plus jeunes et les plus âgés ont une fréquence plus élevée. On observe une convexité similaire dans le cas des pannes, ceci peut s'expliquer par le fait que les jeunes conducteurs possède

généralement des voitures plus anciennes ayant un risque de pannes plus élevé (cf. Annexe A.2). Cette interaction sera prise en compte dans la modélisation de la fréquence des pannes. En ce qui concerne la variable bonus-malus, comme prévu, les courbes de fréquence augmentent lorsque le niveau de bonus diminue.

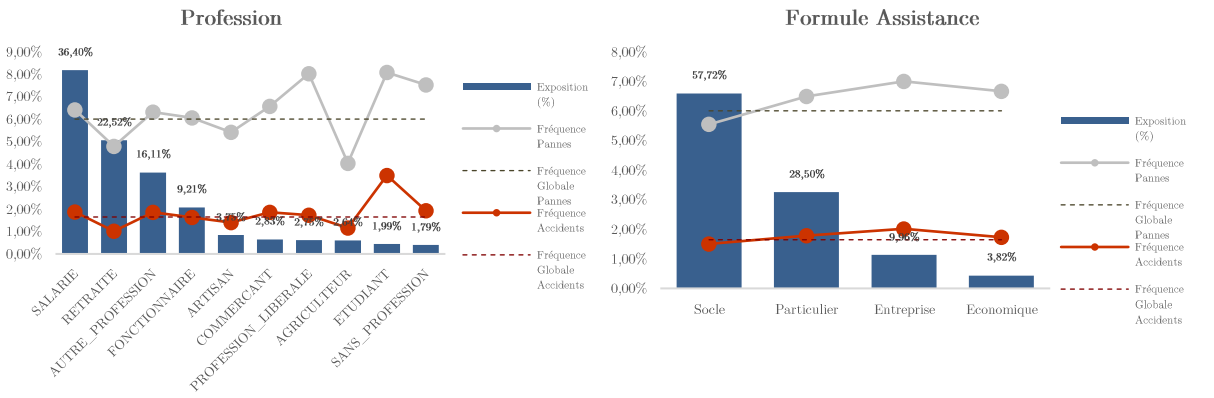


Figure 2.9 - Fréquence par profession et formule d'assistance

Il apparaît que les variables *profession* et *formule d'assistance* ont un impact plus faible sur la fréquence que les variables *ancienneté de permis* et *bonus-malus*, néanmoins, leur impact n'est pas négligeable. Dans le cas des professions, les profils des commerçants, libéraux, étudiants et sans profession ont une fréquence plus élevée que la moyenne. Cela peut s'expliquer par la forte utilisation de leur véhicule par les profils commerçants et libéraux et par le manque d'expérience des profils souvent jeunes comme les étudiants et les personnes sans profession.

Pour la variable *formule d'assistance*, il apparaît que les formules Particuliers, Entreprises et Economiques ont des fréquences légèrement supérieures à la moyenne. Étant donné que, contrairement à la Formule Socle, les profils souscrivant à ces trois formules ont inclus le service secondaire d'un véhicule de remplacement, il est possible que ces profils soient caractérisés par une utilisation élevée de ses véhicules.

Analyses univariées sur le coût moyen

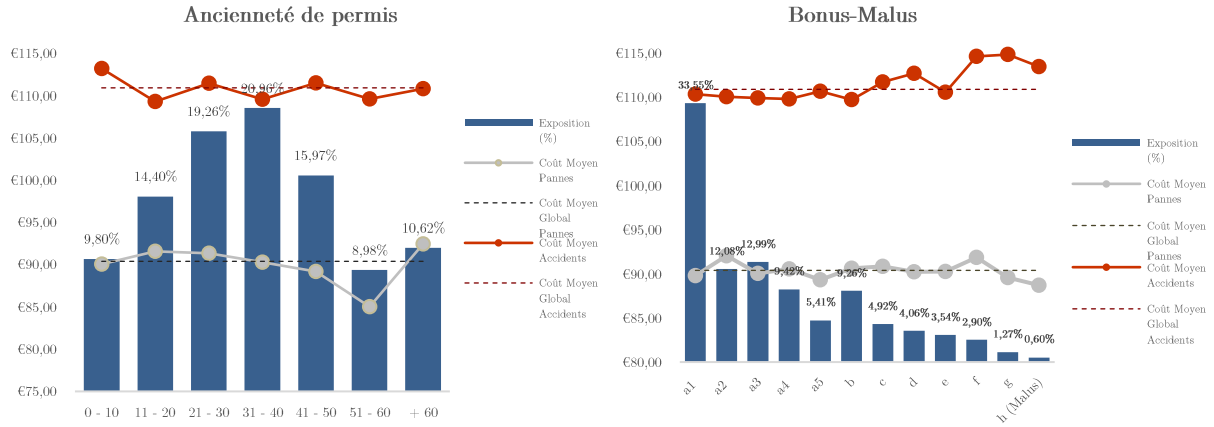


Figure 2.10 - Coût moyen par ancienneté de permis (années) et niveau de bonus-malus

Contrairement à la fréquence, l'ancienneté de permis et le niveau de bonus-malus ne semblent pas avoir un impact significatif sur le coût moyen. Ceci est cohérent avec le fait que le coût moyen d'un dépannage ou d'un remorquage sur place sera davantage lié aux caractéristiques du véhicule qu'au profil du conducteur.

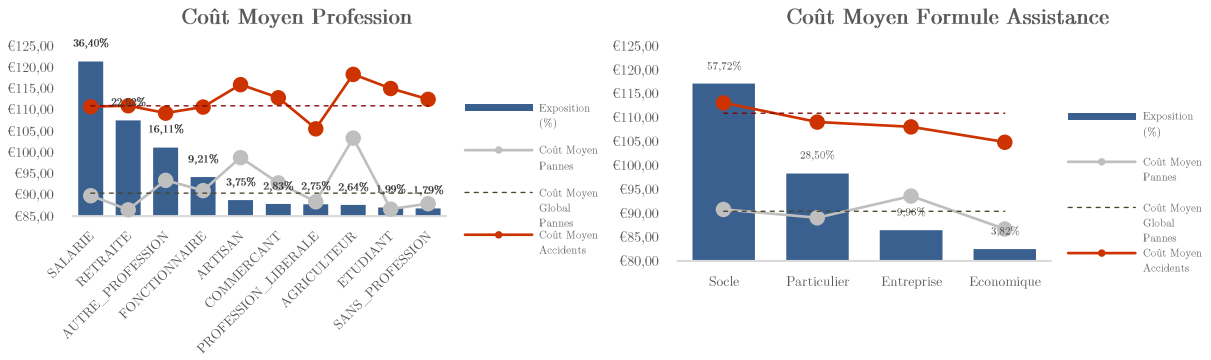


Figure 2.11 – Coût moyen par profession et formule d'assistance

Néanmoins, les caractéristiques d'un véhicule peuvent être liées au profil de son conducteur. Par exemple, les artisans et les agriculteurs auront tendance à avoir des véhicules spécifiques plus grands pour exercer leur métier. Par conséquent, en cas de panne ou d'accident, ils seront difficiles à réparer sur place et le remorquage sera plus coûteux en raison de la taille du véhicule. Quant à la formule d'assistance, aucun impact n'est observé dans le cas des pannes. Dans le cas des accidents, les formules Particulier, Entreprise et Economique ont un coût moyen inférieur à la moyenne. Cette relation est difficile à expliquer et doit être relativisée en raison du faible nombre de sinistres accident pour ces formules.

Enfin, il ressort de cette analyse que les variables liées à l'assuré ont un impact plus important sur la fréquence que sur le coût moyen. Ceci semble cohérent car le coût moyen dépend davantage des caractéristiques du véhicule que de celles du conducteur. De même, en suivant l'analyse par fait générateur, il apparaît qu'il existe une différence significative pour la fréquence et le coût moyen entre les pannes et les accidents. Les accidents sont des sinistres de faible fréquence mais de coût moyen élevé. Néanmoins, la relation entre les variables explicatives et les deux faits générateurs sont relativement similaires.

2.4.2. L'analyse univariée des variables liées au véhicule

Analyses univariées sur la fréquence

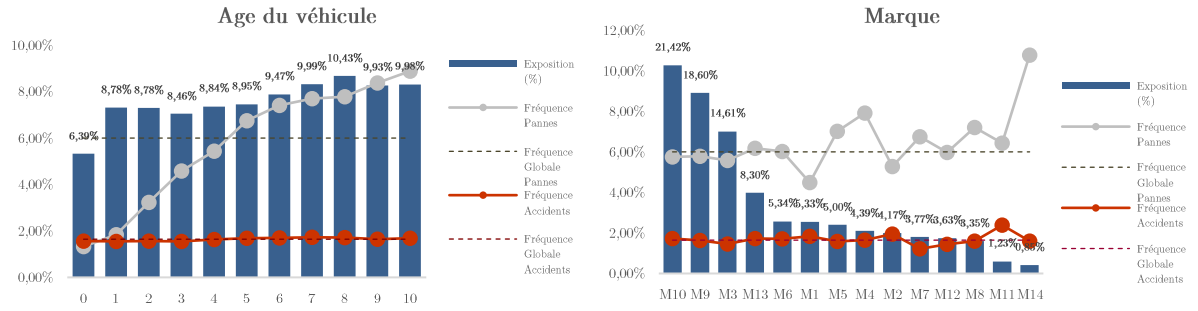


Figure 2.12 - Fréquence par âge du véhicule et par marque

L'âge du véhicule et la marque ont une forte incidence sur la fréquence des pannes. Comme attendu, la fréquence des pannes augmente avec l'âge du véhicule. Ce phénomène s'explique par la diminution des performances du véhicule due à son vieillissement. Il existe également une différence de fréquence des pannes en fonction de la marque, ce qui semble cohérent étant donné les différences de performances de chacune des marques existantes sur le marché. En ce qui concerne les accidents, il n'y a pas d'impact significatif de ces deux variables sur la fréquence.

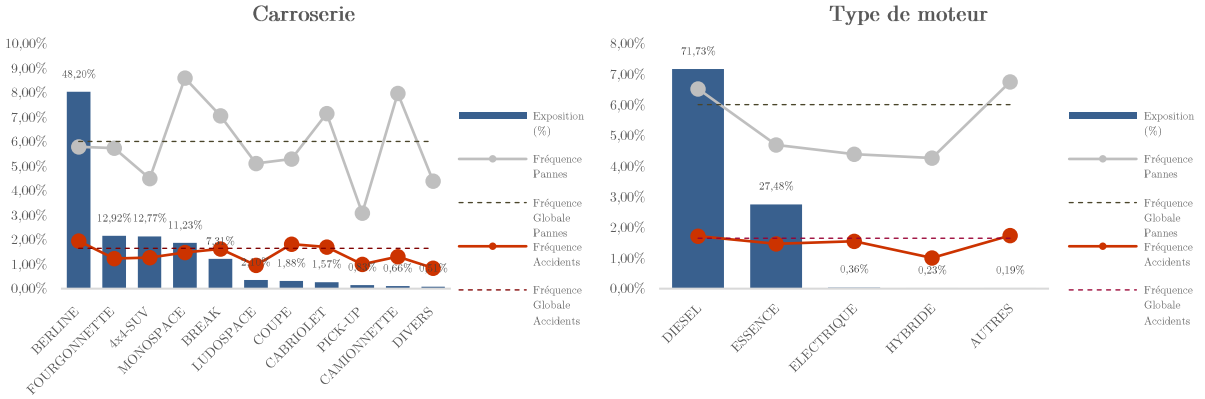


Figure 2.13 - Fréquence par carrosserie et type de moteur

La carrosserie est une variable qui a également un impact sur la fréquence des pannes. Il apparaît que les profils disposant d'un véhicule avec une carrosserie de type monospace, break ou cabriolet auraient plus de risque de tomber en panne. A l'inverse, pour les accidents, il semble que la carrosserie ait peu d'impact sur la fréquence.

Pour le type de moteur, il semble que les véhicules à essence, électriques et hybrides soient moins susceptibles de tomber en panne. Toutefois, cette interprétation doit être prise avec précaution, car environ 72 % de l'exposition est composée de véhicules à moteur diesel. Comme observé pour les autres variables, le type de moteur n'a pas non plus d'impact sur la fréquence des accidents.

Analyses univariées sur le coût moyen

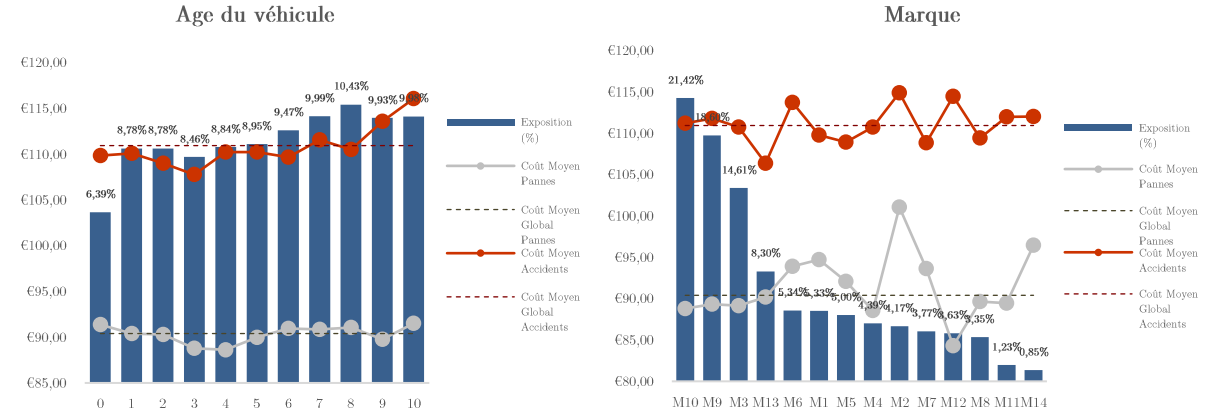


Figure 2.14 - Coût moyen par âge du véhicule et par marque

La figure 2.13 montre que l'âge du véhicule a peu d'impact sur le coût moyen des sinistres, tant pour les pannes que pour les accidents. En contrepartie, la marque a une influence sur le coût moyen, surtout dans le cas des pannes. Il faut noter que la marque est liée au type et à la gamme du véhicule. Par exemple, un véhicule haut de gamme sera plus souvent remorqué pour être réparé dans des garages spécialisés.

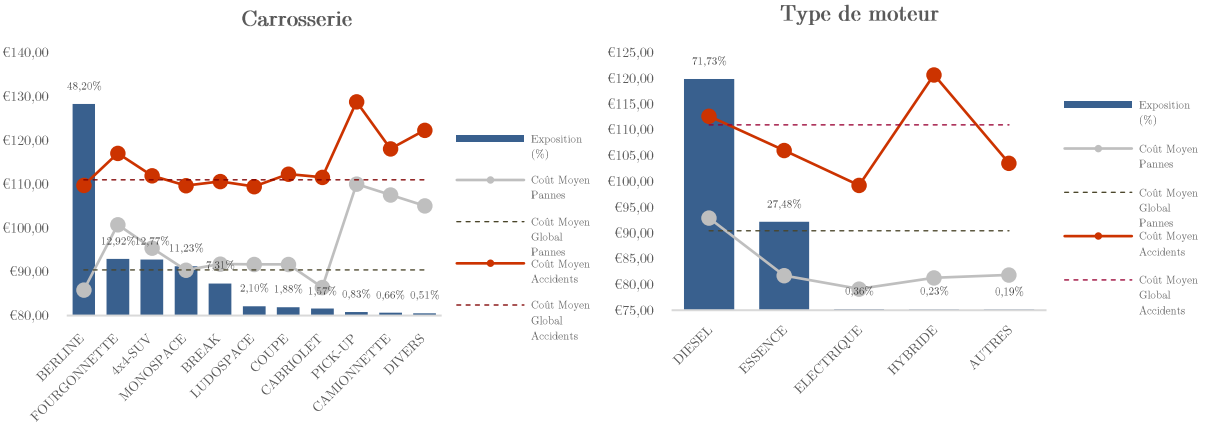


Figure 2.15 - Coût moyen par carrosserie et type de moteur

Le type de carrosserie et de moteur a un impact significatif sur le coût moyen des sinistres. Il est cohérent d'observer une différence de coût entre les carrosseries de taille moyenne (berline ou coupe) et les grosses carrosseries (fourgonnette, pick-up ou camionnette). L'expérience montre que ces dernières sont difficiles à réparer sur place et nécessitent un véhicule de remorquage spécial. Concernant le type de moteur, il apparaît que pour les pannes, les véhicules à essence et rechargeables (électriques et hybrides) ont un coût des sinistres inférieur à la moyenne. Cela pourrait être dû au poids plus important du dépannage sur place par rapport au remorquage, notamment dans le cas des véhicules électriques et hybrides, dont les sinistres sont souvent liés à des problèmes de batterie qui peuvent être facilement résolus sur place. Dans le cas des accidents, le comportement observé pour les pannes est reproduit sauf pour les véhicules hybrides, pour lesquels on observe un coût des sinistres significativement

plus élevé que la moyenne. Aucune conclusion ne peut être tirée, étant donné le faible nombre d'observations disponibles.

Enfin, suite à l'analyse de certaines variables liées au véhicule, on peut constater qu'elles ont un impact sur la fréquence ainsi que sur le coût moyen. Néanmoins, cet impact est plus important dans le cas des pannes, ce qui permet de conclure que les accidents sont plus souvent liés au profil du conducteur qu'au type de véhicule.

Les analyses exploratoires des variables géographiques et les effets d'interaction entre les variables sont également intéressants à aborder. Elles seront évoquées plus loin dans le document.

3. Pré-sélection des variables

Les variables disponibles dans le jeu de données représentent des informations importantes pour expliquer le comportement d'autres variables, dans ce cas, le coût moyen et la fréquence. Bien qu'un grand nombre de variables puisse donner de la robustesse à un modèle de tarification, leur exploitation individuelle ainsi que l'exploitation des liens entre variables devient assez coûteuse en termes de temps de calcul. De plus, la gestion d'un grand nombre de variables, en particulier avec des variables qualitatives à plusieurs modalités rend les modèles complexes et difficiles à interpréter.

Afin de pouvoir traiter une base de données exploitable en termes de temps de calcul et de complexité, trois étapes de sélection des variables ont été appliquées préalablement à la modélisation :

1. Analyse des liens parmi les variables quantitatives : En utilisant l'analyse en composantes principales (ACP), les liens entre les variables quantitatives seront identifiés. Les variables fortement corrélées seront éliminées.

2. Analyse des liens parmi les variables qualitatives[11] : En utilisant l'analyse des correspondances multiples (ACM) et le V de Cramer, les liens entre les variables qualitatives seront identifiés. Les variables fortement corrélées seront éliminées.

3. Matrice de corrélation : Afin d'identifier les relations entre les variables quantitatives et qualitatives, une matrice de corrélation PHI-K a été construite. L'analyse des relations des variables prises deux à deux permettra encore d'éliminer les attributs redondants.

3.1. Analyse des liens parmi les variables quantitatives

Comme réalisé pour le traitement des valeurs manquantes (cf. section 2.2.2), la base des données globale après retraitements a été découpée en deux bases sous-jacentes : une contenant les variables quantitatives et une autre contenant les variables qualitatives.

L'approche descriptive de l'analyse des composantes principales (ACP) a été appliquée pour analyser les relations entre les variables quantitatives. Pour rappel, l'ACP projette une observation de la base de données liée à certaines variables explicatives sur un sous-espace de dimension inférieure tout en maximisant la variance entre les données pour capturer le maximum d'information. Bien que l'intégralité des variables étudiées soient quantitatives, elles restent très hétérogènes en termes d'échelle et unité de mesure. Par exemple, l'âge du véhicule a une échelle et une unité

[11] Etant donné que la base de données dispose de 528 modèles de véhicules différents, la variable « Modèle » a été retirée pour faciliter l'analyse des modalités effectuée par l'ACM. Elle a été réintégrée dans le calcul du V de Cramer.

différentes du poids du véhicule. Cela a un fort impact sur le calcul des variances fait par l'ACP. La normalisation est une solution à ce problème car elle permet de rendre les variables comparables. Lorsque les données sont normées et centrées, elles ont toutes le même écart-type et une moyenne égale à 0.

Équation 3.1: Normalisation des données

$$x_{norm} = \frac{x_i - \mu(x)}{\sigma(x)}$$

Où,

$\mu(x) = Moyenne\ de\ la\ variable\ x$

$\sigma(x) = Ecart - type\ de\ la\ variable\ x$

3.1.1. Choix d'axes principaux à retenir pour l'ACP

Une fois les données normées et centrées, il faut définir le nombre d'axes à prendre en compte pour la projection sur le sous-espace de plus faible dimension généré par la base orthonormale. L'objectif est de choisir les axes qui capturent le maximum de variance, c'est-à-dire le maximum d'information. La variance des points projetés dans la direction $\vec{u}_1$ est définie comme $\vec{u}_1{}^T S\, \vec{u}_1$, dont S correspond à la matrice de corrélation empirique. L'expression qui montre la maximisation de la variance en appliquant la contrainte d'orthogonalité ($\|\overrightarrow{u_1}\|^2 = \vec{u}_1{}^T \vec{u}_1 = 1$) est présentée dans l'équation 3.2 (BUCCI, 2021).

Équation 3.2 : Maximisation de la variance des point projetées dans la direction $\vec{u}_1$

$$S\, \vec{u}_1 = \lambda_1 \vec{u}_1$$

$$\vec{u}_1{}^T S\, \vec{u}_1 = \lambda_1$$

λ_1 correspond à la valeur propre la plus forte associée au vecteur $\vec{u}_1$. La première composante principale représente le vecteur propre de la matrice de corrélation empirique.

Les *valeurs propres* associées aux composantes principales définies par l'ACP seront utilisées pour mesurer la quantité de variance expliquée par chaque axe. Ainsi, le *pourcentage de variance expliquée* sera l'indicateur à considérer pour déterminer le nombre d'axes principaux à retenir dans l'analyse.

Équation 3.3: Pourcentage de variance expliquée par un axe principale

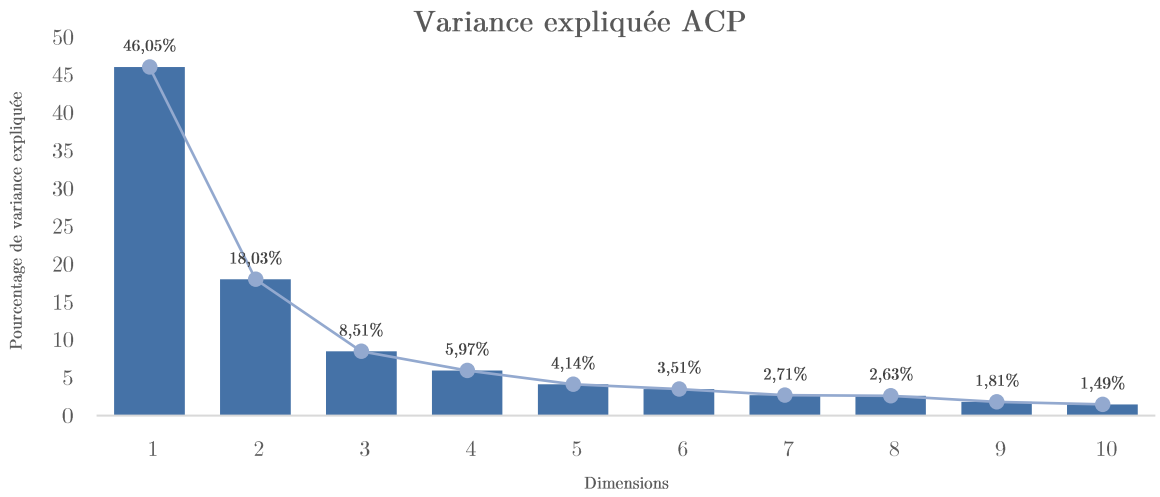
$$var\_exp_k = \frac{\lambda_k}{\sum_{k=1} \lambda_k} = \frac{\lambda_k}{I}$$

Où,

$\lambda_k = Valeur\ propre\ liée\ à\ l'axe\ k$

$I = Variance\ totale\ ou\ inertie\ du\ nuage\ des\ points\ (somme\ de\ tous\ les\ valeurs\ propres)$

La figure 3.1 montre le pourcentage de variance expliquée par les 10 premières composantes principales relatives à la base contenant uniquement les variables quantitatives. Comme environ 73% de la variance est expliquée par les 3 premières composantes, il a été décidé de garder 3 axes pour la projection. Ces axes seront utilisés pour construire le plan factoriel où seront projetés les variables et les observations. Par la suite, l'objectif est d'identifier les variables ayant un comportement similaire (fortement corrélées), ce qui réduira le nombre de variables quantitatives à analyser.



Graphique 3.1 - Histogramme du pourcentage de variance expliquée par les premiers 10 composantes principales (ACP)

3.1.2. Cercle de corrélations

Dans l'ACP, les observations sont représentées par leurs projections (à l'aide de la matrice de variance-covariance empirique) et les variables sont représentées par leurs corrélations (Abdi & Williams, 2010). Par conséquent, la corrélation d'une variable avec l'une des composantes principales est la coordonnée de la variable sur l'un des axes retenus.

Les cercles de corrélation présentés dans la figure 3.1 montrent les relations entre les variables et les trois premiers axes retenus. Le premier et le deuxième axe sont représentés dans le cercle de gauche et le deuxième et le troisième axe sont représentés dans le cercle de droite.

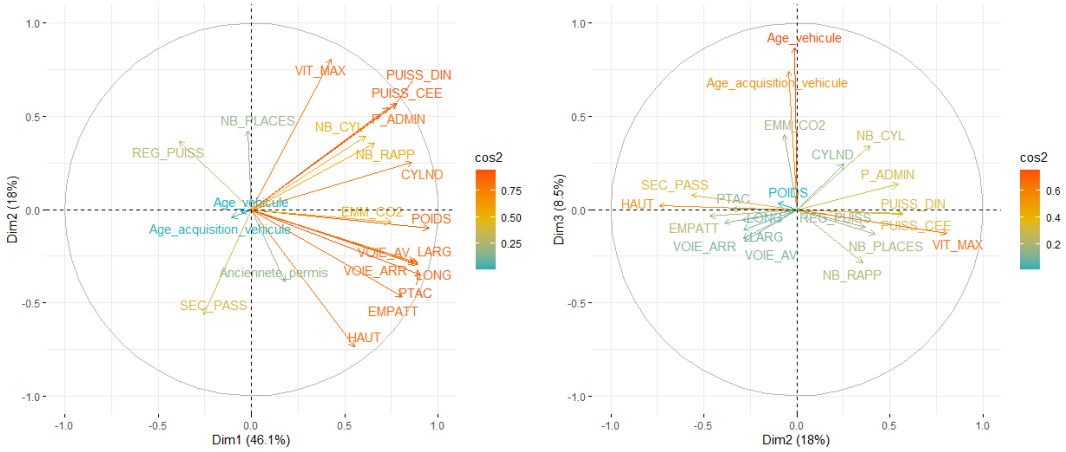


Figure 3.1 - Cercle de corrélation entre les variables et les composantes principales (ACP)

Les cercles de corrélation présentés ci-dessus s'interprètent comme suit (Kassambara, 2017) :

- Les variables regroupées dans le même quadrant du cercle sont positivement corrélées.
- Les variables positionnées avec des directions et quadrants opposés sont négativement corrélées.
- La qualité de la représentation des variables par l'ACP est mesurée par le *cosinus carré* de l'angle que forme la variable avec sa projection. Un *cosinus carré* élevé indique une bonne représentation de la variable sur les axes considérés. De manière visuelle, la flèche représentant la variable dans le cercle de corrélation est positionnée près de la circonférence. Au contraire, les flèches étant proches du centre du cercle ont un *cosinus carré* faible et donc une moins bonne représentation. Dans la figure 3.1, les variables les mieux représentées sont soulignées en rouge/orange tandis que les moins bien représentées sont soulignées en vert/bleu.

L'analyse des deux premiers axes montre que la puissance DIN (PUISS_DIN), la puissance CEE (PUISS_CEE) et la puissance administrative (P_ADMIN) sont des variables fortement corrélées. Comme ces 3 variables mesurent la puissance du véhicule mais avec des unités de mesure différentes, cette corrélation semble cohérente. Pour l'analyse, seule la puissance DIN sera retenue[12]. De la même manière, la voie[13] avant (VOIE_AV), la voie arrière (VOIE_ARR), la largeur (LARG) et la longueur (LONG) du véhicule sont des variables fortement corrélées. Pour l'étude, seule la longueur du véhicule sera retenue. L'âge du véhicule et l'ancienneté d'acquisition du véhicule ne sont pas très bien représentés par le premier et le deuxième axe (cercle de gauche), néanmoins, on observe une forte corrélation de ces deux variables sur les

[12] La puissance DIN est exprimée en chevaux dits din ou vapeur et la puissance CEE est exprimée en kilowatts. La puissance administrative s'exprime en chevaux dits fiscaux et représente la puissance théorique du véhicule

[13] La voie est la distance entre les deux roues d'un même essieu.

représentations du deuxième et du troisième axe (cercle de droite). Seul l'âge du véhicule sera retenu pour l'analyse. **En conséquence, le nombre de variables quantitatives est réduit de 22 à 16 variables.**

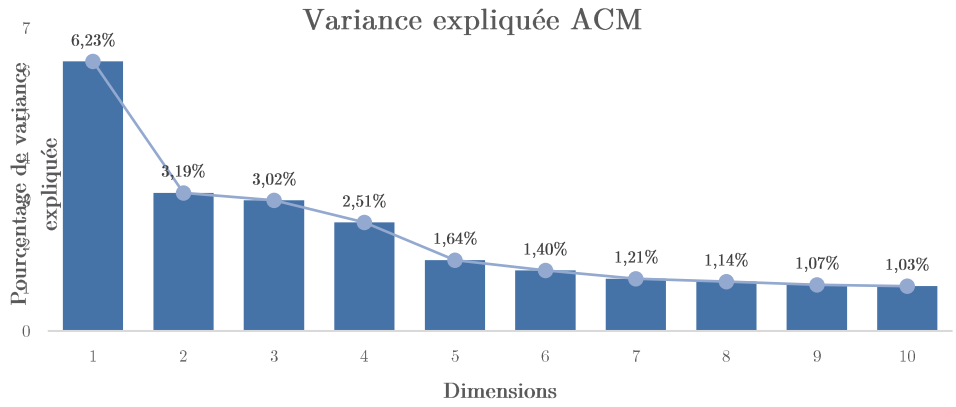
3.2. Analyse des liens parmi les variables qualitatives

L'analyse des correspondances multiples (ACM) a été utilisée pour analyser les relations entre les variables catégorielles. Bien que cette technique repose sur les mêmes principes théoriques que l'ACP, elle est basée sur les tableaux de contingence des modalités des variables au lieu d'utiliser un tableau de données standardisé comme dans l'ACP.

L'ACM applique l'algorithme de l'ACP au *tableau disjonctif complet (TDC)*. Le TDC est défini comme une matrice D dont les lignes correspondent aux individus (polices) et les colonnes aux modalités de chaque variable. Si l'élément $D_{ij} = 1$, cela signifie que l'individu i prend la modalité j de la variable J. Les autres éléments relatifs aux modalités de la variable J pour l'individu i prendront 0 comme valeur. Ainsi, le TDC permet de projeter les individus dans un sous-espace de dimension inférieure. A partir du TDC, il est également possible d'obtenir le *tableau de Burt*. Le *tableau de Burt* est une matrice carrée symétrique B d'ordre p (p est le nombre total de modalités de l'intégralité des variables). L'élément B_{ij} correspond au nombre d'individus prenant les modalités i et j. La *table de Burt* remplacerait la matrice de variance-covariance dans la représentation des individus dans l'avec une ACP.

3.2.1. Choix d'axes principaux à retenir pour l'ACM

Comme mentionné auparavant, les *valeurs propres* et donc le *pourcentage de variance expliquée* par chacune des composantes principales dans l'ACM sont obtenus à partir du tableau de Burt. Le graphique 3.2 montre le *pourcentage de variance expliquée* par les 10 premières composantes principales relatives à la base contenant seulement les variables catégorielles. La variance du nuage de points est expliquée à 22,5% par les 10 premiers axes et à 12,5% par les 3 premiers. Ces pourcentages sont relativement faibles par rapport à l'ACP conduit sur les variables quantitatives, ce qui signifie que les premiers axes de l'ACM expliquent très peu de la variance du nuage de points.



Graphique 3.2 - Histogramme du pourcentage de variance expliquée par les premiers 10 composantes principales (ACM)

3.2.2. Carré de liaisons

Le carré de liaisons est une représentation graphique de l'ACM qui permet d'identifier et de visualiser la corrélation entre les variables et les axes principaux de l'ACM. Étant donné que le but de cette analyse est de réduire le nombre de variables catégorielles, le principal intérêt du carré de liaison est de représenter les variables et non leurs modalités. Les coordonnées correspondent au carré du rapport de corrélation entre les variables et les composantes principales.

Équation 3.4: Calcul du carré du rapport de corrélation entre une variable y et une composante principale u

$$\eta^2(u,y) = \frac{SCE_{inter}}{SCE_{totale}} = \frac{\sum_j \sum_{i \epsilon J_j} (u_{ij} - \bar{u})^2}{\sum_j \sum_{i \epsilon J_j} J_j \times (\bar{u}_{.j} - \bar{u})^2}$$

Où,

u_{ij} = *Valeur prise par la composante principale u pour l'individu i et la modalité j de la variable y*

$\bar{u}$ = *Valeur moyenne de la composante principale u*

$\bar{u}_{.j}$ = *Valeur moyenne de la composante principale u pour la modalité j*

J_j = *Nombre d'individus appartenant à la modalité J*

SCE_{inter} = *Variabilité inter − classes*

SCE_{totale} = *Variabilité totale*

Le carré du rapport de corrélation correspond au pourcentage de variabilité de la composante principale u attribuable aux différences entre modalités de la variable y (PAGES, 2010). Le carré du rapport de corrélation est une valeur entre 0 et 1 et s'interprète comme la corrélation linéaire utilisée dans l'ACP, 0 lorsque la variable et la composante principales ne sont pas liées et 1 lorsqu'elles sont parfaitement liées.

La figure 3.2 montre le carré de liaisons pour les deux premiers axes principaux. Les variables positionnées loin de l'origine du graphique sont fortement corrélées avec les axes et donc mieux représentées par l'ACM. En outre, les variables dont les coordonnées sont proches sont corrélées entre elles. Même si seulement 9,4% de la variance est expliquée par les deux premiers axes, il est possible d'observer des regroupements entre les variables (cercles bleu et violet dans le graphique ci-dessous) dont la représentation est satisfaisante (coordonnées éloignées de l'origine). Le cercle bleu regroupe les variables liées aux positionnement des airbags et le cercle violet regroupe les variables liées au moteur, la boite de vitesse et les freins du véhicule (l'annexe A.1 fournit la définition de l'acronyme pour chacune des variables).

L'analyse des liaisons obtenue par l'ACM sera complétée et confirmée par une mesure de la dépendance des variables : le V de Cramer. Ces deux analyses permettront de sélectionner les variables catégorielles finalement retenues pour la modélisation.

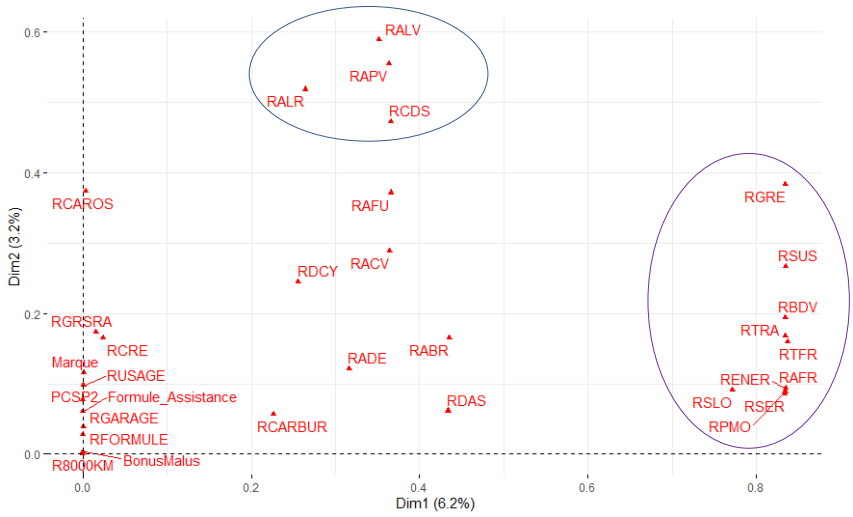


Figure 3.2 - Carré de liaisons entre les axes principaux et les variables catégorielles (ACM)

3.2.3. V de Cramer

Le V de Cramer mesure l'intensité de la relation entre deux variables en utilisant les test d'indépendance de X^2. Le calcul du V de Cramer pour un couple des variables est défini de la manière suivante :

Équation 3.5: Calcul du V du Cramer

Soit un couple de variables catégorielles x et y prenant les modalités respectives $x_1,\ldots,x_r$ et $y_1,\ldots,y_k$. Ces variables sont représentées par un jeu de données de taille n dont n_i, n_j et n_{ij} correspondent respectivement au nombre d'observations prenant la modalité i de la variable x, au nombre d'observations prenant la modalité j de la

variable y et au nombre d'observations prenant les deux modalités i et j sur les deux variables.

$$X^2 = \sum_{ij} \frac{(n_{ij} - \frac{n_i n_j}{n})^2}{\frac{n_i n_j}{n}}$$

$$V = \sqrt{\frac{\frac{X^2}{n}}{min(k-1, r-1)}}$$

D'après la littérature existante sur le V de Cramer (IBM, 2022), l'intensité de la relation entre deux variables peut s'interpréter de la manière suivante :

Valeur du V	Intensité de la relation
V ≤ 0,2	Faible
0,2 < V ≤ 0,6	Moyenne
V> 0,6	Forte

La figure ci-dessous représente la matrice V de Cramer pour les variables étudiées (le lecteur pourra retrouver les acronymes en annexe 1). L'intensité de la relation (comprise entre 0 et 1) est représentée par une échelle de bleus : le plus clair représente une absence de relation et le plus foncé une relation forte.

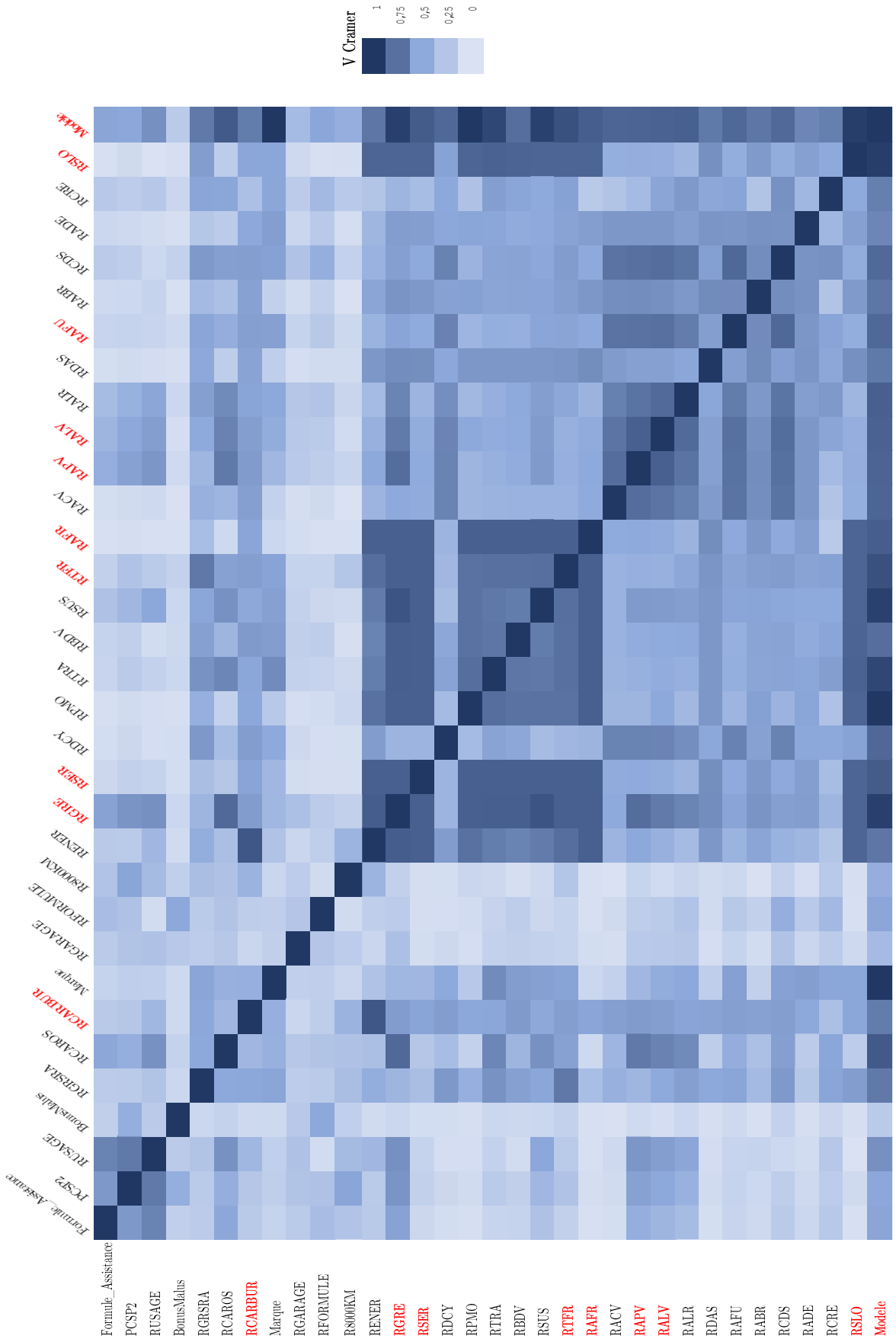


Figure 3.3 - Matrice de corrélation des variables qualitatives (d'après le V de Cramer)

A partir de la matrice de corrélation présentée ci-dessus, des relations à forte intensité (V de cramer supérieur à 0,6) ont été identifiées, notamment entre les variables faisant partie des regroupements observés dans le carré des liaisons généré par l'ACM. Les variables soulignées en rouge dans la matrice de corrélation étant très corrélées à d'autres variables ont été retirées de la base des données. Par exemple, la variable type de carburant (RCARBUR) est très fortement corrélée à la variable type d'énergie (RENER). En outre, la variable modèle est fortement corrélée à plusieurs variables de la base de données, notamment la marque, le type de carrosserie, le positionnement des airbags, le type de boîte de vitesses, etc. De plus, comme le portefeuille compte plus de 500 modèles, il était complexe de gérer une variable avec autant de catégories. Ainsi, **le nombre des variables qualitatives est réduit de 33 à 23** (l'annexe A.1 fournit la définition de l'acronyme pour chacune des variables).

3.3. Matrice de corrélation Phi-K

En 2019, M. Baak, R. Koopman, H. Snoek et S. Klous ont défini le score de corrélation Phi-K (ϕ_K) comme une nouvelle mesure d'intensité de la relation entre deux variables, allant de 0 à 1, où 0 signifie qu'il existe aucune relation et 1 que la relation est parfaite. Il repose sur un ajustement du test d'indépendance X^2. Si les deux variables sont tirées d'une distribution normale bivariée, alors la valeur de ϕ_K reviendra à la valeur absolue du coefficient de corrélation de Pearson.

Cette nouvelle mesure de corrélation présente les avantages suivants :

- Elle s'applique aux variables quantitatives et qualitatives réunies. Jusqu'à présent, les relations entre les variables qualitatives et quantitatives n'ont pas encore été analysées.
- Les relations non-linéaires sont capturées. Les méthodes d'analyse factorielle ACP et ACM utilisées dans les sections précédentes traitent uniquement les relations linéaires entre variables.

Le score de corrélation Phi-K (ϕ_K) n'est pas défini par une formule fermée. Pour avoir plus d'information sur son calcul, le lecteur peut se référer à l'article *« A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics »* (Baak, Koopman, Snoek, & Klous, 2019).

La matrice de corrélation basée sur le score de corrélation Phi-K (ϕ_K) de chaque paire de variables est présentée ci-dessous. Les variables surlignées en rouge dans la matrice ont été retirées de l'analyse car elles présentaient une corrélation supérieure à 0,9 avec d'autres variables. Par exemple, le RGRSRA est une variable liée à la dangerosité du véhicule. Elle est spécifique à la base de données SRA et n'est donc pas applicable à d'autres pays que la France. Etant donné qu'elle est fortement liée à la vitesse maximale (VIT_MAX) et à la puissance DIN (PUISS_DIN), qui sont des

variables universelles, et que le portefeuille du Groupe Europ Assistance est international, la variable RGRSRA a été retirée de l'analyse.

A l'issue de l'étape de présélection des variables, 36 variables ont été retenues, dont 14 quantitatives et 22 qualitatives. On procède alors à l'analyse de la variable géographique par la construction d'un zonier qui sera intégré comme variable qualitative dans la base de données utilisée pour la modélisation.

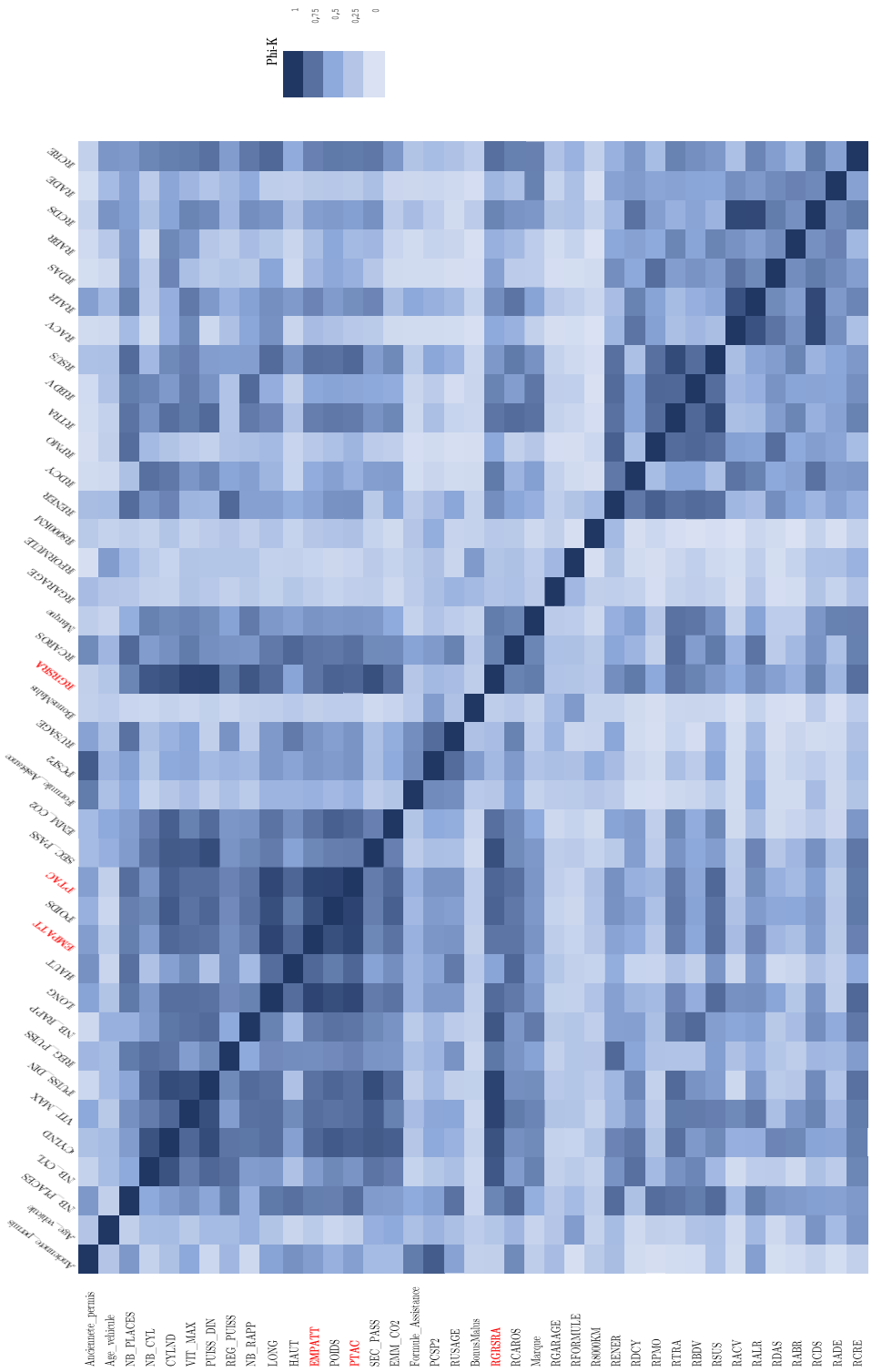


Figure 3.4 - Matrice de corrélation de l'intégralité des variables (d'après le score Phi-K)

L'objectif de ce chapitre était d'identifier les liens entre les variables afin d'éliminer les informations redondantes, c'est-à-dire les variables fortement corrélées entre elles. À cette fin, trois étapes de présélection des variables ont été appliquées avant la modélisation :

- Analyse des liens entre les variables quantitatives à l'aide de l'ACP.
- Analyse des liens entre les variables quantitatives à l'aide de l'ACM et du V de Cramer.
- Analyse des liens entre les variables quantitatives et qualitatives à l'aide de la matrice de corrélation PHI-K.

A l'issue de l'étape de présélection des variables, 36 variables ont été retenues, dont 14 quantitatives et 22 qualitatives. La réduction de dimension de la base des données permet par la suite, de faciliter son exploitation en termes de temps de calcul et complexité. On procède alors à l'analyse de la variable géographique par la construction d'un zonier qui sera intégrée comme variable qualitative dans la base de données utilisée pour la modélisation.

4. Définition des variables de type *zonier*

L'objectif de cette section est d'intégrer dans l'étude les disparités géographiques qui peuvent exister au niveau de la fréquence et le coût moyen du portefeuille étudié par la création d'un *zonier*. Le *zonier* est une variable qualitative dont chaque modalité représente le niveau de risque géographique des assurés. Le principe est de créer une variable *zonier* par type de risque (fréquence et sévérité) et par fait générateur (panne et accident) qui permette de diviser le territoire où se trouve le portefeuille en différentes zones de risque sur la base d'une granularité définie.

4.1. Représentation géographique des données empiriques

La base de données fournie par l'assureur partenaire contenait le code de la commune de stationnement du véhicule lié à chaque police du portefeuille. Le code commune également appelé code INSEE correspond à une des mailles les plus fines pour représenter le risque géographique en assurance automobile, néanmoins, un regroupement par département a été réalisée afin de faciliter la visualisation de la fréquence et du coût moyen empiriques sur la carte de la France.

Les figures 4.1 et 4.2 présentent respectivement pour les pannes et les accidents la fréquence empirique (à gauche) et le coût moyen (à droite) par département sur la carte de la France métropolitaine. Les différences géographiques observées soulignent l'impact de la localisation géographique sur la fréquence et la sévérité des sinistres. Par conséquent, le *zonier* sera une variable discriminante dans la modélisation de la prime pure.

Maintenir une granularité au niveau du département reviendrait à créer une variable *zonier* avec 96 modalités, soit le nombre de départements français métropolitains. Une maille aussi fine rendrait le modèle de tarification très complexe et moins lisible, c'est pourquoi il a été décidé de réduire le nombre de zones de risque en utilisant la méthode de la *classification ascendante hiérarchique (CAH).*

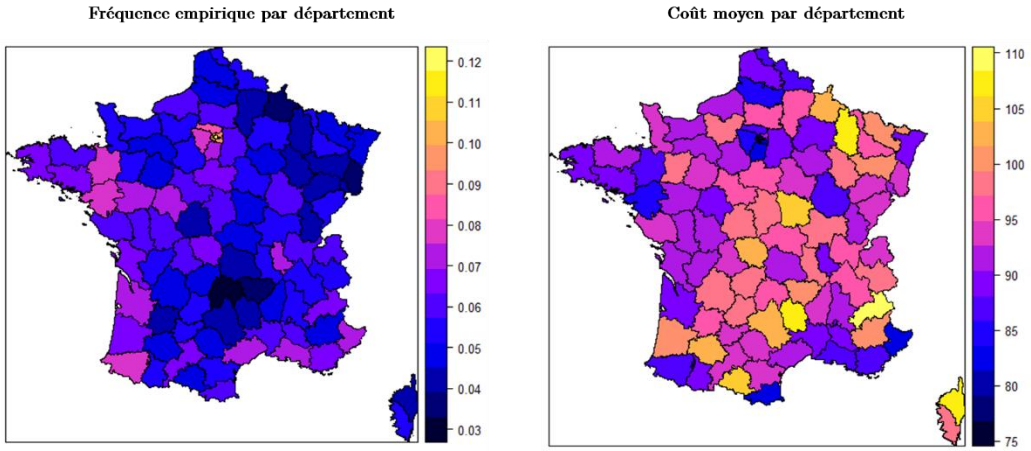


Figure 4.1 - Représentation graphique des données empiriques par département (pannes)

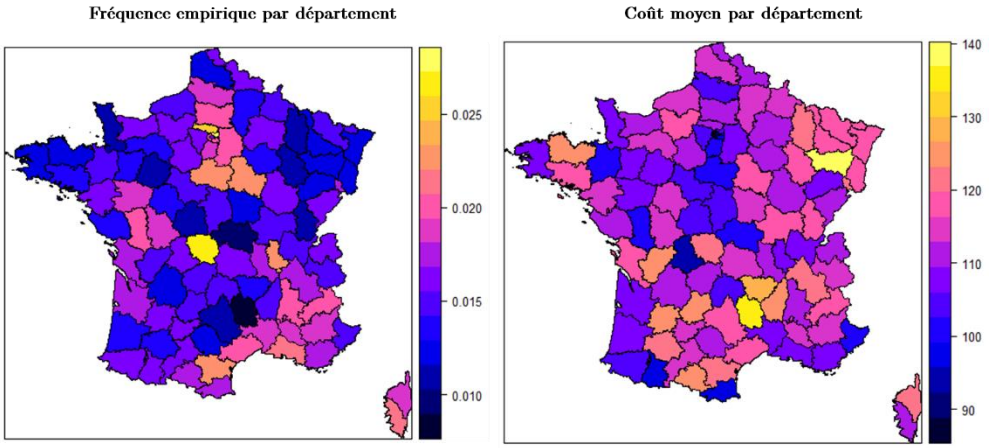


Figure 4.2 - Représentation graphique des données empiriques par département (accidents)

4.2. Définition des zones de risque

L'objectif de la variable *zonier* est de diviser le territoire en zones de risque similaires, c'est-à-dire, avec une fréquence et une sévérité semblables. Pour atteindre cet objectif, l'algorithme de *classification ascendante hiérarchique (CAH)* sera appliqué à la base des données contenant la fréquence et le coût moyen par fait générateur et par département. La CAH a été choisie parmi plusieurs méthodes de classification car elle effectue le regroupement sans avoir besoin de fixer le nombre de classes au préalable. En effet, cette méthode a l'avantage de représenter la classification sous forme d'arbre, ce qui facilite le choix du nombre optimal de zones. C'est optimal ici puisque le nombre de zones de risque n'est pas défini à l'avance.

4.2.1. Le principe de la CAH[14]

La *classification* est l'action de former des groupes ou des classes, et les *classes* sont définies comme un ensemble d'individus ayant des caractéristiques communes. La *classification ascendante hiérarchique (CAH)* est une méthode qui s'intéresse principalement aux jeux des données avec des variables quantitatives. Elle cherche à construire un arbre hiérarchique pour observer comment les individus sont organisés. L'arbre construit par la CAH permet de :

- Faire ressortir les liens hiérarchiques entre individus ou groupe d'individus.
- Repérer le nombre de classes optimal au sein de la population étudiée.

Pour effectuer une classification, il est nécessaire de définir une mesure de similarité entre les individus afin de définir les classes. Pour la création des zoniers, la distance euclidienne sera utilisée car c'est un indicateur commun et facile à

[14] Section basée sur le cours de classification ascendante hiérarchique (CAH) du professeur François Husson du Laboratoire de mathématiques appliquées – Agrocampus Rennes (Husson F. , 2014)

interpréter lors de la visualisation des données. De plus, la distance euclidienne sera utile pour faire le lien avec les méthodes d'analyse factorielle détaillées plus haut dans ce document (section 2.2.2), qui fournissent également des représentations euclidiennes des individus.

Équation 4.1: Calcul de la distance euclidienne entre deux points

Soit p et p' deux points de coordonnées (x, y) et (x', y'), leur distance euclidienne est définie par la formule suivante :

$$\|p - p'\| = \sqrt{(x - x')^2 + (y - y')^2}$$

De même, il est nécessaire de définir la similarité entre les groupes d'individus qui peut être mesurée par plusieurs critères : le saut minimum ou lien simple, le lien complet ou l'indice de Ward. C'est ce dernier qui sera utilisé dans cette étude car il permet d'obtenir les classes les plus homogènes possibles en minimisant la diminution de la variance interclasse à chaque itération lors de la construction de l'arbre hiérarchique. En effet, la variance interclasse ne peut que diminuer lorsque deux ensembles d'individus sont regroupés. Par équivalence, l'indice de Ward permet de maximiser la variance interclasse pour obtenir des classes bien distinctes.

Lorsque l'on veut regrouper les individus de la classe a et de la classe b, on cherche à ce que l'inertie de l'agrégation des deux classes soit la plus proche de la somme des inerties de chacune des classes. La somme des inerties de chacune des classes peut être définie en fonction de l'inertie de l'agrégation des deux classes de la façon suivante :

Équation 4.2: Calcul de la somme des inerties de la classe a et de la classe b

$$Inertie(a) + Inertie(b) = Inertie(a \cup b) - \delta_{WARD}(a, b)$$

Où,

$Inertie\ (a) = Variabilité\ totale\ de\ la\ classe\ a$

$Inertie\ (b) = Variabilité\ totale\ de\ la\ classe\ b$

$Inertie\ (a \cup b) = Variabilité\ totale\ de\ l'agrégation\ des\ classes\ a\ et\ b$

$\delta_{WARD}(a, b) = Variation\ de\ la\ variance\ interclasse\ entre\ les\ classes\ a\ et\ b\ (indice\ de\ Ward)$

Le CAH cherchera donc à minimiser l'indice Ward défini comme suit :

Équation 4.3: Calcul de l'indice de Ward

$$\delta_{WARD}(a, b) = \frac{m_a m_b}{m_a + m_b} d^2(g_a, g_b)$$

Où,

$m_a\ et\ m_b = Nombre\ d'individus\ de\ la\ classe\ a\ et\ nombre\ d'individus\ de\ la\ classe\ b$

$d^2(g_a, g_b) = Distance\ euclidienne\ entre\ les\ centres\ de\ gravité\ de\ la\ classe\ a\ et\ de\ la\ classe\ b$

La minimisation de l'indice de Ward permettra de regrouper des individus de faible poids et éviter les effets de chaine[15] (premier terme de la formule ci-dessus) ainsi que de regrouper des classes ayant des centres de gravité proches (deuxième terme de la formule ci-dessus).

4.2.2. Résultats de la CAH pour la création des zoniers

Les zoniers seront construits sur 4 tableaux de données dont les individus correspondent aux départements où sont stationnés les véhicules du portefeuille étudié et dont les variables sont présentées dans le tableau 4.1

Tableau	Fait Générateur	Type de Risque	1er Variable de classification	2ème variable de classification
1	Panne	Fréquence	Somme de l'exposition	Moyenne de la fréquence empirique
2	Panne	Sévérité	Somme du coût de sinistre	Moyenne du coût moyen empirique
3	Accident	Fréquence	Somme de l'exposition	Moyenne de la fréquence empirique
4	Accident	Sévérité	Somme du coût de sinistre	Moyenne du coût moyen empirique

Tableau 4.1 - Variables de classification pour la création des zoniers

Une analyse en composantes principales (ACP) a été effectuée avant la classification. La CAH réalisée pour chacun des tableaux de données est basée sur les deux premières composantes principales générées par l'ACP. Les arbres hiérarchiques et les plans factoriels illustrant le processus d'agrégation des classes pour la construction des zoniers à intégrer dans la modélisation montrent que les départements peuvent être regroupés en 3 zones sauf pour la fréquence des pannes pour laquelle 4 zones ont été identifiées (Annexe B).

Zoniers relatifs à la fréquence

Que ce soit pour les pannes ou pour les accidents, les zones représentent un niveau de risque progressif, la zone 1 ayant la fréquence la plus faible et les zones 3 respectivement pour les accidents et 4 pour les pannes ayant la fréquence la plus élevée. En ce qui concerne les pannes, on observe une similarité entre les classes 2 et 3 avec un niveau de fréquence très similaire et relativement alignée à la moyenne globale. Ce qui différencie ces deux classes est l'exposition moyenne par département. La zone 2 contient 44 départements et contribue à hauteur de 40% de l'exposition globale alors que la zone 3 ne contient que 9 départements et contribuent à hauteur de 33% à l'exposition globale (cf. tableau 4.2 et tableau 4.3). L'exposition moyenne par département est largement supérieure dans la zone 3 (~ 279 611) que dans la zone 2 (~7 700).

[15] Regroupement des individus des proche en proche ce qui donne un arbre hiérarchique de très mauvaise qualité.

Zone	Fréquence	Nb. Sinistres	Exposition	Nb Départements
Zone 1	4,56%	8 735	191 562	39
Zone 2	6,25%	21 116	337 967	44
Zone 3	6,26%	17 512	279 611	9
Zone 4	10,51%	3 217	30 603	4
Total	**6,02%**	**50 580**	**839 743**	**96**

Tableau 4.2 - Fréquence empirique par zone (pannes)

Zone	Fréquence	Nb. Sinistres	Exposition	Nb Départements
Zone 1	1,42%	5 485	385 211	60
Zone 2	1,74%	4 867	279 611	9
Zone 3	2,01%	3 514	174 921	27
Total	**1,65%**	**13 866**	**839 743**	**96**

Tableau 4.3 - Fréquence empirique par zone (accidents)

Dans le cas des pannes, l'exposition est relativement concentrée géographiquement dans les zone 3 et 4 puisque 37% de l'exposition globale est située dans 14% des départements (13 départements sur 96). En revanche ces deux zones ont une fréquence de sinistralité plus élevée puisqu'elles représentent environ 41% de la sinistralité globale du portefeuille. Ceci peut s'expliquer par la densité de la population dans les départements appartenant aux zones les plus risquées. On observe d'ailleurs une forte similarité entre le zonier représenté sur la carte de la France (cf. figure 4.4) et la carte représentant la densité démographique par département (Annexe B.3).

Pour les accidents, les constats sont similaires. L'exposition est principalement concentrée dans les zones 2 et 3 avec une contribution de 54% de l'exposition globale couvrant 38% des départements (36 départements sur 96). Ces deux zones ont également la sinistralité la plus élevée contribuant à hauteur de plus de 60% de la sinistralité globale. Des rapprochements peuvent être fait effectués entre le zonier réalisé pour la fréquence des accidents et la carte de la mortalité routière par département. En effet, on observe que les départements ayant une mortalité élevée correspond à des départements classés en zone 3 (Annexe B.4).

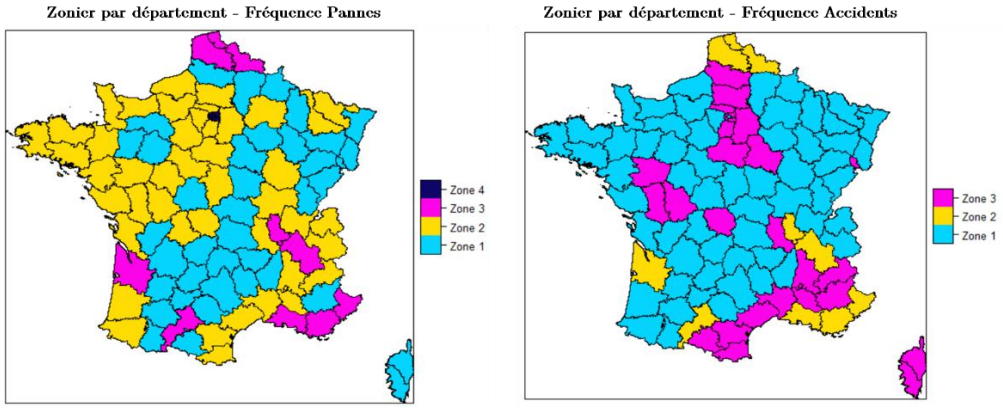


Figure 4.3 - Zonier par département des fréquences liées aux pannes (à gauche) et aux accidents (à droite)

Bien que le CAH n'ait été calibré que sur les variables liées à la sinistralité, il est également intéressant d'examiner les autres variables explicatives pour identifier les relations avec le zonier. On constate que la moyenne des variables explicatives observées reste stable à l'exception de l'ancienneté du permis où l'on observe des anciennetés différentes par zone de risque (cf. Annexe B.5 et B.6). Cette interaction potentiellement significative sera prise en compte dans la modélisation.

Zoniers relatifs au coût moyen

De même que pour les zoniers de fréquence, les zoniers de coût moyen représentent un niveau progressif de sévérité, la zone 1 ayant le coût moyen le plus faible et la zone 3 le coût moyen le plus élevé.

Contrairement à la fréquence, c'est la zone la plus risquée, la zone 3, qui est composée de plus de 60% des départements dans le cas des pannes comme des accidents (voir tableau 4.4 et 4.5).

Zone	Coût Moyen	Coût Total	Nb Sinistres	Nb Départements
Zone 1	88,46 €	1 447 377 €	16 362	8
Zone 2	86,54 €	1 494 416 €	17 269	26
Zone 3	96,23 €	1 630 964 €	16 949	62
Total	**90,41 €**	**4 572 757 €**	**50 580**	**96**

Tableau 4.4 - Coût moyen empirique par zone (pannes)

Zone	Coût Moyen	Coût Total	Nb Sinistres	Nb Départements
Zone 1	109,88 €	534 803 €	4 867	9
Zone 2	100,49 €	225 696 €	2 246	24
Zone 3	114,85 €	775 577 €	6 753	63
Total	**110,78 €**	**1 536 076 €**	**13 866**	**96**

Tableau 4.5 - Coût moyen empirique par zone (accidents)

Ceci est dû au fait que les départements classés dans cette zone correspondent à des départements moins denses démographiquement et donc caractérisés par la présence de plus de zones rurales que de zones métropolitaines. Le réseau de partenaires d'une compagnie d'assistance fournissant des prestations de base dans ce type de zones est moins granulaire ce qui amène aux dépanneurs à parcourir plus de distance pour attendre le bénéficiaire qui a eu une panne ou un accident. Ce phénomène augmente le coût de la prestation car la distance moyenne entre le garagiste et le lieu du sinistre est plus importante dans les zones rurales que dans les zones métropolitaines où le réseau de partenaires est plus dense. En outre, le nombre de remorquages dans les zones rurales est supérieur au nombre de dépannages sur place, ce qui a également une incidence sur le coût moyen des prestations de base.

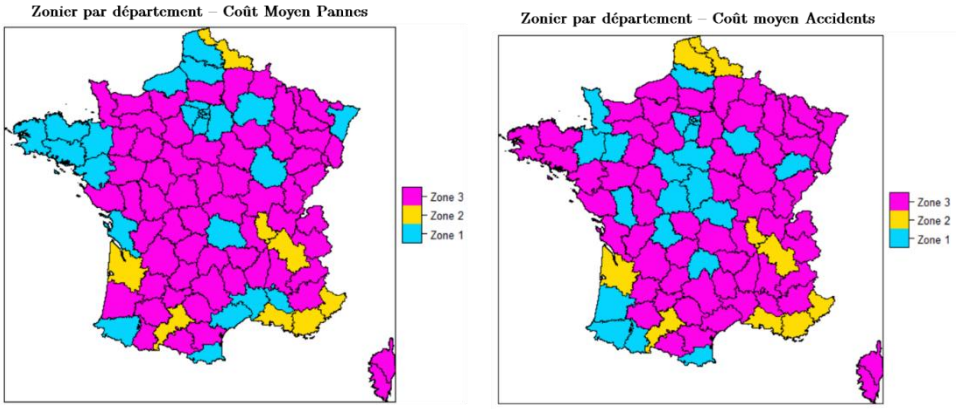


Figure 4.4 - Zonier par département du coût moyen lié aux pannes (à gauche) et aux accidents (à droite)

Le fait de n'inclure que les variables liées à la sinistralité (nombre de sinistres, coût total des sinistres, exposition, fréquence empirique et coût moyen empirique) impliquent que les zoniers et les autres variables explicatives sont *iid*. Une interaction potentielle a été déjà repérée avec la variable *Ancienneté de permis*, néanmoins, une matrice de corrélation Phi-k a été construite en incorporant les 4 variables zoniers pour tester cette hypothèse.

Le niveau de corrélation observé ne permet pas d'affirmer avec certitude la dépendance entre les variables explicatives retenues pour la modélisation et les nouvelles variables zonières (voir Annexe B.7). Cependant, la variable marque présente un niveau moyen de corrélation avec les zoniers (environ 0,68), cette interaction potentielle sera donc également prise en compte dans la modélisation.

Une fois les variables zoniers définies, elles sont intégrées dans la base de données de l'étude et on procèdera à modéliser la fréquence et la sévérité.

5. Modélisation de la fréquence et la sévérité

Les analyses effectuées dans les sections précédentes ont permis de construire des ensembles de données appropriés pour la modélisation. L'objectif de cette section est de calibrer des modèles linéaires généralisés (GLM) pour estimer au mieux la fréquence et le coût moyen des deux faits générateurs (pannes et accidents). De même, une nouvelle prime pure sera calculée et comparée à celle estimée via un modèle de tarification simple.

La section se déroulera comme suit :

1. Introduction à Akur8, l'outil de modélisation utilisé.
2. Description des paramètres d'ajustement des modèles et des méthodes validation.
3. Présentation des modèles linéaires généralisés (GLMs) comme méthode d'ajustement des modèles de fréquence et de sévérité.
4. Définition des critères de performance pour la sélection des modèles.
5. Evaluation de la performance des modèles.
6. Application des modèles retenus pour le calcul de la prime pure.

5.1. Akur8, l'outil de modélisation utilisé

Akur8 est un outil de tarification actuariel pour l'assurance non-vie qui repose principalement sur les GLMs. Akur8 a un champ d'application très large, allant de l'exploration de données au calcul des primes commerciales et au développement de plans de tarification, y compris les tests de scénarios. Il dispose de 4 modules de modélisation différents avec des fonctionnalités variées :

1. Données : Préparer, personnaliser, inspecter et enrichir les bases des données avant de créer les modèles de risque.
2. Risque : Modéliser la fréquence et la sévérité en exploitant la puissance des GLMs.
3. Agrégation[14] : Agréger les modèles de fréquence et de sévérité en un seul modèle de risque pour calculer une prime pure complète.
4. Demande[16] : Modéliser séparément la propension à acheter et l'élasticité des prix.

L'avantage de cet outil est de pouvoir s'attaquer aux aspects techniques de la tarification en générant des modèles de risque plus rapidement et avec un pouvoir

[16] Ces modules ne seront pas utilisés dans le cadre de cette étude.

prédictif égal ou supérieur à celui d'autres logiciels statistiques existants sur le marché. Compte tenu de la taille des ensembles de données à modéliser dans cette étude, la puissance de calcul d'Akur8 était importante pour gagner en efficacité dans la modélisation.

5.2. Construction et validation des modèles

5.2.1. Type et usage des variables

Akur8 n'admet que des variables qualitatives de 2 types :

- Variables ordinales : variables dont les catégories ont un ordre spécifique. Toutes les variables numériques sont automatiquement converties en variables ordinales par l'outil.
- Variables nominales : variables avec des catégories sans ordre spécifique.

Le type de variable a un impact important sur la modélisation : pour les variables ordinales, les coefficients seront ajustés et lissés en tenant compte de l'ordre des modalités. Dans le cas contraire, chaque modalité sera traitée indépendamment.

5.2.2. Paramètres d'ajustement des modèles basés sur l'apprentissage automatique

Akur8 repose sur l'une des méthodes de modélisation les plus courantes : le GLM mais en adoptant une approche différente à celle des solution traditionnelles. L'outil permet d'ajuster des paramètres globaux des modèles plutôt que d'ajuster des coefficients individuels. Trois paramètres principaux doivent être définis :

Intervalle du nombre des variables retenues

Ce paramètre permet de définir le nombre de variables à intégrer dans le modèle. En effet, les modèles créés par l'algorithme de l'outil incluront un certain nombre de variables, appartenant à un intervalle choisi par l'utilisateur. **Dans le cas de cette étude, un intervalle entre 5 (modèles de complexité simple) et 20 variables (modèles de complexité moyenne) a été défini.** Ce choix a été fait dans le but de trouver un modèle performant tout en limitant le nombre de variables. Dans le contexte de la tarification des contrats BtoB, comme les variables tarifaires ne sont pas toujours disponibles et que des négociations doivent être menées avec le client - qui n'est pas un particulier mais une entreprise, il est important de privilégier l'interprétabilité et la simplicité du modèle tout en gardant un niveau de précision acceptable.

Parcimonie

Chaque valeur de parcimonie correspondra à un groupe de modèles avec le même nombre de variables. **Les 7 niveaux spécifiés pour la modélisation de cette étude** vont donc générer 7 groupes de modèles, chaque groupe ayant un nombre spécifique de variables appartenant à l'intervalle défini ci-dessus.

Ce paramètre permettra de définir le nombre de variables à retenir dans chacun des modèles à définir. Considérant que l'ajout d'une variable au modèle a un coût en termes de lisibilité et que la suppression d'une variable tend à réduire le pouvoir prédictif d'un modèle, les 7 niveaux de parcimonie définis permettront de trouver des modèles menant à un compromis entre interprétabilité et complexité (cf. Annexe C.1).

	Faible Complexité	Forte Complexité
Nombre de variables	Faible	Fort
Interprétabilité	Forte	Faible
Performance	Faible	Forte

Tableau 5.1 - Comparaison de niveau de complexité des modèles

Lissage

Toutes les données de la base de données sur laquelle repose la modélisation contiennent deux composantes, le signal, porteur de l'information nécessaire pour expliquer soit la fréquence, soit le coût moyen, et le bruit constitué d'erreurs aléatoires. Ce dernier est bien sûr indésirable, car il réduit l'exactitude et la précision de la modélisation. Le fait de disposer d'une base de données riche est souvent considéré comme un avantage mais peut aussi présenter des inconvénients si les données ont tendance à présenter beaucoup de bruit. Akur8 permet de définir le niveau de lissage des modèles permettant de trouver un second compromis, celui de la robustesse et de la sensibilité.

Pour la modélisation, **10 niveaux de lissage ont été définis**, ce qui signifie que l'outil générera, pour chaque niveau de parcimonie, 10 modèles différents. Ce choix a été fait pour avoir un spectre de modèles plus large, allant de modèles très sensibles (avec un risque élevé de de surajustement ou *overfitting*) à des modèles très robustes (avec un risque élevé de sous-ajustement ou *underfitting*).

Les paramètres de parcimonie et lissage contribuent à la définition du nombre de modèles créés : 70 modèles ont été initialement crées par l'outil (7 niveaux de parcimonie x 10 niveaux de lissage) sur l'ensemble complet de la modélisation (cf. Annexe C.2). Des modèles additionnels ont été créés dans le cadre de la validation croisée.

5.2.3. Techniques de validation appliquées

La validation est une étape essentielle dans la sélection des modèles et l'évaluation de leurs performances. Deux types de méthodes de validation ont été appliquées à la structure de modélisation décrite dans la section précédente :

- Validation non-croisée : La base de données de l'étude a été divisée de manière aléatoire en deux sous-ensembles ; le premier, appelé sous-ensemble d'apprentissage ou *train*, composé de 80% des données et le second, appelé sous-ensemble de validation ou *test*, contenant les 20% des données restantes. Le modèle a été construit en utilisant uniquement les données du sous-ensemble *train* et validé sur le sous-ensemble *test*.

- Validation croisée ou *k-fold* : Cette méthode a été appliquée uniquement au sous-ensemble *train* défini lors de la validation non-croisée. La procédure générale est la suivante (Brownlee, 2020) :

    1. L'ensemble des données doit être mélangée de façon aléatoire.

    2. Les données sont divisées en k sous-ensembles. Pour la modélisation $k = 4$ a été choisi pour que chaque groupe de données *train/test* soit suffisamment grand pour être statistiquement représentatif.

    3. Pour un sous-ensemble de données i :

        a. Le sous-ensemble i est considéré comme une base *test*.

        b. Les sous-ensembles restants sont considérés comme des bases *train*.

        c. Le modèle est ajusté sur l'ensemble *train* et évalué sur l'ensemble *test*.

    4. La moyenne des métriques de performance sur les $k = 4$ modèles est calculée, ce qui fournit des prédictions plus robustes.

L'annexe C.3 présente de manière graphique les techniques de validation non-croisée et croisée (*k-fold*). Dans le cadre de la validation croisée, 280 modèles additionnels ont été créés (7 niveaux de parcimonie x 10 niveaux de lissage x $k = 4$ sous-ensembles de validation croisée). En conclusion, on aura un spectre de 350 modèles parmi lesquels on devra choisir le plus adapté en termes de précision et interprétabilité.

5.3. Ajustement d'un modèle

5.3.1. La vraisemblance pour l'apprentissage

Les modèles crées correspondent à des *modèles linaires généralisées* (GLMs). Les GLMs reposent sur 2 hypothèses (Gorrand, 2021) :

- Pour un profil d'assuré X donné, la distribution du nombre et du coût de sinistres Y (donc $Y|X$) appartient à la famille exponentielle. Cette hypothèse est respectée dans le cas de cette étude car il a été montré dans la section

2.2.3 de ce document que les distributions du nombre et du coût de sinistres suivent respectivement une loi poisson et une loi gamma.

- L'espérance de Y sachant X est liée par une combinaison linéaire des composantes de X au travers d'une fonction bijective g :

  Équation 5.1 : Calcul de l'espérance $E[Y|X]$ sous un GLM

  $$g(E[Y|X]) = \beta_0 + \sum_i \beta_i \times x_i$$

  Cela signifie que :

  $$E[Y|X] = g^{-1}(\beta_0 + \sum_i \beta_i \times x_i)$$

  Où, g^{-1} correspond à la fonction de lien inverse (logarithme pour un modèle multiplicatif.

Ainsi, la modélisation GLM du nombre et du coût des sinistres par police, équivalant à la fréquence et au coût moyen lorsqu'ils sont rapportés à l'exposition et au nombre de sinistres respectivement, peut être définie comme suit :

Équation 5.2: Estimation de la fréquence dans le cadre d'un GLM

$$log(\frac{nb_{sinistres}}{Exposition}) \sim \beta_0 + \sum_i \beta_i \times x_i + e$$

$$log(nb_{sinistres}) \sim log(Exposition) + \beta_0 + \sum_i \beta_i \times x_i + e$$

En mettant l'expression $log(Exposition)$ en *offset*[17], l'estimateur suivant est obtenu :

$$\widehat{nb_{sinistres}} = exp(\beta_0 + \sum_i \beta_i \times x_i + e)$$

Équation 5.3: Estimation du coût moyen dans le cadre d'un GLM

$$log(\frac{C_{sinistres}}{nb_{sinistres}}) \sim \beta_0 + \sum_i \beta_i \times x_i + e$$

$$log(C_{sinistres}) \sim log(nb_{sinistres}) + \beta_0 + \sum_i \beta_i \times x_i + e$$

En mettant l'expression $log(Exposition)$ en *offset*[15], l'estimateur suivant est obtenu :

[17] Un élément de type *offset* dans le GLM correspond à un élément fixe dans le modèle pour lequel aucun coefficient ne sera attribué.

$$\widehat{CM_{sinistres}} = exp(\beta_0 + \sum_i \beta_i \times x_i + e)$$

Trouver un modèle signifie trouver les coefficients β_i appropriés, en tenant en compte qu'exclure une variable signifie que tous les coefficients liés à chacune de ses modalités seront nuls.

Les coefficients β sont estimés à partir de la méthode de maximum de vraisemblance :

Équation 5.4: Estimation des coefficients β par le maximum de vraisemblance

$$\hat{\beta} = ArgMax\ \rho(Y|\hat{Y}_\beta) = Argmax\ Log(\ L(X,Y,\beta))$$

$$\rho(Y|\hat{Y}_\beta) \sim Loi\ de\ poisson\ dans\ le\ cas\ du\ nombre\ des\ sinistres$$

$$\rho(Y|\hat{Y}_\beta) \sim Loi\ Gamma\ dans\ le\ cas\ du\ coût\ des\ sinistres$$

$$Log(\ L(X,Y,\beta) \sim Log - vraisemblence$$

Dans le cas des variables à valeur réelle (poids, longueur, puissance, âge, etc.), trouver les coefficients β en maximisant la log-vraisemblance signifierait une forte probabilité de surajustement ou *overfitting* du modèle. Cela s'explique par le fait que l'outil considère les variables numériques comme des variables qualitatives ordinales et donc, le nombre de paramètres à estimer serait très élevé (plus de 100 par variable). Des contraintes supplémentaires telles que les regroupements de modalités doivent être intégrées à la modélisation pour réduire la complexité.

5.3.2. Les tests statistiques pour la robustesse

Akur8 s'appuie sur une approche bayésienne pour effectuer les regroupements, une hypothèse *à priori* est prise et elle est vérifiée ou non par les données disponibles. *A priori*, l'outil suppose que tous les coefficients β sont constants. Des tests statistiques khi-deux (X^2) sont réalisés pour vérifier cette hypothèse :

Équation 5.5: Test statistique X^2 pour vérifier si deux coefficients β voisins sont constants

$$H_0: \beta_{ij} = \beta_{(i+1)j} \quad i \in X, j \in J$$

$$H_1: \beta_{ij} \neq \beta_{(i+1)j} \quad i \in X, j \in J$$

Où,

$$X = Groupe\ des\ variables\ ordinales$$

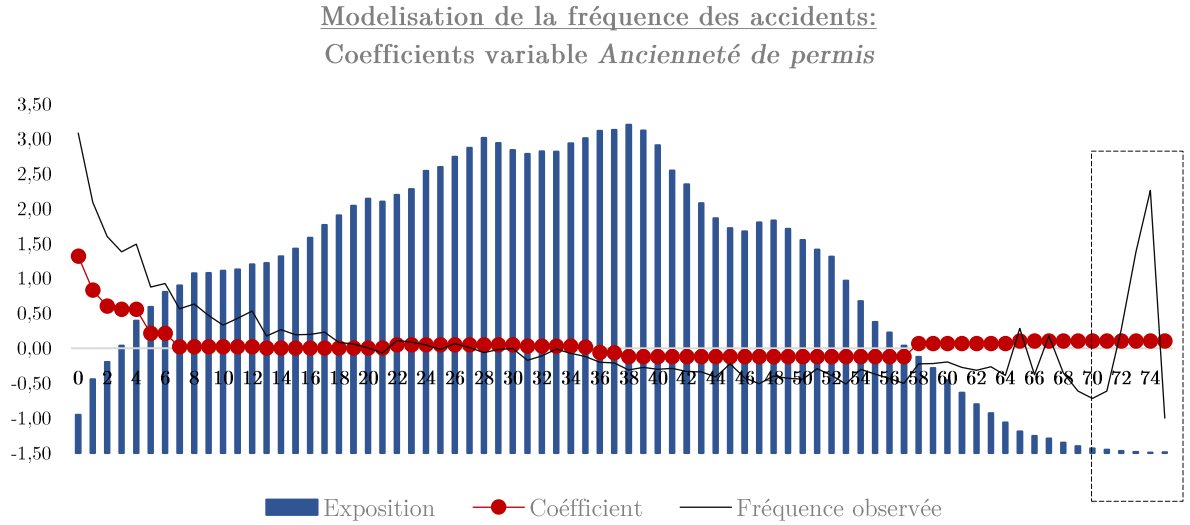
$$J = Groupe\ de\ modalités\ de\ la\ variable\ ordinale\ i$$

L'hypothèse nulle est confrontée statistiquement aux données disponibles, si elle est rejetée alors les coefficients voisins ne sont pas les mêmes, si elle n'est pas rejetée les coefficients voisins peuvent être considérés comme égaux et donc les modalités

peuvent être regroupées. Le rejet de l'hypothèse nulle dépend du niveau de significativité du test α. Une valeur faible de α regroupera très peu des modalités et donc estimera beaucoup de coefficient (modèles sensibles avec du bruit). Dans le cas contraire, une forte valeur de α créera de grands regroupements avec peu de coefficients (modèles robustes). La définition du niveau α est lié au paramètre de lissage défini auparavant, plus un modèle est lisse plus la valeur de α est grande.

Les regroupements fournis par Akur8 sont cohérents pour toutes les variables. Si un effet est significatif sur une variable, un autre effet avec le même niveau de significativité présent dans une autre variable est également considéré. Cela veut dire que toutes les modalités, pour toutes les variables, seront groupées avec les mêmes critères.

Néanmoins, dans certains cas, l'hypothèse *a priori* selon laquelle la fonction des coefficients est constante ne conduit pas toujours à des résultats fiables pour certaines variables et certaines modalités. Par exemple, lorsque les modalités ont une faible exposition, leur effet est souvent considéré comme non significatif et donc regroupé. C'est le cas de la variable ancienneté de permis dans la modélisation de la fréquence des accidents où l'on observe que les coefficients des anciennetés supérieures à 60 ans sont constants alors qu'en réalité à partir de 70 ans une sinistralité plus importante est observée (cf. Graphique 5.1). De même, lorsque l'exposition est élevée, les coefficients ont tendance à être indépendants.



Graphique 5.1 - Coefficients variable ancienneté de permis (modélisation pannes)

5.3.3. La sélection des variables pour la lisibilité

L'approche statistique décrite précédemment permet de contrôler l'*overfitting*, néanmoins, il est également important de rendre les modèles lisibles par des

interlocuteurs qui ne sont peut-être pas familiers avec l'orientation technique des méthodes utilisées dans cette étude. En utilisant l'intervalle de variables défini dans la section 5.2.2, Akur8 n'inclura que les variables les plus significatives dans les modèles. Ceci n'est pas destiné à améliorer le pouvoir prédictif des modèles mais à les rendre plus lisibles. Néanmoins, il peut arriver que certaines variables ne soient pas statistiquement significatives mais qu'elles le soient d'un point de vue actuariel. Afin d'éviter cela, une révision manuelle pour intégrer les variables pertinentes et laisser de côté celles qui le sont moins a été faite avant la modélisation.

5.4. Critères de performance

La sélection et comparaison des modèles GLM pour l'estimation de la fréquence et de la sévérité reposera sur l'analyse des critères de performance suivants :

5.4.1. Courbe de Lorenz

La courbe de Lorenz a été initialement créée pour évaluer l'inégalité de la distribution des revenus dans une population donnée. Dans le cas de cette étude, elle est utilisée pour évaluer la qualité des prédictions du modèle et sa capacité à segmenter le niveau de risque des profils du portefeuille étudié. Cette courbe est construite en triant les observations en fonction de leur prédiction par ordre décroissant, en traçant la valeur observée cumulée, en fonction de la prédiction triée.

Dans le domaine de l'économie sociale, lorsque la distribution des revenus est parfaitement alignée sur la distribution de la population, la courbe de Lorenz se traduit par une ligne à 45 degrés connue sous le nom de "courbe d'égalité" (Frees, Meyers, & Cummings, 2013). Dans le cadre de cette étude, une courbe de Lorenz égale à la courbe d'égalité signifiera que le modèle construit considère que tous les profils ont le même niveau de risque.

A titre d'exemple, la figure 5.1 montre la courbe de Lorenz basée sur la distribution des primes d'un portefeuille. La flèche marque le point où 60 % des assurés paient 40% des primes. La ligne à 45 degrés correspond à la "courbe d'égalité" qui représente le scenario où chaque assuré paie la même prime, la distribution des primes serait la même.

5.4.2. Coefficient de Gini

Le coefficient de Gini représente deux fois la valeur de la superficie entre la courbe de Lorenz et la courbe d'égalité - qui est représentée par la droite linaire sur la même figure 5.1. Le coefficient de Gini représente la capacité du modèle à trier les niveaux risque de plus haut au plus faible : un coefficient de Gini élevé signifie un pouvoir de classement plus important.

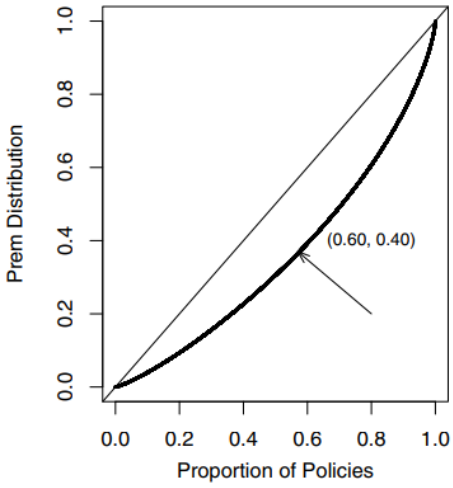


Figure 5.1 - Courbe de Lorenz basée sur la distribution des primes d'un portefeuille (Frees, Meyers, & Cummings, 2013)

Il est à noter que même avec une prédiction parfaite, il n'est pas possible d'avoir un indice de Gini de 100%. Il est donc intéressant d'analyser le coefficient de Gini normalisé. Celui-ci correspond au rapport entre l'indice de Gini du modèle étudié et l'indice qui serait obtenue avec une prédiction parfaite. Plus le coefficient normalisé est proche de 100%, plus le pouvoir de segmentation du modèle est bon.

5.4.3. La déviance

La déviance reflète la comparaison entre le modèle étudié et le modèle dit *saturé,* c'est-à-dire le modèle possédant autant de paramètres que d'observations et estimant donc exactement les données. Une déviance faible signifie que le modèle étudié s'ajuste bien aux données, néanmoins, l'addition des variables supplémentaire aux modèles diminue systématiquement la valeur de la déviance. L'ajout d'un paramètre augmente le nombre de degrés de liberté du modèle, indépendamment de la pertinence de la variable (BUCCI, 2021). Ainsi, la déviance est un critère qui ne tient pas en compte de la complexité du modèle. La déviance sera utilisée pour comparer des modèles avec le même nombre des variables.

Équation 5.6: Calcul de la déviance Poisson et Gamma

$$D\ (Poisson) = 2 \times (y \times ln\frac{y}{\hat{y}} - \ (y - \hat{y}))$$

$$D\ (Gamma) = 2 \times (ln\frac{\hat{y}}{y} - \frac{\hat{y} - y}{\hat{y}})$$

Où, y et $\hat{y}$ se réfèrent respectivement aux valeurs observées et prédites.

Les critères AIC et BIC sont souvent utilisés dans l'ajustement et la comparaison des GLMs car à différence de la déviance ils prennent en compte la complexité du modèle. L'AIC pénalise la déviance prenant en compte le nombre de paramètres inclus

dans le modèle. Le BIC pénalise la déviance par le nombre de paramètres et la taille de l'échantillon (BUCCI, 2021).

Néanmoins, dans cette étude, on s'appuiera sur les concepts de parcimonie, de lissage et de validation croisée pour gérer la complexité des modèles. Contrairement aux critères statistiques tels que l'AIC et le BIC, ils peuvent être utilisés pour tout type de modèle car ce sont des critères beaucoup plus généraux.

5.4.4. Pseudo-R^2

Le Pseudo-R^2 représente la réduction de la déviance apportée par le modèle par rapport à la déviance du modèle nul. La déviance du modèle nul est la déviance d'un modèle dans lequel il n'y a pas de prédicteurs.

Équation 5.7: Calcul du critère Pseudo-R2

$$Pseudo - R^2 = 1 - \frac{Deviance\ du\ modèle\ étudié}{Déviance\ du\ modèle\ nul}$$

5.4.5. La racine de l'erreur quadratique moyenne (RMSE)

La racine de l'erreur quadratique moyenne est la racine carrée de la somme des carrés de la différence entre l'observé et le prédit. Elle est équivalente à la log-vraisemblance d'un modèle gaussien. Cette métrique est faussée par les valeurs ou erreurs extrêmes. De même, il faut être critique dans l'analyse de la RMSE car bien qu'il s'agisse d'un critère d'évaluation du pouvoir prédictif du modèle, une RMSE trop faible peut signifier que les prédictions sont trop proches des observations et donc qu'il y a du surapprentissage.

Équation 5.8: Calcul du RMSE

$$RMSE = \frac{1}{\sum_i \omega_i} \times \sum_i \omega_i \times (y_i - \hat{y}_i)^2$$

Où, y_i et $\hat{y}_i$ se réfèrent respectivement aux valeurs observées et prédites et ω_i à l'exposition.

5.4.6. Courbe *lift*

La courbe *lift* est construite en triant les prédictions de la plus basse à la plus haute et en les regroupant en 20 groupes qui représentent chacun 5% des prédictions Pour chaque groupe, la prédiction moyenne du modèle et l'observation moyenne sont affichées, afin d'évaluer la qualité des prédictions.

5.5. Sélection et évaluation des modèles

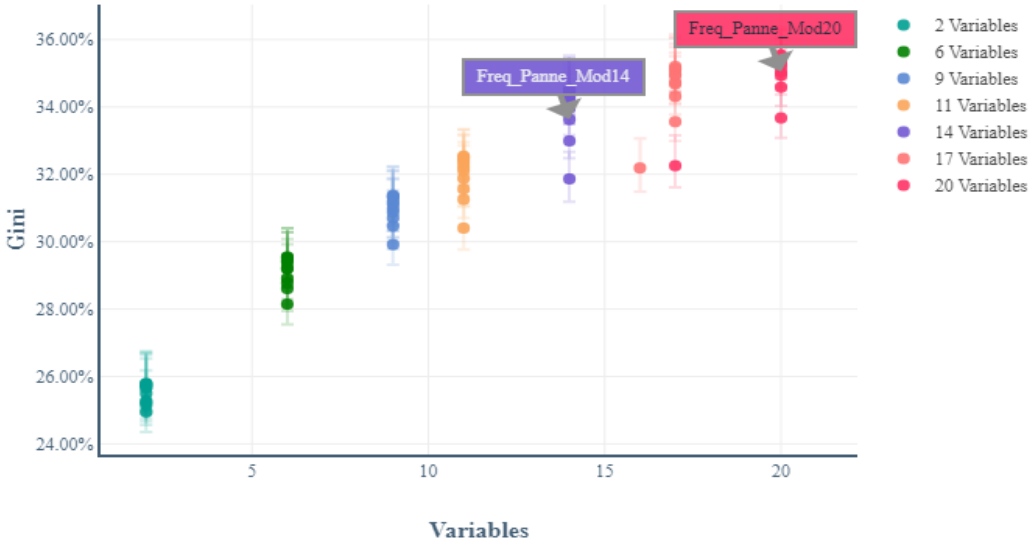
5.5.1. Modélisation de la fréquence des pannes

Le graphique 5.2 montre le résultat de la modélisation de la fréquence des pannes. Chaque point du graphique représente un modèle différent. A partir des 7 niveaux de parcimonie et 10 niveaux de lissage définis précédemment, 70 solutions au problème de modélisation ont été élaborées.

- L'axe horizontal représente le nombre de variables dans le modèle et l'axe vertical représente les performances moyennes estimées sur les 4 sous-ensembles des validation croisée.
- Le critère de performance présenté par défaut est le coefficient de Gini normalisé, néanmoins, les autres critères de performance décrits précédemment ont également été pris en compte dans la sélection.
- Les barres d'erreur représentent les variations de performances entre les sous-ensembles de validation croisée (et donc la fiabilité des performances mesurées).

Le but de ce graphique est de visualiser le compromis entre la complexité et la performance du modèle. Plus le modèle se trouve sur le côté droit du graphique, plus il est complexe puisqu'il contiendra plus de variables et sera donc plus difficile à interpréter. De la même manière, les modèles situés plus haut dans le graphique sont les plus performants (sur la validation croisée *k-fold*). Néanmoins, ce sont aussi les modèles les moins lisses et donc avec plus de coefficients, ce qui ajoute également de la complexité au modèle.

Dans le cadre de la modélisation de la fréquence des pannes, on procédera à la comparaison de deux modèles de complexité différente : un modèle à 14 variables et un autre à 20 variables. Ces modèles se trouvent sur le côté droit du graphique 5.2 et ont le même niveau de lissage. Le tableau 5.2 montre le coefficient de Gini normalisé et la RMSE et l'annexe C.4 montre les autres critères de performance.



Graphique 5.2 - Structure de modélisation de la fréquence des pannes (Source : Akur8)

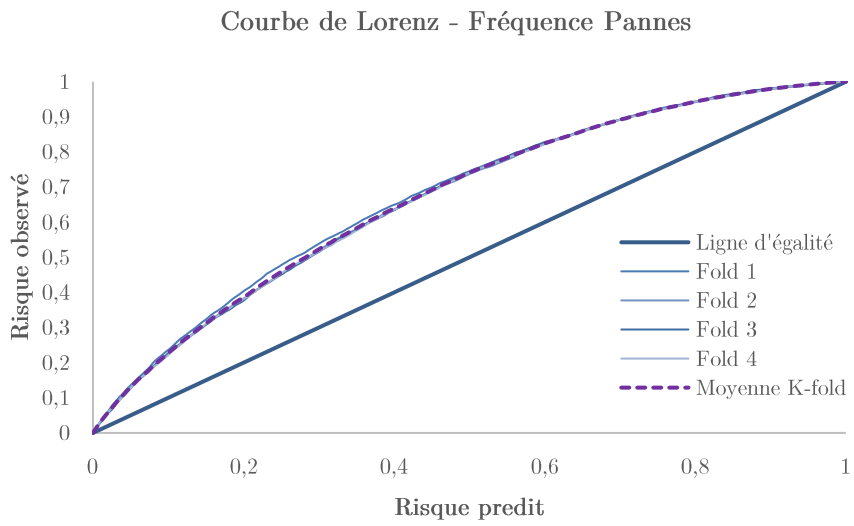
	Gini Normalisé		
	Validation K-Fold	Validation Globale	Var. (%)
14 Variables	34,87%	35,58%	-2,00%
20 Variables	36,34%	37,23%	-2,39%
Var. (%) Mod14 vs Mod20	-4,05%	-4,43%	

	RMSE		
	Validation K-Fold	Validation Globale	Var. (%)
14 Variables	0,3434	0,34	1,00%
20 Variables	0,3432	0,34	0,94%
Var. (%) Mod14 vs Mod20	0,06%	0,00%	

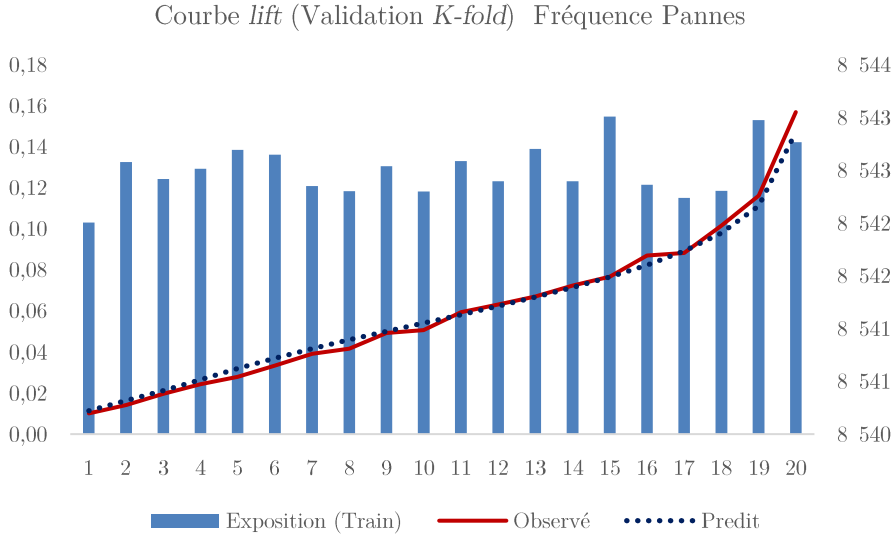
Tableau 5.2 Comparaison des modèles à 14 et 20 variables

On observe que le modèle à 20 variables segmente mieux les profils de risque, néanmoins, de manière non significative car la variation du coefficient de Gini normalisé n'est que de 4% en moyenne. De plus, la variation de la RMSE est presque nulle, ce qui signifie que les deux modèles ont le même pouvoir prédictif. Ainsi, le modèle à 14 variables est choisi pour modéliser la fréquence des pannes car à performance et niveau de lissage similaires, les modèles plus simples (à gauche du graphique 5.2) sont préférables.

Les graphiques 5.3 et 5.4 montrent la courbe de Lorenz et la courbe *lift* du modèle retenu pour l'estimation de la fréquence des pannes. En ce qui concerne la courbe de Lorenz, on observe par exemple qu'en moyenne les 20% cumulés du risque prédit le plus élevé représentent en fait environ 39% du risque observé, ce qui laisse supposer une segmentation favorable du portefeuille par le modèle. De plus, la variabilité entre les courbes de chaque *fold* et la courbe moyenne est faible ce qui apporte de la fiabilité à la modélisation. Etant donné que les courbes des valeurs observées moyennes et des prédictions moyennes sont très proches, on conclut que ce modèle a également un bon pouvoir prédictif.



Graphique 5.3 - Courbe de Lorenz du modèle de fréquence des pannes



Graphique 5.4 - Courbe de Lorenz du modèle de fréquence des pannes

Le tableau 5.3 montre les variables retenues comme discriminantes dans la modélisation de la fréquence des pannes.

No. Variable	Description	Type
1	Age du véhicule	Véhicule
2	Marque	Véhicule
3	Zonier	Géographique
4	Bonus-Malus	Assuré
5	Nombre des places	Véhicule
6	Nombre de rapports	Véhicule
7	Formule d'assistance	Contrat
8	Hauteur du véhicule	Véhicule
9	Type de garage pour le véhicule	Assuré
10	Poids du véhicule	Véhicule
11	Offre 8000 km (le conducteur s'engage à faire moins de 8000km par an)	Contrat
12	Etoiles de sécurité passive	Véhicule
13	Ancienneté du permis	Assuré
14	Année de survenance	Temporelle

Tableau 5.3 - Variables retenues pour la modélisation de la fréquence des pannes

Parmi les résultats notables, on remarque que la variable type d'énergie (essence, diesel, hybride, électrique) n'est pas retenue par le modèle comme variable discriminante ; contrairement à ce qu'on aurait pu penser puisque l'on observe dans des portefeuilles actuels des différences significatives sur la fréquence de pannes entre les véhicules thermiques, hybrides et électriques. Ceci est une limite du modèle qui peut s'expliquer par la faible exposition des véhicules non thermiques dans le portefeuille étudié, qui s'étend sur la période 2017 - 2019. Au vu des incitations de l'état pour augmenter la part de véhicules rechargeables dans le parc automobile français afin de réduire les émissions de CO2, la proportion de ces véhicules devrait augmenter progressivement. Il serait donc intéressant, en complément de cette étude, de réaliser une modélisation similaire sur un portefeuille comprenant une part plus importante de véhicules rechargeables, ou sur le même portefeuille sur une période plus récente (2020-2022 par exemple). Cela démontre également l'importance de retester la pertinence des modèles d'une année sur l'autre sur un segment en évolution comme l'automobile.

Analyse des interactions

Les interactions reflètent la relation entre deux ou plusieurs variables lorsqu'une variable exerce une influence sur d'autres. Créer une variable d'interaction dans un modèle signifie incorporer une nouvelle variable qui dépend des valeurs des variables en interaction. L'analyse des interactions parmi les variables explicatives dans la modélisation a été réalisée en deux étapes :

1. Détection et étude des interactions : Akur8 testera automatiquement toutes les interactions possibles dans le modèle et proposera à l'utilisateur celles qui sont le plus prédictives. Les interactions qui ont plus de sens d'un point de vue actuariel et business ont été intégrées à la modélisations (cf. Annexe C4.3).

2. Ajuster un nouveau modèle en intégrant les interactions sélectionnées.

Le modèle avec 3 interactions s'est avéré être le modèle montrant l'amélioration la plus significative en termes de performance (cf. tableau 5.4). Les interactions retenues sont les suivantes :

- Age du véhicule x Marque
- Nombre de cylindres x Marque
- Puissance DIN x Marque

On observe que la variable marque est présente dans toutes les interactions retenues comme significatives. Cela semble correspondre à la réalité, car la marque du véhicule a un fort impact sur les caractéristiques et les performances d'un véhicule. C'est ce qui fait les avantages concurrentiels de chaque constructeur automobile, et permet également d'évaluer le risque de panne. C'est à cause de cet effet qu'il n'est pas très précis de baser la tarification d'un nouveau constructeur automobile sur les données d'un autre constructeur automobile déjà dans le portefeuille.

	Gini Normalisé		
	Validation K-Fold	Validation Globale	Var. (%)
14 Variables	34,87%	35,58%	**-2,00%**
14 Variables + interactions	36,01%	36,69%	**-1,87%**
Var. (%) Mod14 vs Mod14+int	**-3,17%**	**-3,04%**	

	RMSE		
	Validation K-Fold	Validation Globale	Var. (%)
14 Variables	0,3434	0,34	**1,00%**
14 Variables + interactions	0,3433	0,34	**0,97%**
Var. (%) Mod14 vs Mod14+int	**0,03%**	**0,00%**	

Tableau 5.4 - Comparaison du modèle à 14 variables et du modèle à 14 variables + interactions (pannes)

Le tableau 5.4 montre que comme observé dans la comparaison avec le modèle à 20 variables, le gain de performance en termes de pouvoir de segmentation et prédictif avec le modèle incorporant les interactions n'est pas significatif. Par conséquent, on préfèrera retenir le modèle le plus simple, celui à 14 variables. L'annexe D.1 montre, à titre d'information, les graphiques représentant la fréquence observée, la fréquence prédite, les coefficients et l'exposition des variables considérées comme les plus discriminantes : l'âge du véhicule, la marque et la zone.

5.5.2. Modélisation de la fréquence des accidents

La même approche a été entreprise pour la modélisation de la fréquence des accidents. La figure 5.3 présente le spectre des modèles créés par Akur8. On peut observer que les barres d'erreur pour chaque niveau de parcimonie sont plus larges que celles observées pour la fréquence des pannes, ce qui signifie que la performance des modèles est plus volatile et donc légèrement moins fiable que celle observée pour les pannes. Néanmoins, il est compliqué de comparer deux modèles essayant de prédire des variables cibles différentes, c'est pourquoi, on s'est concentré sur l'analyse et comparaison des modèles disponibles prédisant uniquement la fréquence des accidents.

Un modèle à 14 variables a été également sélectionné dans le cas des accidents, comme observé dans la figure 5.3, il s'agit du modèle ayant le meilleur compromis entre performance et complexité.

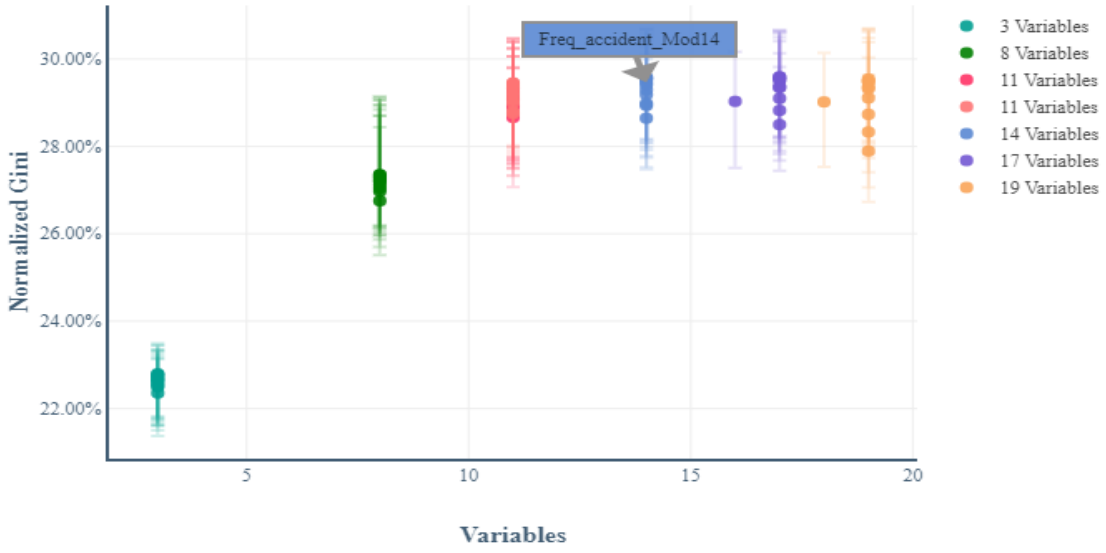
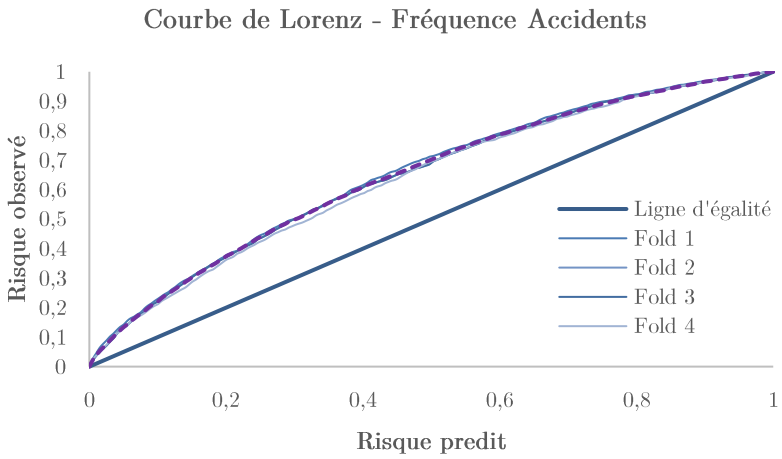


Figure 5.2 - Structure de modélisation de la fréquence des accidents (Source : Akur8)

Le tableau 5.5 présente les critères de performance appliqués aux sous-ensembles de validation *K-Fold* et Global. De même, les graphiques 5.5 et 5.6 montrent la courbe de Lorenz et la courbe *lift* construites de manière similaire sur les sous-ensembles de validation.

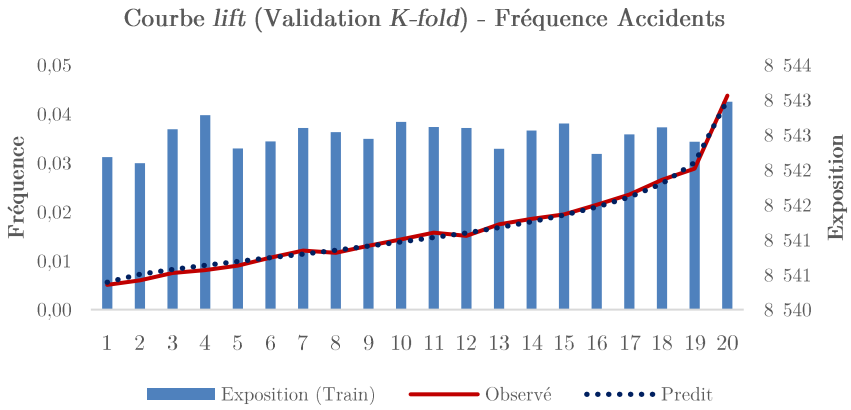
	Validation K-Fold	Validation Globale	Var. (%)
Coefficient de Gini Norm.	29,47%	29,37%	**0,34%**
Pseudo-R2	3,02%	3,02%	**0,00%**
RMSE	0,2047	0,205	**-0,15%**

Tableau 5.5 - Critères de performance du modèle de la fréquence des accidents



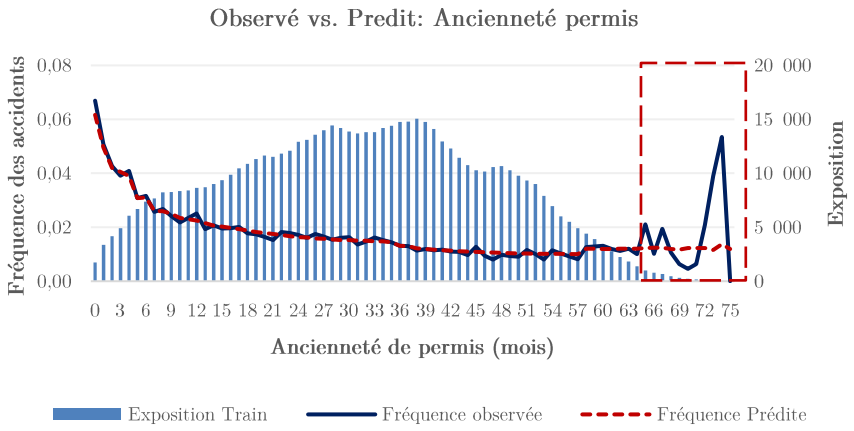
Graphique 5.5 - Courbe de Lorenz du modèle de fréquence des accidents

Le coefficient de Gini normalisé et la courbe de Lorenz mettent en avant un niveau favorable de segmentation du portefeuille. En effet, de manière similaire aux pannes, les 20% cumulés du risque prédit le plus élevé représentent environ 38% du risque observé. Le Pseudo-R^2 se situe lui autour de 3%, une valeur inférieure à celle du modèle de fréquence des pannes qui est également faible (~6%). Bien qu'il semble que ces modèles fournissent une très faible réduction de la déviance, il est normal d'observer ce niveau de Pseudo-R^2 dans des ensembles de données bruyantes comme celles de cette étude.

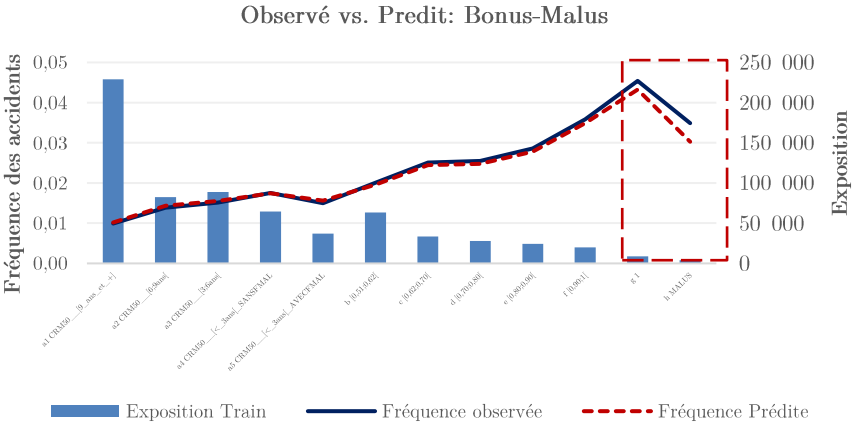


Graphique 5.6 - Courbe lift de la modélisation de la fréquence des pannes

En ce qui concerne la RMSE, la faible différence observée entre la valeur obtenue avec la validation *k-fold* et celle obtenue avec la validation globale indique que le modèle s'adapte correctement aux données. On observe également que la courbe lift des valeurs prédites est plus lisse que celle des valeurs observées. Cela montre qu'il n'y a pas de problèmes de sur-apprentissage dans le modèle. Néanmoins, lorsque l'on regarde les prédictions par modalité de certaines variables retenues dans la modélisation, on observe que la fréquence des modalités considérées comme « mauvais risque » est sous-estimée. Les figures 5.7 et 5.8 montrent cette limite pour les variables ancienneté de permis et bonus-malus.



Graphique 5.7 - Fréquence observée vs fréquence prédite variable ancienneté de permis



Graphique 5.8 - Fréquence observée vs fréquence prédite variable bonus-malus

Comme évoqué dans la section 5.3.2, le tests statistiques appliqués pour regrouper les coefficients ne donnent pas toujours de résultats fiables. Etant donné que l'exposition des modalités considérées comme des mauvais risques est souvent assez faible, leur effet est considéré comme non significatif et donc regroupé. Ceci explique la sous-estimation de la fréquence de ces modalités. Un ajustement manuel des coefficients pour ces modalités pourrait être une solution potentielle à cette limite.

Les interactions potentielles ont également été analysées (cf. Annexe C4.4), mais leur incorporation n'améliore pas significativement les performances du modèle (cf. tableau 5.6). Le modèle sans interactions a donc été retenu pour estimer la fréquence des accidents.

	Gini Normalisé		
	Validation K-Fold	Validation Globale	Var. (%)
Modèle 14 variables	29,47%	29,37%	**0,34%**
Modèle 14 variables + interaction	29,43%	30,41%	**-3,22%**
Var. (%)	**0,14%**	**-3,42%**	

	RMSE		
	Validation K-Fold	Validation Globale	Var. (%)
Modèle 14 variables	0,2047	0,205	**-0,15%**
Modèle 14 variables + interaction	0,205	0,204	**0,49%**
Var. (%)	**-0,15%**	**0,49%**	

Tableau 5.6 - Comparaison du modèle à 14 variables et du modèle à 14 variables + interactions (accidents)

Le tableau 5.7 présente les variables discriminantes retenues dans la modélisation. Il y a une légère différence entre les variables retenues pour l'estimation de la fréquence des pannes et celle pour l'estimation de la fréquence des accidents. Par exemple, l'âge du véhicule n'est pas retenu dans le modèle d'accident alors qu'il l'est dans le modèle de panne. Au contraire, la profession du conducteur principal est retenue dans le modèle d'accident alors qu'elle ne l'est pas dans le modèle de panne. Ceci est conforme à la réalité, car on sait *a priori* que les accidents sont principalement liés au profil du conducteur et que les pannes sont principalement liées aux caractéristiques du véhicule. En France, par exemple, les premières causes d'accidents sont la vitesse excessive ou inadaptée et la consommation illégale d'alcool (OISR, 2022), alors que dans le cas des pannes, les problèmes de batterie et de moteur ressortent comme les principales causes (Lauriol, 2022). Il est probable que l'ajout d'autres variables liées au profil de l'assuré, comme l'âge de l'assuré ou l'historique des sinistres passés, améliorerait la performance du modèle de fréquence des accidents.

Le fait que très peu de variables liées au profil de l'assuré soient disponibles dans la base de données représente une limite dans la modélisation des accidents dans cette étude mais plus généralement dans la modélisation des accidents dans un portefeuille BtoB pour un constructeur automobile. En effet, les prix étant négociés avant même la commercialisation des polices, les variables tarifaires liées au profil de l'assuré ne sont pas disponibles lors de la souscription d'un contrat d'assurance collectif comme celui-ci. Dans un deuxième temps, qui ne sera pas abordé dans cette étude, il faudrait construire un nouveau modèle sans considérer les variables liées au profil de l'assuré et analyser ses performances par rapport au modèle complet.

No. Variable	Description	Type
1	Ancienneté du permis	Assuré
2	Bonus-Malus	Assuré
3	Offre 8000 km (le conducteur s'engage à faire moins de 8000km par an)	Contrat
4	Hauteur du véhicule	Véhicule
5	Profession du conducteur principal	Assuré
6	Nombre des places	Véhicule
7	Longueur (mm)	Véhicule
8	Type de garage pour le véhicule	Assuré
9	Zonier	Géographique
10	Formule Assistance	Contrat
11	Régime Puissance Maximum	Véhicule
12	Vitesse Maximale	Véhicule
13	Année de survenance	Temporelle
14	Etoiles de sécurité passive	Véhicule

Tableau 5.7 - Variables retenues pour la modélisation de la fréquence des accidents

5.5.3. Modélisation du coût moyen des pannes et des accidents

La même approche de modélisation appliquée pour la fréquence a également été appliquée pour le coût moyen des pannes et des accidents. Cependant, le nombre de variables explicatives incluses dans la modélisation a été réduit car, contrairement à la modélisation de la fréquence où on a essayé d'identifier les variables les plus discriminantes, dans le cas du coût moyen, on connait en amont de la modélisation les principaux effets l'impactant. Ceci est dû au fait que la société d'assistance s'engage à garantir des services d'assistance routière (les *prestations de base* dans le cas de cette étude) et non une compensation financière (coût des réparations et des pièces détachées par exemple). Par conséquent, il est plus facile de déterminer amont les variables impactant la sévérité de sinistres.

Les principaux effets impactant le coût moyen de la garantie d'assistance routière sont énumérés ci-dessous :

- Type de prestation fournie : Le coût du dépannage sur place est souvent inférieur à celui du remorquage, et de plus, ces deux services ne sont pas exclusifs. Ceci veut dire que ce n'est pas parce qu'il y a un dépannage sur place qu'il n'y aura pas forcément un remorquage, et vice versa. Une des limites de cette étude est que ces deux services ne sont pas différenciés dans la base de données de sinistres. Le coût des accidents sera automatiquement plus cher car la part du remorquage sera plus importante que celle observée pour les pannes. En effet, il est moins probable dans le cas d'un accident que le véhicule soit réparé sur place. Idéalement, les deux services devraient être modélisés séparément.

- Carrosserie du véhicule : Le coût moyen des services de base dépend de la carrosserie du véhicule ou plus précisément de sa taille et de son poids. Plus le véhicule est grand, plus il faudra utiliser de grues et de dispositifs spécifiques lors de la prestation de service d'assistance.

- Type d'énergie : Les véhicules rechargeables (hybrides et électriques) ont besoin de services d'assistance spécifiques notamment liés au rechargement de batteries sur place, ce qui peut avoir un impact sur le coût.

- Inflation : Comme mentionné dans la section 2.2.4, le coût est fortement impacté par l'inflation annuelle, notamment dans le contexte inflationniste actuel où l'on observe une forte tendance à la hausse des prix du carburant et des coût du transport en général. Cet effet est partiellement absorbé par l'intégration de l'année d'occurrence dans la modélisation.

- Zone géographique : Lors de la création du zonier au Chapitre 4, il a été démontré que la zone géographique a un impact sur le coût moyen des sinistres. En particulier, il existe une différence dans le coût des sinistres entre les zones urbaines et les zones rurales. Par exemple, dans ces dernières, le réseau de partenaires est moins granulaire et les distances parcourues par les dépanneuses sont donc plus longues, ce qui a un impact à la hausse sur le coût moyen.

- Négociations des prix avec le réseau de partenaires : Cet aspect est celui qui a le plus d'impact sur le coût des services de base. En fonction des volumes de prestations ramenés au réseau de dépannage et des relations commerciales entre la société d'assistance et ses partenaires, des négociations sont menées périodiquement, souvent annuellement mais pas exclusivement. Même si ce point est partiellement pris en compte par la variable année d'occurrence qui considère l'évolution des coûts d'une année sur l'autre, cet effet est difficile à intégrer dans la modélisation car les conditions de négociation de chaque partenaire sont indépendantes les unes des autres et une société d'assistance peut avoir des centaines de partenaires.

Dans cette optique, les variables intégrées dans la modélisation sont présentées dans le tableau 5.8. Les autres variables ont été retirées de la base de données afin de ne pas ajouter de bruit à la modélisation.

No. Variable	Description	Type
1	Poids à vide (en Kg)	Véhicule
2	Longueur (mm)	Véhicule
3	Hauteur (mm)	Véhicule
4	Zonier	Géographique
5	Type d'Energie	Véhicule
6	Année de survenance	Temporelle
7	Type de Carrosserie	Véhicule
8	Usage du véhicule	Véhicule
9	Marque	Véhicule
10	Formule d'assistance	Contrat

Tableau 5.8 - Variables intégrées à la modélisation du coût moyen des pannes et des accidents

Même si les fréquences des pannes et des accidents sont impactées par des effets différents, les pannes étant plus liées aux caractéristiques du véhicule et les accidents plus liés au profil de l'assuré, le coût moyen des prestations de base dépend très peu du fait générateur. Par exemple, dans le cas du remorquage, le coût sera le même quel que soit l'état du véhicule ou le motif du service. Cependant, le fait de traiter ensemble le service de remorquage et le service de dépannage sur place conduit à une fausse corrélation du coût moyen avec le fait générateur car, comme on l'a déjà mentionné, la part du remorquage dans les accidents est plus élevée que dans les dépannages sur place. Par conséquent, dans cette étude, on créera deux modèles de sévérité différents pour chaque fait générateur, mais dans la pratique, cela n'est pas nécessaire.

Les figures 5.3 et 5.4 présentent le spectre de modèles de sévérité crées par Akur8 respectivement pour les pannes et accidents. Les variables retenues par les modèles sélectionnés sont assez similaires (cf. tableau 5.9), en revanche, leurs performances sont différentes. On observe que les barres d'erreur et les moyennes du coefficient de Gini normalisé pour chaque groupe de parcimonie sont plus volatiles pour les accidents.

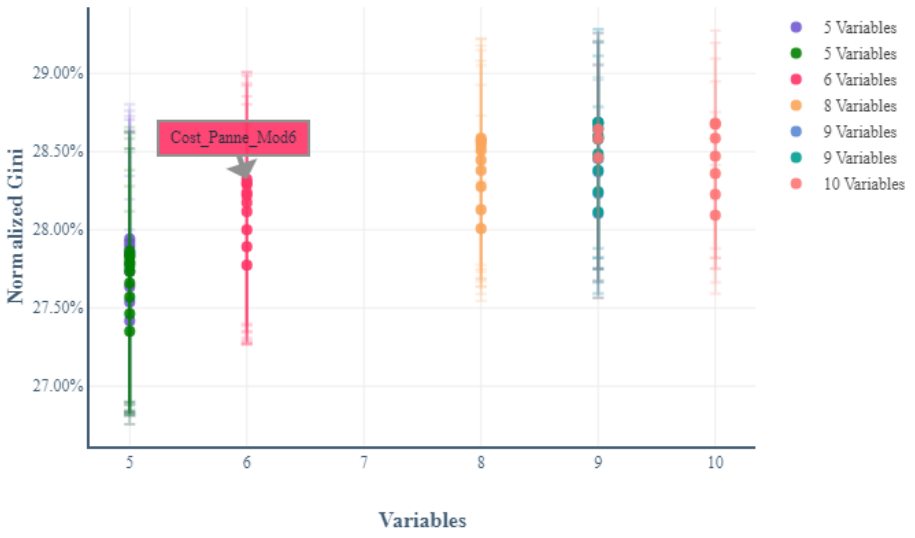


Figure 5.3 - Structure de modélisation de la sévérité des pannes (Source : Akur8)

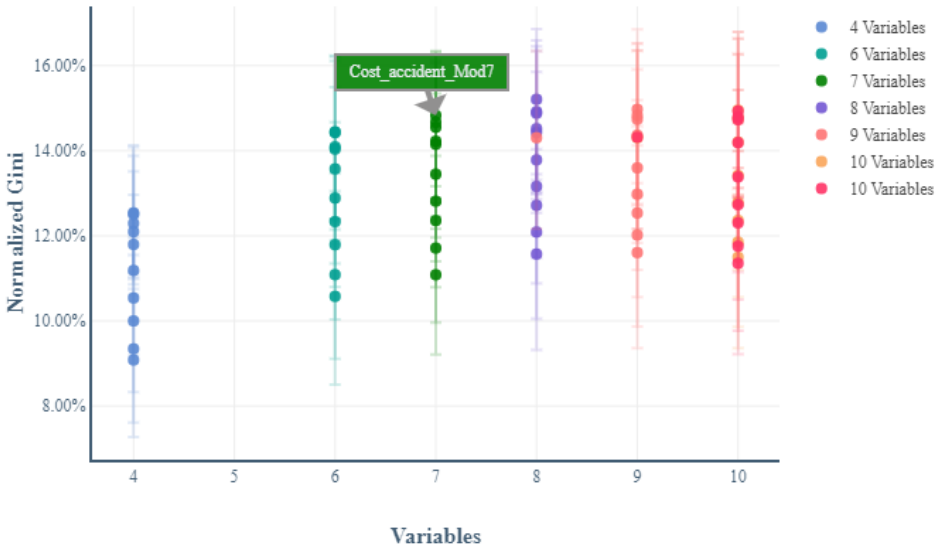


Figure 5.4 - Structure de modélisation de la sévérité des pannes (Source : Akur8)

No. Variable	Description	Pannes	Accidents
1	Poids à vide (en Kg)	X	X
2	Longueur (mm)	X	X
3	Hauteur (mm)	X	X
4	Zonier	X	X
5	Type d'Energie	X	X
6	Année de survenance	X	X
7	Type de Carrosserie		
8	Usage du véhicule		
9	Marque		
10	Formule d'assistance		X

Tableau 5.9 - Variables explicatives retenues pour la modélisation des pannes et des accidents

Dans les deux cas, la courbe de Lorenz est proche de la ligne d'égalité et le coefficient de Gini associé est très faible (cf. figure 5.6 et tableau 5.10), soulignant un faible pouvoir de segmentation du risque de sévérité du modèle. Toutefois, il n'est pas possible de tirer des conclusions sur la qualité du modèle à partir de cette courbe et du coefficient de Gini associé. En effet, contrairement à l'assurance automobile, la sévérité des prestations d'assistance routière présente une faible variance : les coûts de remorquage sont relativement similaires quel que soit le profil de risque. On note d'ailleurs que le coefficient de Gini normalisé est nettement supérieur au coefficient de Gini. Ceci montre que dans le cadre d'une prédiction parfaite, le coefficient de Gini obtenu serait faible.

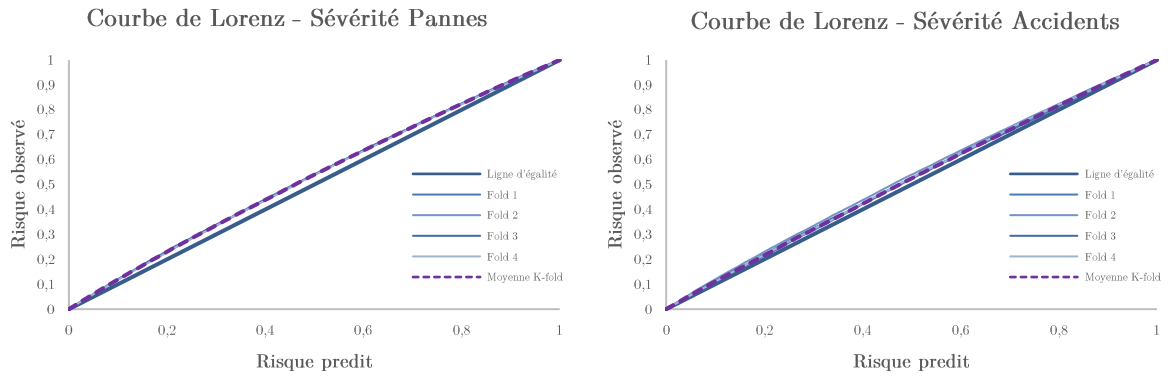


Figure 5.5 - Courbe de Lorenz du modèle de la sévérité des pannes et des accidents

	Pannes		
	Validation K-Fold	Validation Globale	Var. (%)
Coefficient de Gini	5,49%	5,24%	**4,77%**
Coefficient de Gini Norm.	28,32%	27,03%	**4,77%**

	Accidents		
	Validation K-Fold	Validation Globale	Var. (%)
Coefficient de Gini	3,41%	3,64%	**-6,32%**
Coefficient de Gini Norm.	14,83%	15,83%	**-6,32%**

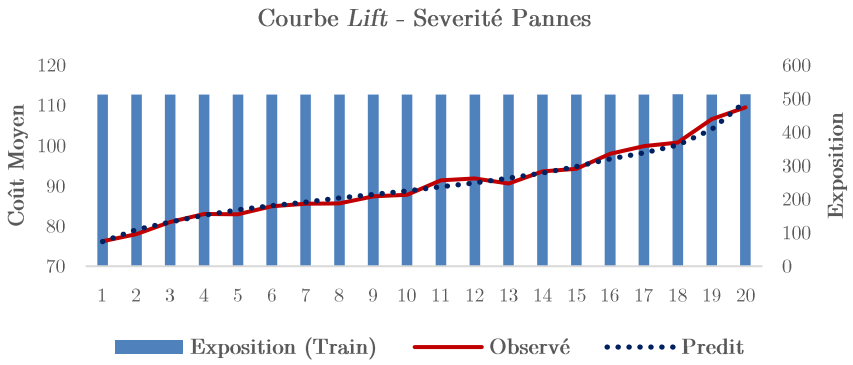
Tableau 5.10 - Coefficient de Gini du modèle de sévérité des pannes et des accidents

En ce qui concerne la qualité prédictive des modèles de sévérité, les constats suivantes peuvent être tirés[18] :

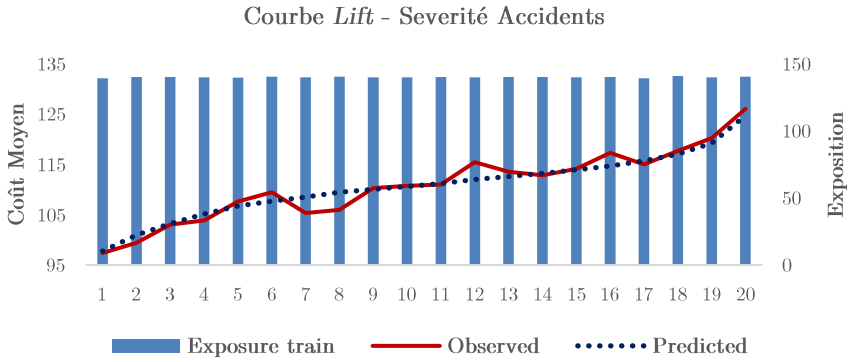
- Les figures 5.9 et 5.10 font apparaître une courbe des valeurs observées moins lisse que celle observée dans le cas de la fréquence et donc des prédictions moins fiables, plus spécifiquement dans les cas d'accidents.
- L'analyse RMSE révèle une cohérence entre la validation K-Fold et la validation globale, soulignant que les modèles s'adaptent de la même manière aux données d'entraînement (Train) et aux données de validation (Test) et évitent ainsi le surajustement (cf. Tableau 5.11).
- Même si le pseudo-R2 obtenu dans la modélisation de la sévérité des pannes lors de la validation K-Fold est inférieur d'environ 7% à celui de la validation globale, on observe une légère amélioration par rapport au modèle de fréquence de l'ordre de 6%. Ce n'est pas le cas pour le modèle de sévérité des accidents où l'on observe une valeur très faible du Pseudo-R2 autour de 2% et une variation de 20% entre la validation croisée et la validation globale du modèle (cf. Tableau 5.11).
- La faible stabilité et fiabilité de la modélisation des accidents s'explique en partie par la faible fréquence de cet événement dans le portefeuille étudié. La loi Gamma n'acceptant que des valeurs positives, la modélisation n'est effectuée que sur le jeu de données contenant des polices ayant eu au moins un accident et donc un coût de sinistre non nul. Le portefeuille étudié n'est plus composé que de 3000 polices ayant eu un accident durant la période d'étude.

[18] Le détail des critères de performance de la modélisation de la sévérité se trouve dans les annexes C.4.5 et C.4.6

- L'intégration des interactions potentielles n'a pas amélioré de manière significative la qualité des modèles.



Graphique 5.9 - Courbe lift du modèle de sévérité des pannes



Graphique 5.10 - Courbe lift du modèle de sévérité des accidents

	Pannes				Accidents		
	Validation K-Fold	Validation Globale	Var. (%)		Validation K-Fold	Validation Globale	Var. (%)
RMSE	32,25	32,43	**-0,56%**	RMSE	45,87	45,94	**-0,15%**
Pseudo-R^2	7,86%	7,34%	**7,08%**	Pseudo-R^2	2,15%	2,70%	**-20,37%**

Tableau 5.11 - Critères de performance des modèles de sévérité des pannes et des accidents

Les résultats de la modélisation de la sévérité et de la fréquence analysés et commentés précédemment permettent de conclure que la modélisation des pannes est plus stable que celle obtenue pour les accidents. Par conséquent, le calcul de la prime pure ne sera effectué que sur ce fait générateur afin d'obtenir une estimation plus fiable.

5.6. Application du modèle pour le calcul de la prime pure

L'analyse réalisée dans les sections précédentes a fait ressortir les variables impactant la fréquence et la sévérité des prestations de base au sein d'une garantie d'assistance routière. Ainsi, en utilisant ces résultats, il est possible de calculer une nouvelle prime de risque pour le portefeuille étudié. Il est important de se rappeler ici que l'objectif de la tarification d'un contrat en BtoB pour un constructeur automobile est de déterminer une prime de risque moyenne qui sera appliquée à l'ensemble des bénéficiaires. Par conséquent, il n'est pas possible d'appliquer une prime segmentée suivant le profil de risque du client final.

La méthode de tarification actuelle de ce type de contrat repose sur le cadre usuel du modèle simple fréquence x coût moyen. L'estimation de la prime pure par le biais du modèle simple est décrite par l'équation ci-dessous :

Équation 5.9: Estimation de la prime pure par une modélisation simple *(Sauveplane, 2020)*

- On note j un sinistre d'un assuré i tel que $0 \leq j \leq k_i$
- k_i correspond au nombre de sinistres de l'assuré i. Il s'agit d'une variable aléatoire prenant des valeurs positives où nulle lorsqu'aucun sinistre n'est survenu.
- Y_{ij} désigne le coût du j*ème* sinistre de l'assuré i
- La charge annuelle imputable à un assuré est définie de la manière suivante :

$$X_i = Y_{i1} + Y_{i2} + \ldots + Y_{ik}$$

- Le nombre total de sinistres d'un portefeuille est défini de la manière suivante :

$$N = k_1 + k_2 + \ldots + k_n$$

Où, n nombre total d'assurés du portefeuille

Sous l'hypothèse que les Y_{ij} sont *iid* tels que $E(Y_{ij}) = E(Y)$ et que les Y_{ij} et N sont indépendantes, l'espérance du coût de total de sinistres peut être défini comme :

$$E(X) = E(Y) \times E(N)$$

$$E(Y) = \frac{\sum_i^n \sum_j^k Y_{ij}}{E(N)}$$

$$E(N) = \frac{\sum_i^n k_i}{n}$$

Où $E(Y)$ et $E(N)$ sont calculés sur la base de données historiques pour un portefeuille donné sur une période prédéfinie.

Le prime pure est donc définie comme le produit de la fréquence et le coût moyen :

$$PP = \frac{E(X)}{n} = E(Y) \times \frac{E(N)}{n}$$

Le problème du modèle simple l'absence de prise en compte des facteurs de risque impactant la prime pure.

Afin de considérer les facteurs de risque impactant le risque de fréquence et de sévérité, la nouvelle prime pure calculée tiendra compte de l'estimation segmentée de fréquence et de coût moyen défini à l'aide des modèles précédemment développés. Ces modèles permettent de calculer une fréquence et un coût moyen pour chaque assuré selon son profil de risque. Comme l'objectif est de tarifier un contrat de type BtoB pour un constructeur automobile, des nouveaux sous-ensembles de données seront créés chacun correspondant à une marque de véhicule au sein du portefeuille d'étude. La fréquence et le coût moyen de chaque sous-ensemble représentant un portefeuille fictif seront calculés comme la moyenne des valeurs estimées par le biais des modèles GLM au prorata de l'exposition de chaque police (cf. Equation 5.10). Enfin, la nouvelle prime sera calculée par la méthode usuelle de tarification fréquence x coût moyen.

Équation 5.10: Estimation de la prime pure considérant les profils de risque

$$PP_m = \frac{1}{W_m} \times \sum_i \omega_{im} \times E(Y_{im})_{GLM} \times \frac{E(N_{im})_{GLM}}{\omega_i}$$

$$PP_m = \frac{1}{W_m} \times \sum_i E(Y_{im})_{GLM} \times E(N_{im})_{GLM}$$

Où,

$$PP_m = Prime\ pure\ pour\ la\ marque/constructeur\ automobile\ m$$

$$E(Y_{im})_{GLM} = Coût\ moyen\ estimé\ par\ le\ GLM\ pour\ l'assuré\ i\ \ ayant\ un\ vehicule\ de\ marque\ m$$

$$E(N_{im})_{GLM} = Nombre\ de\ sinistres\ estimé\ par\ le\ GLM\ pour\ l'assuré\ i\ ayant\ un\ vehicule\ de\ marque\ m$$

$$\omega_{im} = Exposition\ de\ l'assuré\ i\ ayant\ un\ vehicule\ de\ marque\ m$$

$$W_m = Exposition\ total\ du\ sous-ensemble\ correspondant\ à\ la\ marque\ m$$

Les tableaux 5.12, 5.13 et 5.14 ci-dessous permettent de comparer la fréquence, le coût moyen et la prime pure appliquée au portefeuille sur l'année 2019 dans le cadre de la modélisation simple et des GLMs développés dans cette étude. La variation de l'estimation fréquence par les deux modèles est plus importante que celle du coût moyen. Par exemple, la fréquence de la marque n°2 estimée avec le GLM est environ 17% plus élevée que celle estimée par le modèle simple, ce qui signifie que la méthode de tarification actuelle sous-estimerait le risque de fréquence pour cette marque. En

revanche, on observe que pour la marque n° 11, la fréquence estimée par le GLM est inférieure d'environ 15% à celle du modèle simple. Ceci représente une surestimation du risque de fréquence par la méthode actuelle. Une évaluation de la fréquence au moyen des GLMs développés dans cette étude permettrait de mieux segmenter la fréquence et d'avoir une marge de négociation avec un gain potentiel en compétitivité en termes de prix.

En ce qui concerne le coût moyen, des variations plus faibles sont observées entre les estimations des deux modèles. Comme mentionné précédemment, le coût moyen est davantage impacté par des effets conjecturaux (inflation, conditions du réseaux partenaires d'assistance, productivité des plateformes des opérateurs, etc.) que par des effets liés au véhicule ou au profil de l'assuré en tant que tel.

Marque	Modèle Simple	GLM	Var. (%)
M1	3,83%	4,37%	14,03%
M2	4,27%	4,98%	16,70%
M3	5,02%	5,25%	4,59%
M4	7,14%	6,91%	-3,19%
M5	6,42%	6,06%	-5,56%
M6	5,59%	5,56%	-0,56%
M7	5,86%	6,02%	2,65%
M8	6,37%	6,11%	-4,01%
M9	4,78%	5,16%	7,88%
M10	5,36%	5,29%	-1,45%
M11	6,23%	5,30%	-14,93%
M12	5,58%	5,32%	-4,60%
M13	5,72%	5,51%	-3,62%
M14	8,75%	8,65%	-1,17%

Tableau 5.12 - Comparaison d'estimation des fréquences entre le modèle simple et le GLM

Marque	Modèle Simple	GLM	Var. (%)
M1	89,03 €	89,41 €	0,42%
M2	95,57 €	92,56 €	-3,14%
M3	84,80 €	86,27 €	1,73%
M4	85,89 €	84,32 €	-1,83%
M5	89,52 €	88,85 €	-0,75%
M6	89,22 €	92,52 €	3,70%
M7	90,17 €	88,01 €	-2,40%
M8	87,94 €	86,28 €	-1,89%
M9	86,24 €	86,52 €	0,33%
M10	85,15 €	86,78 €	1,92%
M11	84,48 €	85,03 €	0,66%
M12	83,24 €	84,62 €	1,66%
M13	88,06 €	86,94 €	-1,28%
M14	91,85 €	93,21 €	1,49%

Tableau 5.13 - Comparaison d'estimation du coût moyen entre le modèle simple et le GLM

Marque	Modèle Simple	GLM	Var. (%)
M1	3,41 €	3,91 €	14,51%
M2	4,08 €	4,61 €	13,03%
M3	4,25 €	4,53 €	6,40%
M4	6,13 €	5,83 €	-4,96%
M5	5,74 €	5,38 €	-6,26%
M6	4,99 €	5,14 €	3,12%
M7	5,29 €	5,30 €	0,19%
M8	5,60 €	5,27 €	-5,83%
M9	4,12 €	4,46 €	8,24%
M10	4,57 €	4,59 €	0,45%
M11	5,26 €	4,50 €	-14,37%
M12	4,65 €	4,51 €	-3,02%
M13	5,04 €	4,79 €	-4,85%
M14	8,04 €	8,06 €	0,31%

Tableau 5.14 - Comparaison d'estimation de prime pure entre le modèle simple et le GLM

A noter que le cadre de l'étude s'intéresse à un portefeuille constitué de plusieurs constructeurs automobiles. Il conviendrait d'adapter le modèle en cas d'application sur un portefeuille constitué d'une seul constructeur :

- Si le constructeur à étudier est déjà présent dans le portefeuille actuel de la compagnie d'assistance, il est nécessaire de répliquer le modèle à la base de données correspondante afin de réévaluer les coefficients du modèle GLM et ajuster l'estimation de fréquence et coût moyen.

- S'il s'agit d'un nouveau portefeuille (par exemple lors d'un appel d'offres), en l'absence de base de données en ligne à ligne, il est nécessaire de faire une étude de marché afin d'évaluer les caractéristiques du portefeuille clients. Les coefficients du modèle du portefeuille présentant le plus de similitudes peut être alors appliqué.

De même, cette étude s'est intéressée uniquement à la méthode de calcul de la prime pure liées aux prestations de base dans le cadre d'une garantie d'assistance automobile. Pour arriver à calculer une prime commerciale il faudrait considérer les points suivants :

- Mettre à jour la base des données avec des données plus récentes, notamment pour considérer les tendance les plus récentes du secteur automobile (e.g. véhicules).

- Considérer l'impact de la forte tendance à la hausse de l'inflation observée dans la dernière année.

- Intégrer le coût opérationnel liés aux plateformes de gestion de sinistres

- Appliquer les chargements nécessaires pour couvrir le frais de structure et rémunérer la compagnie d'assistance.

En conclusion, ce chapitre a permis de mettre en avant les points suivants :

- Au vu du grand nombre de modèles testés avec l'outil Akur8, il est primordial de définir les critères de performance pour la sélection des modèles retenus. Les principaux critères utilisés dans cette étude sont la courbe de Lorenz et le coefficient de Gini associé, le RMSE et la courbe *Lift*.

- L'application des GLMs retenus a permis de mettre en avant les variables discriminantes pour la modélisation de la fréquence et du coût moyen des pannes et accidents.

- Les résultats de la modélisation de la sévérité ont montré que la modélisation des pannes est plus stable que celle obtenue pour les accidents. Par conséquent, le calcul de la prime pure a été effectué que sur ce fait générateur afin d'obtenir une estimation plus fiable.

- Le développement de GLMs à l'aide de l'outil Akur8 a ensuite permis de calculer une prime pure des pannes par marque de véhicule qui considère les facteurs de risque du portefeuille, contrairement à ce qui est réalisé habituellement en BtoB. Des différences notables ont été observées notamment dans l'estimation de la fréquence entre les deux méthodes, démontrant ainsi l'utilité de la mise en place du modèle proposé.

Conclusion

Cette étude a porté sur le développement d'un modèle de tarification de la garantie d'assistance routière pour un portefeuille BtoB en utilisant des modèles linéaires généralisés (GLMs). L'objectif a été de prendre en compte les facteurs de risque impactant la prime pure des deux faits générateurs : les pannes et les accidents. Pour ce faire, une base de données riche et hétérogène en termes de variables et de contrats (21 millions d'individus et 82 variables sur la période 2017-2019) a été mise à disposition par un assureur partenaire. Néanmoins, les résultats de l'étude ont vocation à être utilisés sur les portefeuilles des constructeurs automobiles, un segment de clientèle très concurrentiel et très présent dans le portefeuille des sociétés d'assistance en France.

La caractérisation de la base de données par des analyses exploratoires a démontré que les variables liées au véhicule ont un fort impact sur le risque de pannes et celles liées à l'assuré ont un fort impact sur le risque d'accidents. Des retraitements sur la volumétrie, les valeurs manquantes et les valeurs aberrantes ont été effectués avant les analyses exploratoires afin de rendre les données plus faciles à gérer et plus précises. De même, l'étape de présélection des variables entreprise à l'aide des méthodes factorielles et de la matrice de corrélation Phi-K a permis d'éliminer les informations redondantes et de rendre les données exploitables en termes de temps et de complexité de calcul. Ainsi, la taille de la base de données de l'étude a été réduite à 36 variables et 1,1 million de lignes.

La conduction d'une classification ascendante hiérarchique (CAH) en amont de la modélisation a permis d'identifier les disparités géographiques existantes au niveau de la fréquence et le coût moyen du portefeuille étudié par la création de la variable de type *zonier*. Enfin, différents GLMs ont été calibrés à l'aide de l'outil Akur8 permettant d'estimer au mieux la fréquence et le coût moyen des deux faits générateurs (pannes et accidents). Les modèles retenus ont mis en avant les variables impactant le risque de panne et d'accident. Néanmoins, les résultats de la modélisation de la sévérité ont montré que le GLM retenu pour l'estimation des pannes est plus stable que celui obtenu pour les accidents. Par conséquent, le calcul de la prime pure n'a été effectué que sur les pannes afin d'obtenir une estimation plus fiable.

La prime pure calculée pour les pannes a été comparée à celle estimée via un modèle de tarification simple. Des différences significatives ont été observées, notamment dans l'estimation de la fréquence entre les deux méthodes. Ces différences se traduisent soit par une sous-estimation du risque par la méthode de tarification simple actuellement utilisée pour certaines marques, ce qui signifierait un S/P supérieur à 100% et donc un portefeuille non rentable ; soit par une surestimation du risque par la méthode actuelle, ce qui signifie qu'une évaluation de la fréquence au moyen des GLMs développés dans cette étude permettrait une marge de négociation

et, un gain potentiel de compétitivité en termes de prix. Ceci démontre l'utilité de la mise en œuvre du modèle proposé.

Limites et ouvertures de l'étude

Il convient également de rappeler limites principales de l'étude et les ouvertures potentielles :

- **Les services de remorquage et dépannage ne sont pas distingués dans la base des données**. L'assureur partenaire enregistre le dépannage sur place et le remorquage, mais sans distinction du type de sinistre. Toutefois, la fréquence et la sévérité des deux prestations peut être significativement différente. La fréquence du remorquage sera plus importante dans le cas des accidents puisqu'il est moins probable que le véhicule soit réparé sur place en cas d'accident. Le coût du dépannage sur place est souvent inférieur à celui du remorquage, et de plus, ces deux services ne sont pas exclusifs. Ceci veut dire que ce n'est pas parce qu'il y a un dépannage sur place qu'il n'y aura pas forcément un remorquage, et vice versa. Idéalement, les deux services devraient être modélisés séparément.

- **L'indisponibilité des variables liées à l'assuré qui est une particularité du marché de l'assistance routière en BtoB**. Les prix étant négociés avant même la commercialisation des polices, les variables tarifaires liées au profil de l'assuré ne sont pas disponibles lors de la souscription d'un contrat d'assurance en BtoB. Dans un deuxième temps, qui ne sera pas abordé dans cette étude, il faudrait construire un nouveau modèle sans considérer les variables liées au profil de l'assuré et analyser ses performances par rapport au modèle complet.

- **Le type de carburant/énergie du véhicule ne ressort pas comme une variable discriminante dans la modélisation de la fréquence.** Ceci est une limite du modèle qui peut s'expliquer par la faible exposition des véhicules non thermiques dans le portefeuille étudié, qui s'étend sur la période 2017 - 2019. Au vu des incitations de l'état pour augmenter la part de véhicules rechargeables dans le parc automobile français afin de réduire les émissions de CO_2, la proportion de ces véhicules devrait augmenter progressivement. Il serait donc intéressant, en complément de cette étude, de réaliser une modélisation similaire sur un portefeuille comprenant une part plus importante de véhicules rechargeables, ou sur le même portefeuille sur une période plus récente (2020-2022 par exemple). Cela démontre également l'importance de retester la pertinence des modèles d'une année sur l'autre sur un segment en évolution comme l'automobile.

- **L'étude ne s'applique qu'au marché français.** Comme mentionné au chapitre 1, le portefeuille BtoB des constructeurs automobiles du groupe Europ Assurance est constitué de contrats internationaux, c'est-à-dire de contrats présents sur plusieurs pays. Il serait pertinent d'étendre cette analyse à d'autres marchés européens.

Bibliographie

Luc Charbonnier et C-Ways. (2020). *La Voiture: Le divorce impossible?* Observatoir Cetelem. Patricia Bosc. Récupéré sur https://observatoirecetelem.com/lobservatoire-cetelem-de-lautomobile/voiture-divorce-impossible/partenaires-et-methodologie

Abdi, H., & Williams, L. (2010). Principal component analysis. *John Wiley & Sons, Inc.*, 433-59. Récupéré sur http://staff.ustc.edu.cn/~zwp/teach/MVA/abdi-awPCA2010.pdf

ACEA. (2021). *Véhicules électriques : Avantages fiscaux et incitations à l'achat dans l'Union européenne (2021).* Bruxelles: Association des constructeurs européens d'automobiles.

ACEA. (2022). *Average age of the EU vehicle fleet, by country.* Bruxelles: European Automobile Manufacturers' Association.

ACEA. (2022). *Types de carburants des voitures neuves : part de marché des voitures électriques à batterie 10,0 %, des voitures hybrides 25,1 % et des voitures à essence 36,0 % au premier trimestre 2022.* Bruxelles: Association des constructeurs européens d'automobiles.

Atlas Magazine. (2012, Novembre 21). *Qu'est-ce qu'une assurance assistance ?* Récupéré sur Site web Atlas Magazine. L'actualité de l'assurance dans le monde: https://www.atlas-mag.net/article/l-assurance-assistance

Baak, M., Koopman, R., Snoek, H., & Klous, S. (2019). *A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics.*

Brian Collie, A. W. (2020). *COVID-19's Impact on the Automotive Industry.* Boston Consulting Group (BCG). Récupéré sur https://www.bcg.com/publications/2020/covid-automotive-industry-forecasting-scenarios

Brownlee, J. (2020, Août 3). *Machine Learning Mastery.* Récupéré sur A Gentle Introduction to k-fold Cross-Validation: https://machinelearningmastery.com/k-fold-cross-validation/

BUCCI, S. (2021). *Étude et implémentation de techniques d'analyse de sensibilité dans les modèles de tarification Non-Vie. Application à la tarification à l'adresse.* PAris: Institut des Actuaires.

CORONE, S., & PERRIN, C. (2005, Mai 13). *L'assistance, des prestations étendues et des obligations souples.* Récupéré sur L'argus de l'assurance: https://www.argusdelassurance.com/reglementation/legislation/l-assistance-des-prestations-etendues-et-des-obligations-souples.31183

Europ Assistance. (2022, Mai 2). *Nous connaitre - Histoire.* Récupéré sur Site Web Europ Assistance: https://www.europ-assistance.com/fr/who-we-are-history

Finaccord. (2019). *Manufacturer-Branded and Dealer-Intermediated Motor Insurance and Road Assistance in Europe.*

Frees, E. (., Meyers, G., & Cummings, A. (2013). INSURANCE RATEMAKING AND A GINI INDEX. *The Journal of Risk and Insurance*, 335-366.

Gorrand, R. (2021). *Assurance dommage : Tarification a priori et a posteriori.* Paris: Cours CEA.

Husson, F. (2014). *Classification ascendante hiérarchique (CAH).* Rennes: Laboratoire de mathématiques appliquées - Agrocampus Rennes. Récupéré sur https://husson.github.io/MOOC_AnaDo/classif_cours_slides.pdf

Husson, F., Josse, J., & Pagès, J. (2009). Gestion des donnees manquantes en Analyse des Composantes Principale. *Journal de la Société Française de Statistique*, 28-51.

IBM. (2022, 01 11). *IBM.* Récupéré sur IBM Cognos Analytics 11.1: https://www.ibm.com/docs/en/cognos-analytics/11.1.0?topic=terms-cramrs-v

IHS, ACEA, Autovista Group simulation. (2020). *Used-car transactions Europe + UK 2019-2030, by age group, incl. UK, in million units.* Autovista24. Récupéré sur https://autovista24.autovistagroup.com/news/what-store-europes-car-dealers/

Kassambara, A. (2017, Octobre 15). *Méthodes des Composantes Principales dans R: Guide Pratique.* Récupéré sur Statistical Tools For High-Throughput Data Analysis (STHDA): http://www.sthda.com/french/articles/38-methodes-des-composantes-principales-dans-r-guide-pratique/73-acp-analyse-en-composantes-principales-avec-r-l-essentiel/#standardisation-des-donnees

Lauriol, B. (2022, avril 27). *Fiabilité. Analyse des pannes les plus courantes en 2021.* Récupéré sur L'argus: https://www.largus.fr/actualite-automobile/fiabilite-analyse-des-pannes-les-plus-courantes-en-2021-10925402.html

OISR. (2022). *Accidentalité routière 2021 (Données definitives).* France: Insee.

PAGES, J. (2010). *Statistique générale pour utilisateurs. Méthodologie* (éd. 2ème). Rennes: Presses Universitaires Rennes.

PWC. (2020). *Digital Auto Report 2020: Navigating through a post- pandemic world.* Strategy&. Récupéré sur https://www.strategyand.pwc.com/fr/fr/assets/pdf/2020/11/en-stratyand-digital-auto-report-2020.pdf

Sauveplane, P. (2020). *Assurance non-vie: Théorie du risque et réassurance.* Paris: Cours CEA 1ère année.

Sécurité et Réparation Automobiles. (2022). *Qui sommes nous ?* Consulté le 09 01, 2022, sur SRA: https://www.sra.asso.fr/Qui-sommes-nous

Syndicat National des Sociétes d'Assistance. (2021). *Au coeur de la crise sanitaire, l'assistance adapte son activité et mène des opérations qui feront date dans son histoire.* PAris: SNSA.

Technavio Research Firm. (2020). *Global Vehicle Roadside Assistance Market.* Toronto: Infiniti Research Limited.

Thomas Gersdorf, P. H. (2020). *McKinsey Electric Vehicle Index: Europe cushions a global plunge in EV sales.* McKinsey & Company. Récupéré sur https://www.mckinsey.com/industries/automotive-and-assembly/our-

insights/mckinsey-electric-vehicle-index-europe-cushions-a-global-plunge-in-ev-sales

Vernet, A., & Grafe, S. (2021). *La mutation du modèle des sociétés d'assistance à l'heure de la crise.* Paris: XERFI PERCEPTA.

Liste des abréviations, sigles et acronymes

- ACEA : Association des constructeurs européens d'automobiles
- ACM : Analyse des composantes multiple
- ACP : Analyse des composantes principales
- ADAS : Advanced Driver Assistance Systems
- AIC : Critère d'information d'Akaike
- APAC : Asie-Pacifique
- BEV : Véhicule Battery Electric
- BIC : Critère d'information bayésien
- CA : Chiffre d'affaires
- CAGR : Compound Annual Growth Rate
- CAH : Classification ascendante hiérarchique
- GLM : Modèle linaire généralisé
- HEV : Véhicule Hybrid Electric
- IID : Indépendant et identiquement distribué
- MOA : Moyen Orient et Afrique
- OEM : Original Equipment Manufacturer
- PHEV : Véhicule Plug-in Hybrid
- RMSE : Racine de l'erreur quadratique moyenne
- SARA : Service Activated Roadside Assistance
- SNSA : Syndicat National de Sociétés d'assistance
- UE : Union Européenne

Liste de figures

Liste de tableaux

Liste de graphiques

Liste d'Equations

Annexes

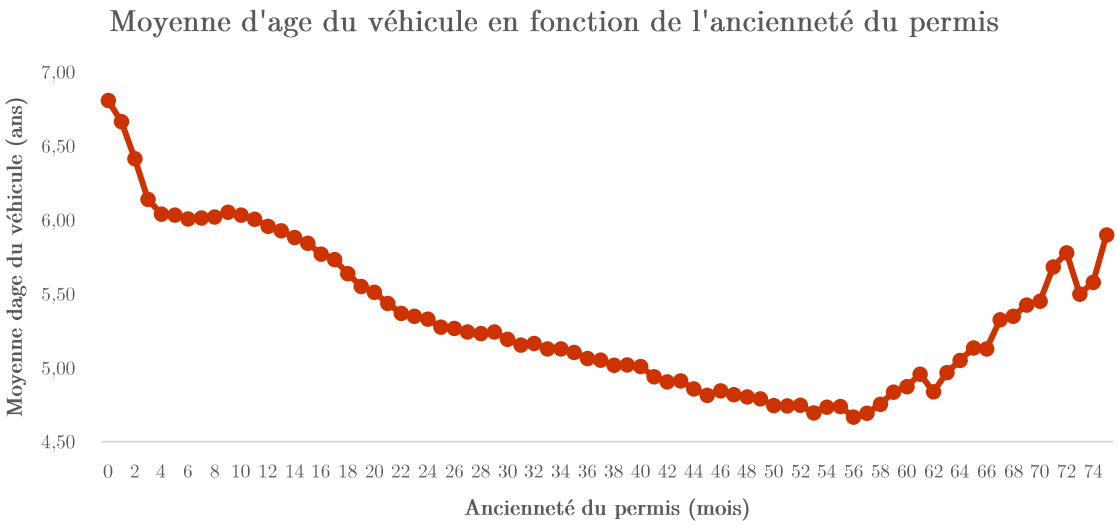
Annexe A : Présentation de la base de données

Annexe A.1 : Description des variables de la base de données d'étude

Nom de la variable	Description	Type	Modalités
ID_VEHICULIER	Identifiant SRA	ID	
RANNEE	Année de survenance	Factor	3
ID_BASE	Identifiant Assureur	ID	
expo_ANASB	Exposition du véhicule aux risques assurés par la formule socle, proratisée	Num	
expo_ANVDR	Exposition du véhicule aux risques assurés par les autres formules, proratisée	Num	
ID_GEO_COMMUNE	Code de la commune du lieu de stationnement du véhicule	Num	
Formule_Assistance	Formule de garanties Europ Assistance	Factor	4
Anciennete_permis	Ancienneté de permis du conducteur principal	Num	
Anciennete_permis_secondaire	Ancienneté de permis du conducteur secondaire le plus aggravant	Num	
PCSP2	Profession du conducteur principal	Factor	10
RUSAGE	Usage fait du véhicule	Factor	4
BonusMalus	Tranches de coefficient bonus-malus	Factor	12
RGRSRA	Groupe SRA du véhicule	Factor	31
Age_vehicule	Ancienneté de commercialisation du véhicule	Num	
Age_acquisition_vehicule	Ancienneté d'acquisition du véhicule	Num	
RCAROS	Carrosserie / Silhouette	Factor	14
RCARBUR	Alimentation / Carburant	Factor	8
Marque	Marque constructeur	Factor	15
RGARAGE	Type de garage pour le véhicule	Factor	6
RFORMULE	Formule de garanties Generali	Factor	3
R8000KM	Offre 8000 km (le conducteur s'engage à faire moins de 8000 ou 4000 km par an)	Factor	2
target_SNB_accident_remorquage	Nombre de sinistres accident X remorquage	Claims	
target_SNB_accident_taxitrain	Nombre de sinistres accident X taxitrain	Claims	
target_SNB_accident_voitureloc	Nombre de sinistres accident X voitureloc	Claims	
target_SNB_panne_remorquage	Nombre de sinistres panne X remorquage	Claims	
target_SNB_panne_taxitrain	Nombre de sinistres panne X taxitrain	Claims	
target_SNB_panne_voitureloc	Nombre de sinistres panne X voitureloc	Claims	
target_SMT_accident_remorquage	Montant de sinistres accident X remorquage	Claims	
target_SMT_accident_taxitrain	Montant de sinistres accident X taxitrain	Claims	
target_SMT_accident_voitureloc	Montant de sinistres accident X voitureloc	Claims	
target_SMT_panne_remorquage	Montant de sinistres panne X remorquage	Claims	
target_SMT_panne_taxitrain	Montant de sinistres panne X taxitrain	Claims	
target_SMT_panne_voitureloc	Montant de sinistres panne X voitureloc	Claims	
Modele	Modèle	Factor	109
RGEN	Generation	Factor	11
RENER	Energie	Factor	10
P_ADMIN	Puissance Administrative	Num	
RGRE	Genre	Factor	2
NB_PLACES	Nb Places	Num	

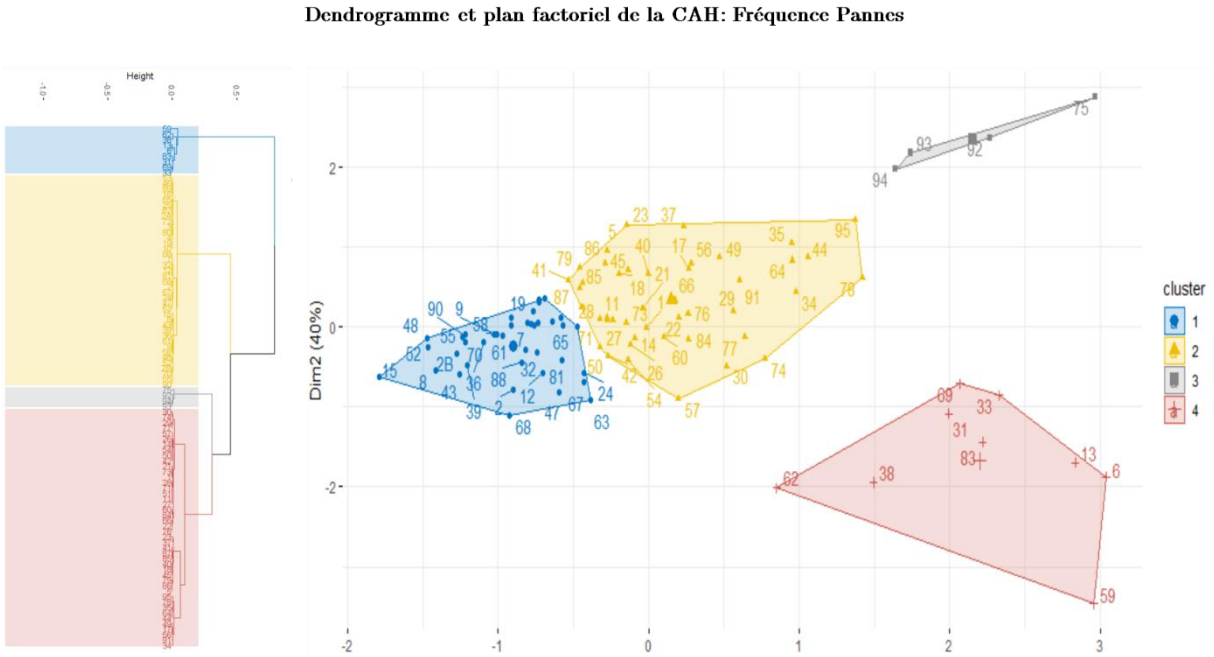
RSER	Série Limitée	Factor	2
NB_CYL	Nombre de cylindres	Num	
RDCY	Disposition de cylindres	Factor	5
CYLND	Cylindrée (en cm3)	Num	
VIT_MAX	Vitesse Maximum en Km/h	Num	
RPMO	Position du Moteur	Factor	3
PUISS_CEE	Puissance CEE (en KW)	Num	
PUISS_DIN	Puissance DIN (en CV DIN)	Num	
REG_PUISS	Regime Puissance Maximum	Num	
COUPL_MOT_MAX	Couple Moteur Max	Num	
REG_COUPL_MOT_MAX	Regime de couple moteur max	Num	
RTRA	Traction	Factor	4
RBDV	Boite de Vitesse	Factor	6
NB_RAPP	Nombre de rapport	Num	
RSUS	Suspension	Factor	4
RTFR	Type de freins	Factor	3
RAFR	Assistance Freinage	Factor	2
LONG	Longueur (mm)	Num	
LARG	Largeur (mm)	Num	
HAUT	Hauteur (mm)	Num	
EMPATT	Empattement (en mm)	Num	
VOIE_AV	Voie avant (en mm)	Num	
VOIE_ARR	Voie arrière (en mm)	Num	
POIDS	Poids à vide (en Kg)	Num	
PTAC	PTAC (en Kg)	Num	
CHARGE_UT	Charge utile (en Kg)	Num	
RACV	Airbag conducteur	Factor	3
RAPV	Airbag passager avant	Factor	3
RALV	Airbags latéraux avant	Factor	4
RALR	Airbags latéraux arrière	Factor	4
RDAS	Direction assistée	Factor	4
RAFU	Assistance de freinage d'urgence	Factor	4
RABR	Antiblocage de roues	Factor	5
RCDS	Contrôle dynamique de stabilité	Factor	4
RADE	Antidémarrage actuel	Factor	9
SEC_PASS	Note de sécurité passive	Num	
RCRE	Classe de réparation actuelle	Factor	28
RSLO	Système de localisation	Factor	2
EMM_CO2	Emission CO2 (g/km)	Num	
ESP	Etoiles de securité passive	Num	
SYST_SS	Système Stop Start	Num	
RSAO	Système anticipation obstacle	Factor	2
RDP2	Dispositif 2	Factor	5

Annexe A.2 : Age du véhicule en fonction de l'ancienneté de permis

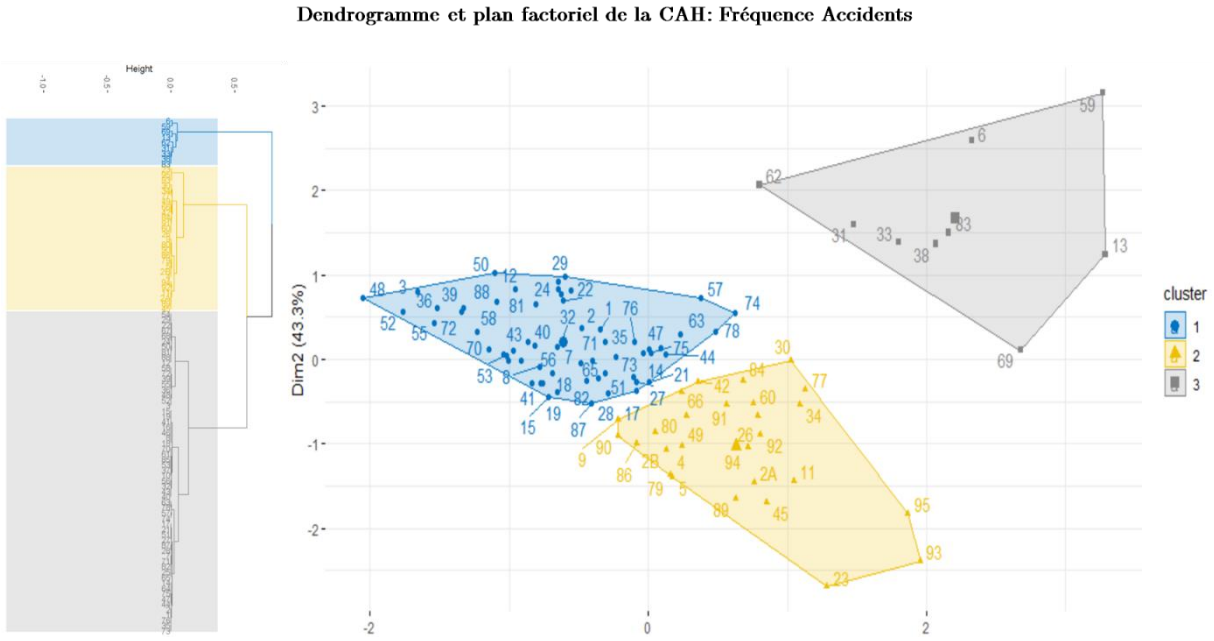


Annexe B : Résultats de la CAH

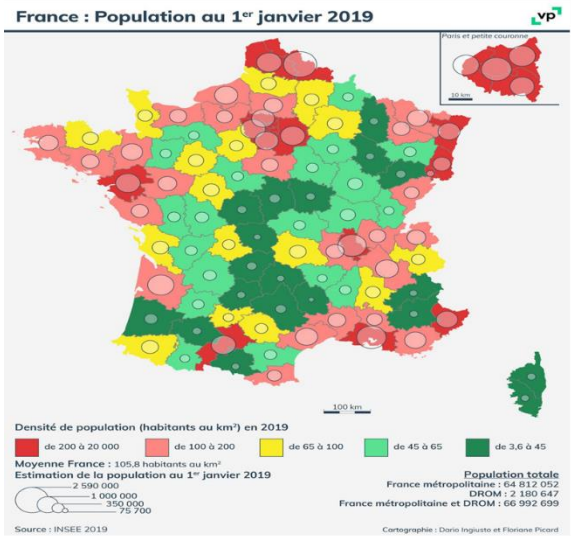
Annexe B.1 : Dendrogramme et plan factoriel pour l'identification des classes du zonier de la fréquence liée aux pannes

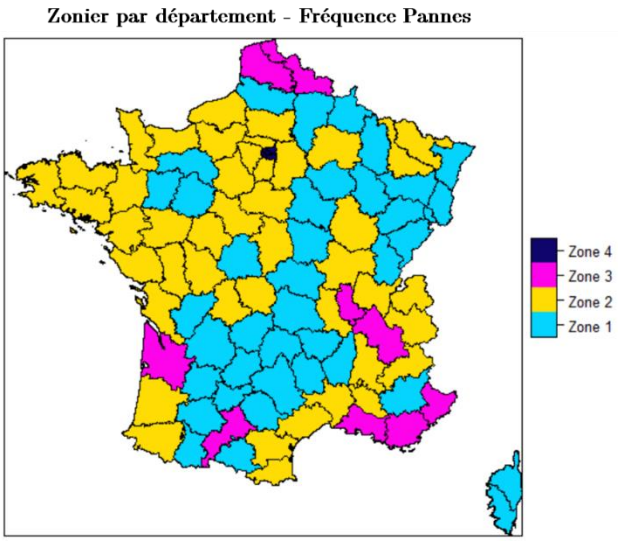


Annexe B.2 : Dendrogramme et plan factoriel pour l'identification des classes du zonier de la fréquence liée aux accidents.

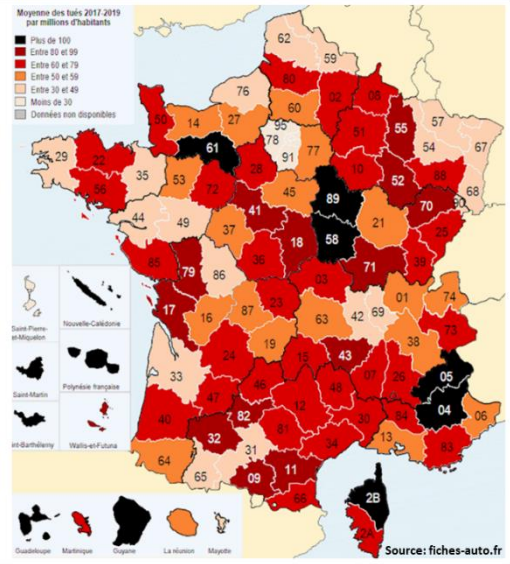


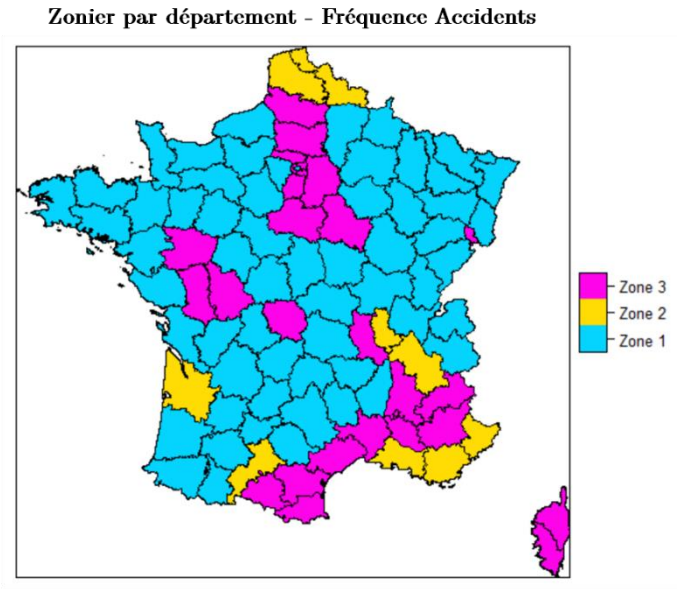
Annexe B.3 : Rapprochement entre la densité démographique par département et le zonier de la fréquence liée aux pannes





Annexe B.4: Rapprochement entre la mortalité routière (moyenne entre 2017 et 2019) par département et le zonier de la fréquence liée aux accidents





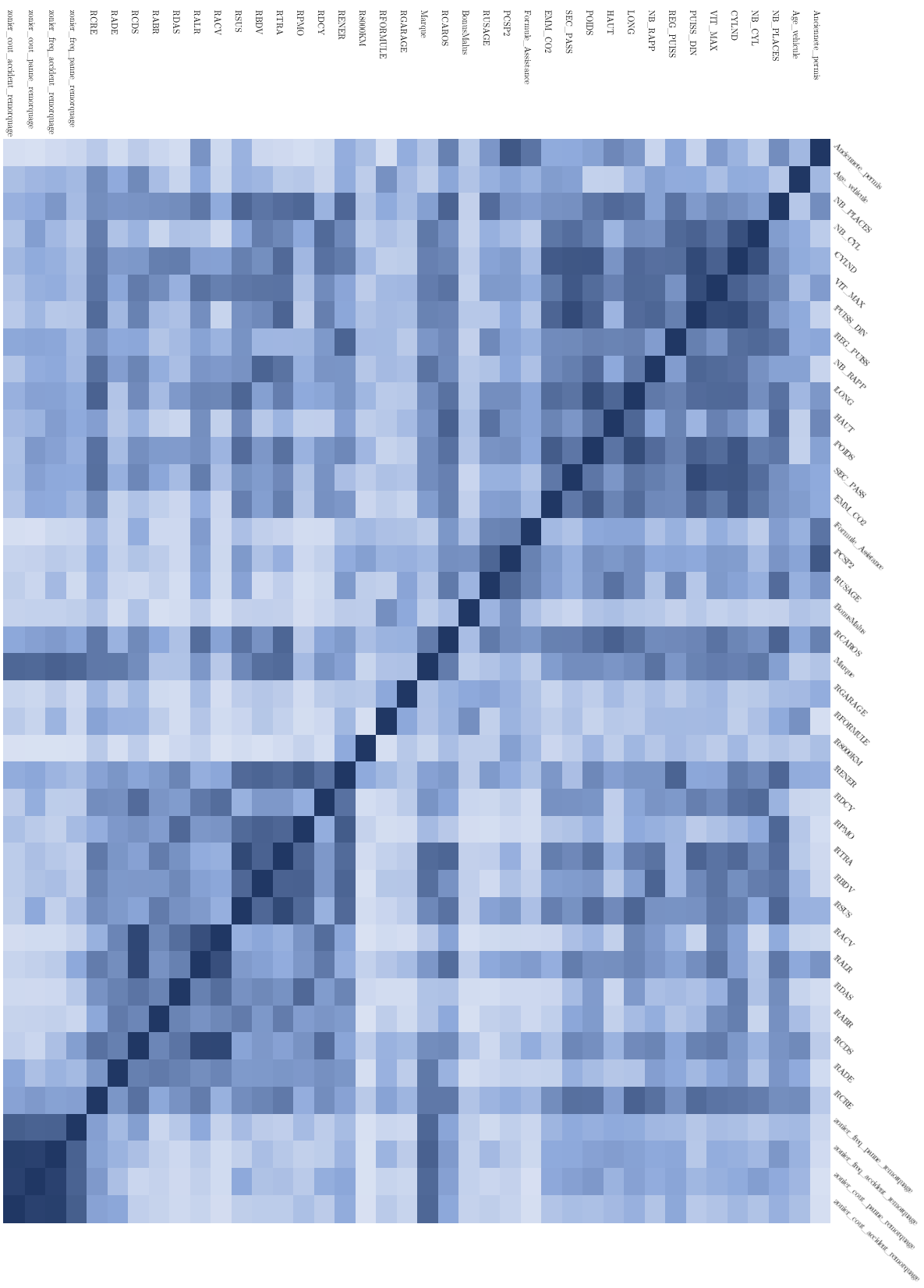
Annexe B.5 : Caractérisation des zones par la moyenne de certaines variables explicatives (pannes)

Zone	Age Véhicule (Années)	Ancienneté Permis (Années)	Vitesse Maximale (Km/h)	Puissance DIN (Cheval DIN)	Poids (Kg)	Longueur (mm)
Zone 1	5,58	6,08	182,08	115,98	1 341,24	4 325,52
Zone 2	5,45	6,94	182,52	117,89	1 357,70	4 350,15
Zone 3	5,31	7,78	182,02	116,35	1 333,51	4 312,43
Zone 4	5,18	12,79	185,34	124,69	1 357,64	4 327,23
Total	**5,48**	**6,91**	**182,41**	**117,25**	**1 348,74**	**4 335,66**

Annexe B.6 : Caractérisation des zones par la moyenne de certaines variables explicatives (accidents)

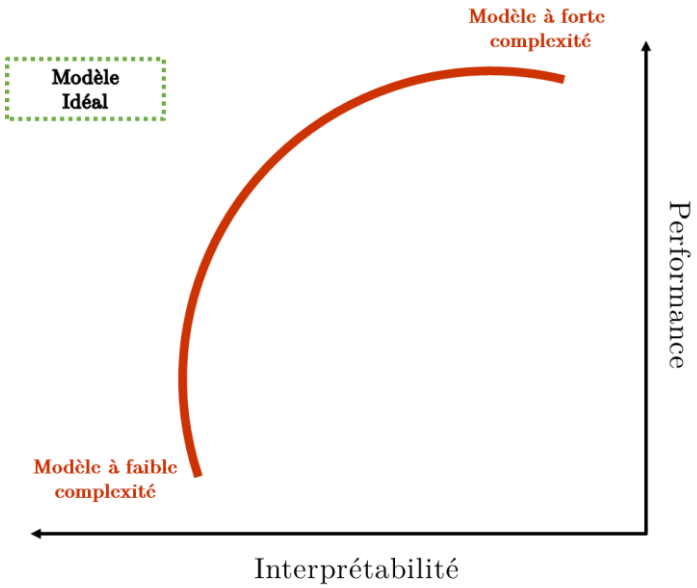
Zone	Age Véhicule (Années)	Ancienneté Permis (Années)	Vitesse Maximale (Km/h)	Puissance DIN (Cheval DIN)	Poids (Kg)	Longueur (mm)
Zone 1	5,51	6,46	182,67	117,33	1 352,16	4 343,36
Zone 2	5,31	7,78	182,02	116,35	1 333,51	4 312,43
Zone 3	5,48	7,63	181,97	117,38	1 346,22	4 326,28
Total	**5,48**	**6,91**	**182,41**	**117,25**	**1 348,74**	**4 335,66**

Annexe B.7 : Matrice de corrélation Phi-K en incluant les variables de type zonier

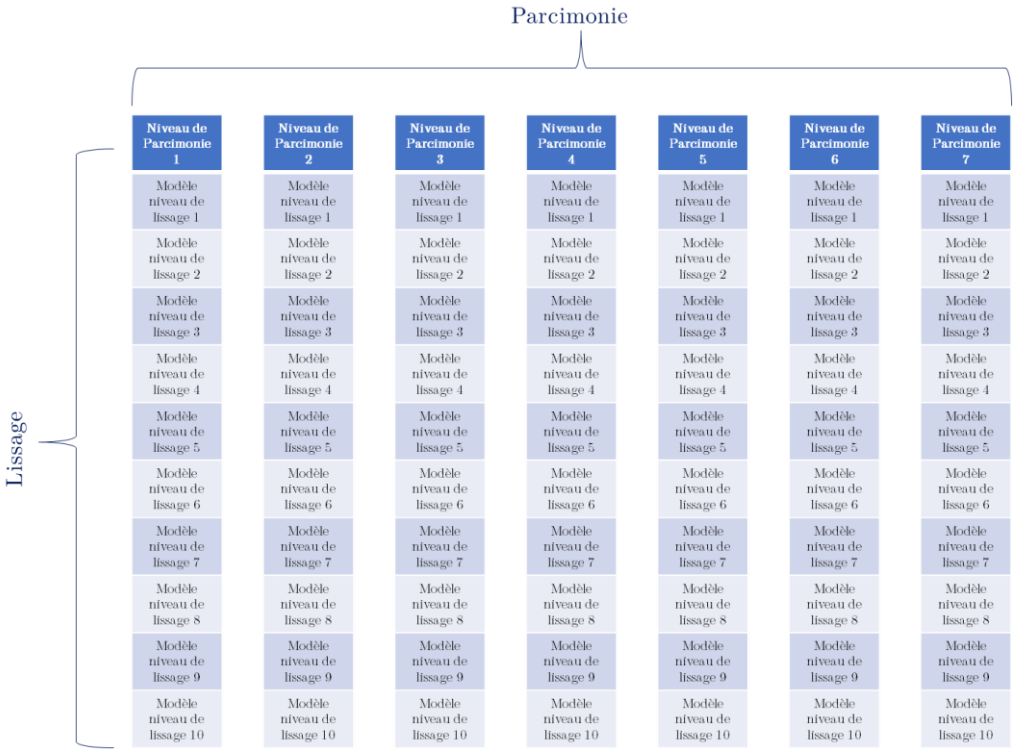


Annexe C : Modélisation de la prime pure

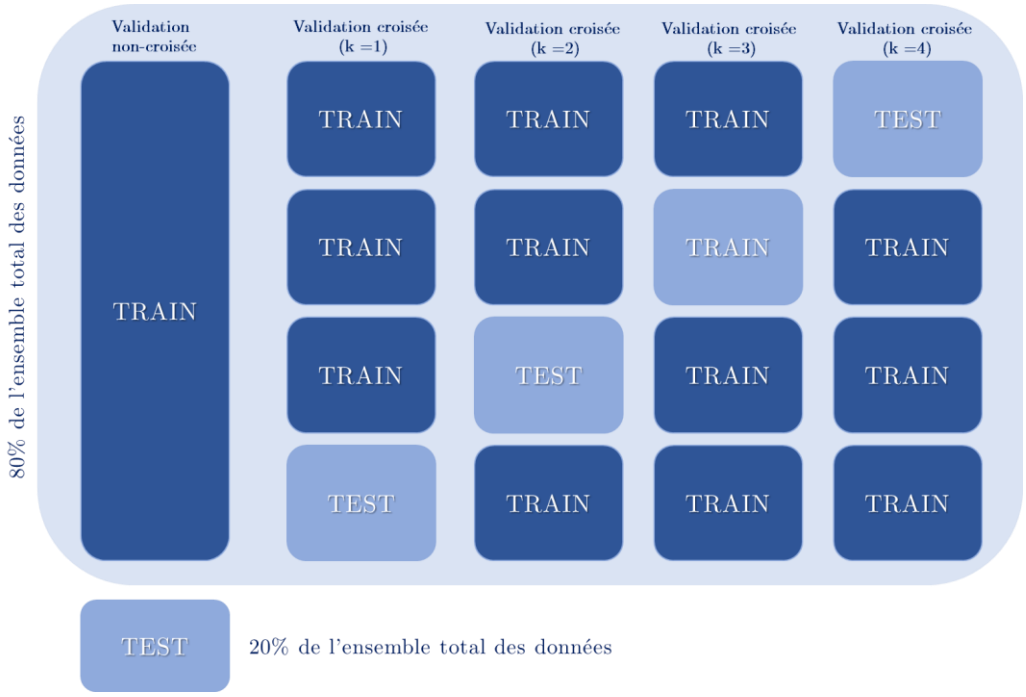
Annexe C.1 : Impact de la parcimonie et le compromis complexité, performance et interprétabilité



Annexe C.2 : Structure de la modélisation sur Akur8



Annexe C.3: Techniques de validation appliquées à la modélisation



ANNEXE C.4: Critères de performance des modèles

Annexe C.4.1 : Critères de performance du modèle à 14 variables pour la fréquence des pannes

	Train K-Fold	Train Global	Validation K-Fold	Validation Globale
Coefficient de Gini	33,92%	33,82%	34,32%	33,63%
Coefficient de Gini Norm.	35,17%	35,07%	35%	36%
Pseudo-R2	5,70%	5,66%	5,92%	5,60%
Déviance	187,6k	250,2k	62,91k	62,6k
RMSE	0,3434	0,34,35	0,3634	0,3723
Valeur moyenne observée	0,06	0,06	0,061	0,06
Valeur moyenne prédite	0,06	0,06	0,06	0,06

Annexe C.4.2 : Critères de performance du modèle à 20 variables pour la fréquence des pannes

	Train K-Fold	Train Global	Validation K-Fold	Validation Globale
Coefficient de Gini	35;52%	35.44%	35,05%	36,00%
Coefficient de Gini Norm.	36,83%	36,75%	36,34%	37,23%
Pseudo-R2	6,29%	6,26%	6,11%	6,50%
Déviance	186,4k	248,6k	62,3k	62,53k
RMSE	0,3432	0,3432	0,3432	0,34
Valeur moyenne observée	0,06	0,06	0,06	0,061
Valeur moyenne prédite	0,06	0,06	0,06	0,06

Annexe C.4.3 : Critères de performance du modèle à 14 variables + interactions pour la fréquence des pannes

Les interactions potentielles analysées dans le cadre de la modélisation de la fréquence des pannes sont les suivantes :

- Age du véhicule x Ancienneté de permis
- Zonier x Age du véhicule
- Zonier x Ancienneté de permis
- Zonier x Marque
- **Marque x Age du véhicule (interaction retenue pour la modélisation)**
- **Puissance DIN x Marque (interaction retenue pour la modélisation)**
- **Nombre de cylindres x Marque (interaction retenue pour la modélisation)**
- Boîte de vitesse x Marque
- Vitesse Maximale x Marque
- Type de Carrosserie x Marque
- Poids x Marque
- Hauteur du Véhicule x Marque

	Train K-Fold	Train Global	Validation K-Fold	Validation Globale
Coefficient de Gini	35,22%	35,10%	34,72%	35,38%
Coefficient de Gini Norm.	36,52%	36,40%	36,01%	36,69%
Pseudo-R2	6,17%	6,12%	5,99%	6,29%
Déviance	186,7k	249k	62,3k	62,67k
RMSE	0,3432	0,3433	0,3433	0,34
Valeur moyenne observée	0,06	0,06	0,06	0,061
Valeur moyenne prédite	0,06	0,06	0,06	0,06

Annexe C.4.4 : Critères de performance du modèle à 14 variables fréquence accidents

Les interactions potentielles analysées dans le cadre de la modélisation de la fréquence des accidents sont les suivantes :

- Age du véhicule x Ancienneté de permis
- Zonier x Age du véhicule
- **Zonier x Ancienneté de permis (interaction retenue pour la modélisation)**
- Zonier x Profession du conducteur principal
- Ancienneté permis x Bonus-Malus
- Régime de puissance maximale x vitesse maximale
- Régime de puissance maximale x Boite de vitesse

	Train K-Fold	Train Global	Validation K-Fold	Validation Globale
Coefficient de Gini	30,38%	30,17%	29,30%	29,20%
Coefficient de Gini Norm.	30,65%	30,44%	29,47%	29,37%
Pseudo-R2	3,27%	3,22%	3,02%	3,02%
Déviance	75,3k	100,6k	25,2k	25,2k
RMSE	0,2047	0,2047	0,2047	0,205
Valeur moyenne observée	0,0164	0,0164	0,0164	0,016
Valeur moyenne prédite	0,0164	0,0164	0,0164	0,016

Annexe C.4.5 : Critères de performance modèle de sévérité des pannes

	Train K-Fold	Train Global	Validation K-Fold	Validation Globale
Coefficient de Gini	5,58%	5,57%	5,49%	5,24%
Coefficient de Gini Norm.	28,78%	28,70%	28,32%	27,03%
Pseudo-R2	8,16%	8,11%	7,86%	7,34%
Déviance	3,3k	4,4k	1,1k	1,1k
RMSE	32,2	32,21	32,25	32,43
Valeur moyenne observée	90,44	90,44	90,44	90,27
Valeur moyenne prédite	90,39	90,39	90,39	90,38

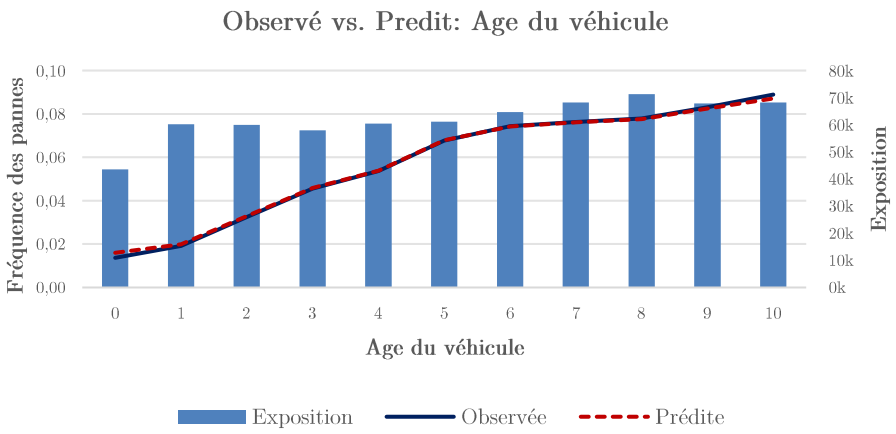
Annexe C.4.6 : Critères de performance modèle de sévérité des accidents

	Train K-Fold	Train Global	Validation K-Fold	Validation Globale
Coefficient de Gini	4,10%	3,86%	3,41%	3,64%
Coefficient de Gini Norm.	17,81%	16,70%	14,83%	15,83%
Pseudo-R2	2,95%	2,95%	2,15%	2,70%
Déviance	1,4k	1,8k	0,5k	0,5k
RMSE	45,68	45,75	45,87	45,94
Valeur moyenne observée	110,8	110,8	110,8	111,3
Valeur moyenne prédite	110,7	110,8	110,7	110,8

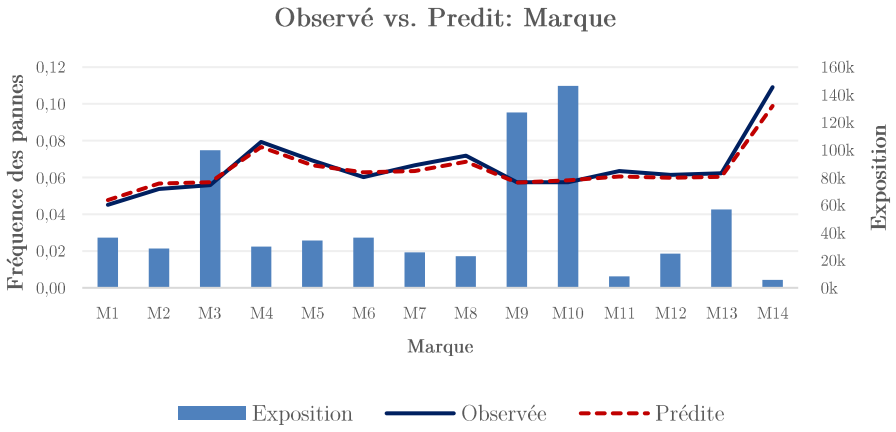
Annexe D : Visualisation du comportement des variables discriminantes

Annexe D.1. : Modélisation des fréquences des pannes

Annexe D.1.1 : Age du véhicule - Fréquences des pannes Observée vs. Prédite



Annexe D.1.2 : Marque - Fréquences des pannes Observée vs. Prédite



Annexe D.1.3 : Zonier - Fréquences des pannes observée vs. Prédite