duas
Diplôme Universitaire des Actuaires de Strasbourg

UFR Mathématique Informatique Strasbourg

Mémoire présenté devant

l'UFR de Mathématique et Informatique

pour l'obtention du Diplôme Universitaire d'Actuaire de Strasbourg

et l'admission à l'Institut des Actuaires

le 18/01/2022

Par : PRUDENT PIERRE

Titre: Modélisation sans contrainte opérationnelle de la garantie Responsabilité Civile Automobile du produit parc dénommé

Confidentialité : ☐ NON ☒ OUI Durée : ☐ 1 an ☒ 2 ans ☐ 3 ans ☐ 4 ans ☐ 5 ans

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

Signature :

Membres du jury de l'Unistra :

P. ARTZNER
J. BERARD
A. COUSIN
K.-T. EISELE
M. MAUMY-BERTRAND

Entreprise :AXA France

Directeur de mémoire en entreprise:
Nom :
Signature :
Invité :

Nom : DECAUX Brice
Signature :

Jury de l'Institut des Actuaires :

A. YOU
Y. RONG

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

Signature du responsable entreprise

Signature du candidat

Secrétariat : Mme Stéphanie Richard

Bibliothèque : Mme Christine Disdier

Résumé

Depuis l'obligation d'assurance sur la garantie Responsabilité Civile, l'automobile est devenue un marché très concurrentiel. Les automobilistes, obligés de s'assurer, cherchent les offres les moins chères. Les assureurs ont donc une marge de manœuvre réduite sur leur tarification. Pour garantir la rentabilité de la branche automobile, l'assureur a pour objectif de segmenter au mieux le risque afin d'adapter la prime aux caractéristiques liées au risque porté par l'assuré.

La segmentation du risque est contrainte par des aspects commerciaux ou opérationnels. Un risque de dégradation de la rentabilité impose une tarification sans contrainte pour juger de la performance actuelle du modèle. La fiabilisation de la base de données est essentielle afin d'assurer en premier lieu la qualité de la modélisation. Les variables sont pré-sélectionnées et pré-retraitées par des méthodes de regroupement adaptées afin de consolider l'analyse du risque. L'élaboration d'un véhiculier a pour but de classifier l'information des véhicules en risques homogènes, et de présenter des données nouvelles.

L'objectif de l'étude est d'appréhender au mieux les spécificités de la garantie Responsabilité Civile sur les véhicules légers à quatre roues du produit *parc dénommé*. Une partie de la charge des sinistres introduit un biais dans la modélisation, dû à la notion de responsabilité et d'indemnisation du préjudice. Le but est alors de choisir le modèle adapté. De plus, un traitement spécifique des sinistres graves est nécessaire à cause de la forte volatilité des dommages corporels. Le seuil des graves peut être déterminé par la théorie des valeurs extrêmes. Pour optimiser la performance, et ainsi la segmentation du modèle tarifaire actuel, les travaux reposeront sur le modèle linéaire généralisé qui sera confronté à des méthodes d'apprentissage pour valider les résultats. L'étape de modélisation permet également d'expliquer et de comprendre le risque.

Mots clés : Assurance automobile, Responsabilité Civile, tarification sans contrainte, gestion de bases de données, seuil des graves, théorie des valeurs extrêmes, modèle linéaire généralisé, segmentation, classification des véhicules, machine learning, méthode de regroupement.

Abstract

Since the introduction of compulsory third party liability insurance, the motor industry has become highly competitive. Obliged to insure themselves, drivers look for the cheapest offers. Therefore, insurers have little flexibility in their pricing. To ensure the profitability of the automotive industry, the insurer has to split the risk as best as possible. Hence, the risk premium is being adapted depending on the characteristics of the risk carried by the insured person.

Risk segmentation is affected by commercial or operational constraints. The risk of degradation of profitability requires unconstrained pricing to judge the current performance of the model. The reliability of the database is essential in order to ensure the quality of the modelling. The variables are pre-selected and pre-reprocessed using appropriate clustering methods to strengthen the analysis of the risk. So, the development of a car requires classification of the car's information into homogenous risks and the introduction of new data.

The aim of the study is to understand the specific characteristics of the Third party liability coverage on light four-wheeled vehicles from "Parc dénommé". Part of claims incurred introduces a bias in modelled results related to the concept of liability and compensation of the third party loss. The goal is then to choose the right model. Furthermore, a specific processing of severe casualties is needed due to the high volatility of personal injuries. The threshold for severe claims can be determined by the Extreme-value theory. To optimize efficiency, and thus the segmentation of the current tariff model, the study will be based on the generalized linear model, which will be tested against learning methods to validate the results. The step of modelling gives a better explanation and understanding of the risk.

Keywords : Car insurance, third party liability, unconstrained pricing, database management, severity threshold, Extreme-value theory, generalized linear model, segmentation, vehicle classification, machine learning, clustering method.

Note de synthèse

Contexte

En France, pour circuler légalement, tout véhicule à moteur est tenu d'être assuré par une garantie Responsabilité civile. La responsabilité civile couvre les préjudices corporels et matériels causés par l'assuré, ou subis par des tiers. La couverture indemnise tous les sinistres, quelle que soit leur gravité. En tant que garantie minimale, elle présente une forte exposition. La concurrence sur le marché de l'assurance automobile a nettement fait baisser les prix. La révision continuelle des offres a pour but de mieux décrire le risque et de rendre plus attrayant le produit. Plusieurs produits peuvent être proposés aux entreprises, en fonction de leurs besoins. Le produit *parc dénommé* représente la part la plus importante du chiffre d'affaires chez AXA en auto-entreprise. Il s'agit d'un contrat proposé aux entreprises pour assurer un parc de 5 à 50 véhicules.

Le modèle tarifaire actuel des parcs dénommés est soumis à des contraintes opérationnelles ou commerciales et est mis à jour tous les trois ans. La révision ne concerne pas la segmentation du risque. En 2021, une augmentation majeure de la prime a été réalisée afin de remettre à niveau le tarif et de prendre en compte la forte inflation générée par le coût des pièces automobiles. La rentabilité du produit était en déclin et ne répondait plus aux objectifs fixés. En raison des difficultés d'implémentation informatique, l'introduction de nouvelles variables n'a pas été envisagée, car elle implique un coût conséquent pour une rentabilité qui serait longue et progressive.

Le sujet de l'étude est de modéliser sans contrainte opérationnelle la garantie Responsabilité Civile du produit *parc dénommé* sur les véhicules de catégorie 1 : véhicules terrestres à moteur à quatre roues et de moins de $3,5$ tonnes. En effet, la sinistralité est impactée différemment selon la catégorie du véhicule, ainsi, pour chaque catégorie, un calibrage est effectué. La première catégorie de véhicule représente 60% en nombre et 50% en prime en 2019. Il s'agit donc d'un périmètre conséquent dans l'automobile entreprise. L'objectif est de comparer les performances du modèle actuel aux performances du modèle sans contrainte pour connaître l'impact d'une segmentation à partir de nouvelles données, notamment véhicules ou financières.

En fait, la nouvelle version tarifaire s'applique seulement sur les affaires nouvelles. Pour un véhicule s'ajoutant à un contrat renouvelé, la version tarifaire utilisée est celle qui a été fixée lors de la date d'effet du contrat. Ainsi, les véhicules du parc sont tarifés de manière cohérente et le client comprend mieux l'évolution de son tarif. De plus, lors du renouvellement des contrats, une majoration est appliquée en considérant le jugement d'expert des intermédiaires d'assurance et un ratio. Ce dernier permet d'inflater le tarif et de prendre en compte d'éventuels changements de segmentation du risque. Actuellement, ce ratio est le rapport entre la charge des sinistres et les primes. Dans un futur proche, il pourrait devenir le rapport entre la prime pure modélisée et la prime annuelle.

Un changement radical de segmentation a pour conséquence un coût informatique pour la tarification des affaires nouvelles important. Ce coût est encore plus important en implémentant un tarif avec une nouvelle segmentation dans l'outil de tarification des anciens contrats à cause d'outils dont les dépendances sont nombreuses. Les affaires nouvelles représentent plus de 10% des contrats, l'implémentation d'un nouveau tarif est donc lente et progressive. La rentabilité de la mise en place d'un nouveau tarif n'est pas immédiate. C'est pourquoi, les résultats doivent présenter une réelle différence de performance entre les deux tarifs sans et avec contrainte pour que le modèle sans contrainte soit mis en place.

L'idée du tarif sans contrainte est de mieux segmenter le risque. Un tel tarif peut permettre d'appliquer des majorations plus proches de la réalité lors du renouvellement sur l'ensemble du portefeuille sans coût informatique supplémentaire. Ainsi, un gain de performance faible du tarif sans contrainte est accepté.

Traitement des données

L'estimation de la prime pure d'un risque est réalisée sur un jeu de données composé d'un historique de la sinistralité et de toutes les informations disponibles qui pourraient expliquer le risque. Plus particulièrement, ces données proviennent :

— d'une base sur les sinistres,
— d'une base sur les véhicules,
— d'une base sur les contrats et les distributeurs d'assurance (courtiers et agents),
— d'une base sur les entreprises.

L'ensemble de ces informations doit être cohérent, de bonne qualité et le plus exhaustif possible. Pour assurer la cohérence et l'exhaustivité, le jeu de données est constitué par des informations issues des sources internes puis complété par des sources externes.

La description des véhicules est détaillée dans deux sources externes : le fichier SIV (Système d'Immatriculation des Véhicules) et la base SRA (Sécurité Réparation Automobile). Le fichier SIV est développé par le ministère de l'Intérieur et il recense l'ensemble des informations inscrites sur la carte grise des véhicules immatriculés en France. Les données qui y sont fournies, ont été utilisées pour développer le tarif actuellement mis en place. En revanche, la base SRA est une source d'information nouvelle. Elle est composée des caractéristiques techniques et commerciales des véhicules terrestres à moteur de moins de 4 roues et de moins de 3,5 tonnes. L'étude se concentrera donc sur le gain apporté par ces nouvelles données.

La base SRA est à la maille type de véhicule contrairement au fichier SIV ou à la base interne qui sont à une maille plus fine : l'immatriculation du véhicule. Un algorithme a été amélioré dans le but de récupérer l'information tirée de la base SRA. Cet algorithme permet de compléter l'information relative à la base SRA de +268%. Il juxtapose les informations véhicules du fichier SIV et celles de la base véhicule interne aux informations affectées à la base SRA. L'hypothèse est que les codes SRA les plus pertinents sont ceux déjà présents dans la base. En fait, la juxtaposition des données ne permet pas de définir dans tous les cas un unique code SRA. Ainsi, le code choisi sera celui qui présente la plus grande fréquence dans les données. En effet, il est raisonnable de penser que les mêmes types de véhicules d'une entreprise à l'autre sont sélectionnés. C'est d'ailleurs ce qui est constaté en observant la forte redondance des codes SRA.

Enfin, un second algorithme est implémenté afin de compléter les données. La liaison entre la base sinistre et la base véhicule génère une perte d'informations sinistres due à des erreurs de saisies de données sur les immatriculations de véhicules. Environ cinq sixième des sinistres non captés, représentant $5,5\%$ des sinistres, sont interprétés ou rectifiés. L'algorithme corrige les immatriculations des véhicules de la base sinistre à partir de différentes combinaisons en les comparant aux immatriculations de la base véhicule. En effet, des morceaux du numéro d'immatriculation sont modifiés afin de faire correspondre le sinistre au bon véhicule. L'étude est réalisée à la maille vision et contrat sur toutes les catégories de véhicules et sur les contrats composés uniquement de véhicules de moins de $3,5$ tonnes (parc VL). Cela permet de se concentrer plus particulièrement sur les véhicules terrestres à moteur de moins de $3,5$ tonnes composant essentiellement ce type de parc. Parmi ces sinistres non captés, un quart s'explique par une immatriculation non renseignée liée aux catégories de véhicules 4 et 5 (engins agricoles et engins de chantier).

Hormis ces deux algorithmes, la complétion des données est classiquement produite par la cohérence des informations sorties des bases. La base véhicule et la base contrat sont supposées être de meilleures qualité que les autres bases. C'est pourquoi, ce sont leurs données qui seront considérées comme une référence. En effet, la base contrat et la base véhicule sont construites par des outils de souscription qui réduisent fortement le risque d'erreur, de plus le format des informations suit un certain modèle défini par AXA.

Malgré l'étape de complétion, des données restent manquantes. Pour pallier à ce problème, plusieurs choix s'offrent à nous. Lors de la mise en place d'un tarif, les données nécessaires au calcul de la prime pure doivent être obligatoirement renseignées. Par contre, des données qui ne font pas partie du tarif ne sont pas obligatoirement renseignées par l'assuré. Ainsi, des données sont manquantes. Cette perte d'information peut produire un biais dans la modélisation du risque, car, par définition, ces données peuvent contenir soit une information importante soit une information déjà captée par le modèle. Ainsi, en fonction de la méthode utilisée pour modéliser le risque, les données manquantes seront traitées différemment. Le jeu de données n'est donc pas directement impacté. Par la suite, ce sont des modèles de type GLM qui seront utilisées et les variables quantitatives seront catégorisées. Les données non renseignées seront conservées et seront modélisées comme une modalité à part entière. Ceci permet de modéliser

les données manquantes, d'étudier leur comportement et de réaliser un regroupement possible avec les données décrivant un effet commun. Toutefois, les variables composées de trop peu d'informations sont écartées de l'étude.

Finalement, la collecte et le traitement des données permettent d'assurer la qualité de la modélisation. Le jeu de données est divisé en deux, 20% de la base sera utilisée à des fins de tests de performance et 80% dans un but de calibrage des modèles.

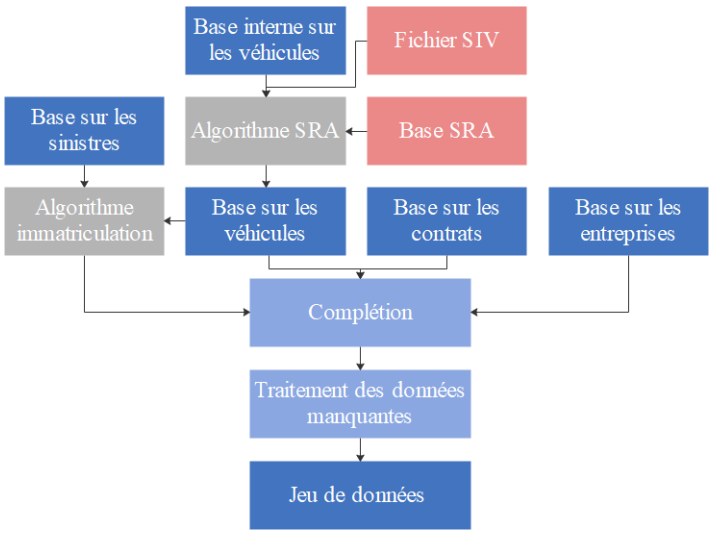


FIGURE 1 – Cartographie des traitements sur le jeu de données

Modélisation

L'assurance est un service qui fournit une prestation lors de la survenance d'un événement incertain et aléatoire. Ce risque est assuré en échange de la perception d'une prime d'assurance. Une prime se définit comme l'addition de la prime pure, de frais et de bénéfice. C'est la prime pure que l'on souhaite estimer. Il s'agit du montant du sinistre moyen en fonction de variables tarifaires. La sinistralité est aléatoire cependant une partie du risque peut-être expliquée à partir des données disponibles, notamment sur les véhicules, sur les contrats et sur les entreprises. Finalement, nous souhaitons modéliser la prime pure liée à la garantie Responsabilité Civile.

La tarification est réalisée sans contrainte opérationnelle et donc avec toutes les données disponibles. Plus de 80 variables peuvent être tarifaires. Créer un modèle composé de l'ensemble des variables aura tendance à être trop ajusté aux données, c'est-à-dire suivre une variabilité similaire aux données sinistres. Cependant, ce genre de modèle se révèle très peu performant sur un nouveau jeu de données, notamment en ce qui concerne les modèles de type GLM (modèle linéaire généralisé) ou GAM (modèle additif généralisé). Ainsi, une sélection parcimonieuse et une étape de lissage des variables sont réalisées avant chaque ajustement à l'aide d'une méthode spécifique à un outil de calibrage. En fait, plusieurs GAM sont construits pour répondre à des questions de calibrage. Les modèles GAM et GLM sont couramment utilisés en assurance.

La Responsabilité Civile (RC) peut être liée à des dommages corporels (RC CORP) ou matériels (RC MAT). Il est possible de distinguer la RC en fonction du dommage, car le comportement des deux sous risques est différent. Cependant, la RC CORP présente une trop faible sinistralité pour être modélisée séparément. De plus, elle est en général associée à des sinistres de charges graves qui seront écartés par la suite. La modélisation de la RC présente une performance similaire à l'agrégation des deux modèles pour un nombre de variables explicatives moindres. Ainsi, l'ajustement qui suivra, sera construit sur la sinistralité de la RC quel que soit le dommage lié.

Un GAM ou un GLM sont des modèles paramétriques faisant intervenir une loi exponentielle pour s'ajuster aux données. Les lois retenues qui semblent le plus en adéquation avec les données, sont respectivement, pour le modèle de fréquence et pour le modèle de coût moyen, la loi de Poisson et loi Gamma. Il s'agit d'un modèle collectif, c'est-à-dire que l'estimation de la prime pure est réalisée à par-

tir de l'agrégation d'un modèle de fréquence et de coût moyen permettant une meilleure description du risque.

Enfin, dans l'étape de pré-calibrage, la charge est retraitée. Les coûts d'ouverture, les recours, ou les sinistres dont la charge est liée aux conventions IRSA et IRCA permettant d'indemniser l'assuré plus rapidement, biaisent le modèle. En effet, leur charge ne correspond pas au réel montant indemnisé ou elle est négatif. Ces sinistres sont donc écartés de la modélisation et font l'objet d'une estimation forfaitaire de la prime pure. En parallèle, les sinistres dont la charge est grave, c'est-à-dire qui dépasse un montant relativement élevé, sont mutualisés au prorata de la prime des sinistrés (hors forfait, recours et coûts d'ouverture) en raison de leur distribution atypique. Le seuil des sinistres graves a été déterminé par la Théorie des Valeurs Extrêmes et plus spécifiquement par la comparaison des résultats des estimateurs de Hill, Pickands et DEdH. C'est un seuil à 18 000€ qui sera retenu.

Dans l'objectif d'améliorer les performances du GLM, une étape de sélection des variables approfondie est produite après regroupement des modalités des variables qualitatives et après catégorisation des variables quantitatives. La re-catégorisation des variables catégorielles est réalisée à l'aide de cinq méthodes :

— test de Wald jugeant de la significativité des coefficients du modèle,
— AIC ou BIC, deux critères de pénalisation, pour comparer des modèles,
— pénalisation de type Lasso par un effet de contraction des coefficients,
— zonier permettant de classifier l'information géographique par projection sur une carte,
— jugement d'expert, les modalités sont regroupées pour leur sens commun opérationnel.

La dernière méthode est également utilisée dans les quatre autres afin de conserver, au-delà d'une cohérence au niveau du risque, une cohérence au niveau de la signification opérationnelle du regroupement.

Enfin, les variables quantitatives sont traitées à l'aide d'une méthode de machine learning : un arbre de décision connu pour son pouvoir de classification. Pour une meilleure robustesse, c'est l'effet lissé de la variable par des fonctions nommées splines qui est mis en entrée de l'arbre.

La catégorisation permet de créer des cases tarifaires, de simplifier le tarif et de le rendre plus interprétable. En assurance, travailler par case tarifaire est en général privilégié et cela n'implique pas de perte de performance.

Par la suite, ce sont en grande partie des variables retraitées qui ont été sélectionnées prouvant la significativité des méthodes de regroupement. De plus, ces méthodes ont permis de mettre en exergue certaines variables qui ont été fortement pénalisées par la sélection liée aux grands nombres de modalités.

De nouvelles variables ont été retenues dans le modèle, notamment des informations sur les véhicules provenant de la base SRA. L'un des objectifs de l'étude est de connaître le gain de performance induit par des variables SRA. Ainsi, une méthode approfondie sur l'effet des véhicule est exploitée, c'est-à-dire les résidus d'un modèle GLM dans lesquels les variables SRA ont été volontairement omises. Cette méthode appelée véhiculier permet de résumer toute l'information véhicule en une unique variable par le biais notamment d'un arbre de décision. Trois différents véhiculiers sont construits. Chacun traite à sa manière l'effet véhicule avant de l'entrer dans l'arbre. Un premier véhiculier utilise l'effet brut, un second l'effet véhicule retiré des données jugées non crédibles à partir de la définition d'un seuil tirée de la théorie de la crédibilité (modèle de Bühlmann-Straub), et le dernier l'effet lissé. Le préambule pour obtenir un effet lissé est de spatialiser les individus à la maille SRA par une Analyse en Composante Multiples (ACM). La projection est réalisée à partir des corrélations des variables véhicules entre elles. Puis, sur cette spatialisation, il est défini le voisinage de chaque individu par la triangulation de Delaunay (géométrie spatiale) pour créer une carte des voisins. Ainsi, l'effet véhicule est lissé à partir du modèle de Bühlmann-Straub crédibilisant l'information à l'aide de la carte des voisins.

Cependant, le gain de performance des véhiculiers est faible, probablement dû à l'application des méthodes de regroupement sur le GLM. De plus, il est plus difficile d'interpréter la sortie des arborescences d'un point de vue opérationnel que d'interpréter des tendances sur les coefficients prédit par le GLM. Ainsi, aucun des véhiculiers n'est retenu.

Tout au long de l'étude, l'unique méthode utilisée pour tarifer le risque est un GLM. D'autres méthodes d'apprentissage existent. Il est donc intéressant de confronter le modèle GLM retenu à des modèles de machines learning tel que le XGBoost ou le Random Forest. Le meilleur modèle au sens des critères de performance est le XGBoost. En revanche, ces ajustements se révèlent être difficiles à interpréter, de par leur construction sur une vaste arborescence, et difficiles à implémenter opérationnellement. La définition de cases tarifaires est moins claire.

L'ensemble du calibrage du modèle de tarification peut être résumé par la cartographie suivante :

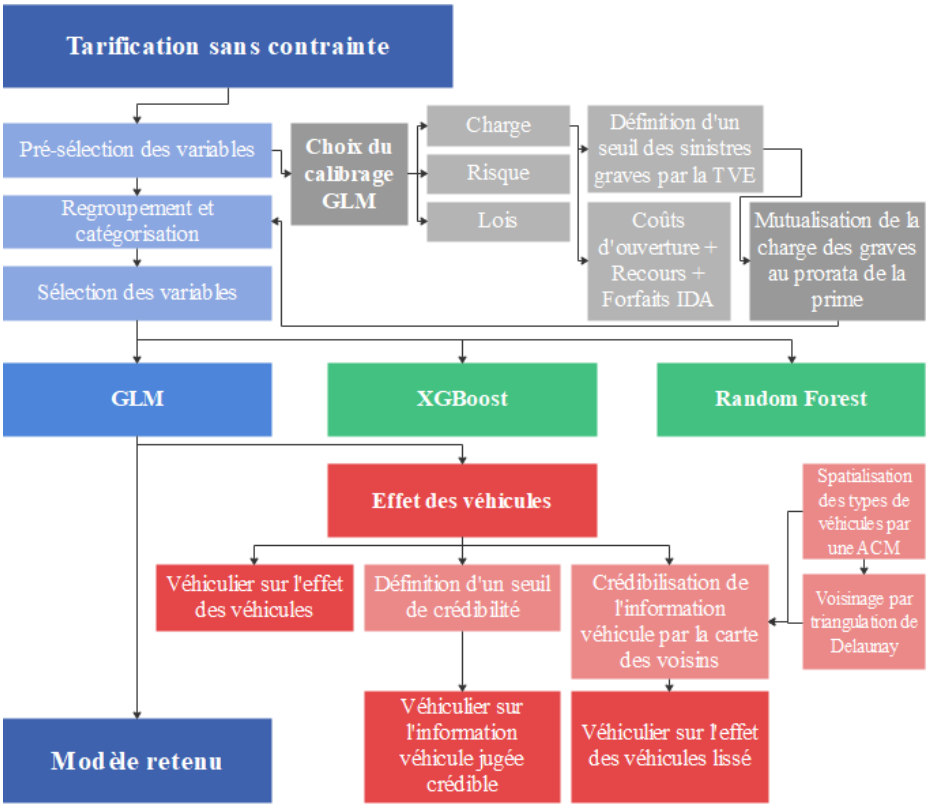


FIGURE 2 – Cartographie de la modélisation

Résultats

Le tarif actuellement commercialisé en auto-entreprise répond au besoin de faciliter la souscription, mais également à des stratégies commerciales. La dernière refonte majeure de tarif a eu lieu en 2016 et introduisait de nouvelles données, notamment véhicules avec le fichier SIV. Depuis, le tarif est mis à jour tous les 3 ans sans apporter de changement majeur dans la segmentation. Le modèle sans contrainte retient une segmentation avec des données nouvelles, notamment des informations sur la santé financière des entreprises, sur le taux d'orientation des assurés vers les garages partenaires par les intermédiaires d'assurance, et sur les véhicules (base SRA).

Le tarif actuel est construit sur un traitement similaire de la charge. L'objectif est donc de juxtaposer les deux tarifs. Un gain de performance est effectivement constaté : une augmentation de +5 points du Gini du modèle de fréquence et de +10 points sur le modèle de coût moyen.

En visualisant les écarts de tarif par un arbre de décision, les variables à but commercial biaisent les performances du tarif actuel. De plus, les primes moyennes des entreprises ayant pour activité le transport de marchandises ou le transport de voyageurs semblent être surestimées. En effet, ces deux classes ont été volontairement majorées lors de la dernière mise à jour de tarification. Ensuite, les nouvelles informations sur les véhicules (carrosserie, hauteur, prix à l'origine du véhicule) et le taux d'orientation vers les garages partenaires se révèlent être significatives dans la segmentation du risque. Ces dernières ne sont pas suffisamment captées par le modèle sous contrainte.

En outre, l'écart de tarif révèle une prime moyenne sur l'ensemble du jeu de données de test de -1%, c'est-à-dire que la prédiction totale du modèle sans contrainte est inférieure à la prédiction du modèle sous contrainte. Cela s'explique en partie par la perte d'informations de la sinistralité due aux sinistres non captés par la base véhicule après application de l'algorithme immatriculation.

En raison des nombreux retraitements réalisés sur la charge, une partie de l'étude se consacre à réconcilier la prédiction du modèle sans contrainte à la charge brute. Les tendances sont similaires, par contre le gain de performance est moindre par rapport au tarif actuel. Cela s'explique par la forte présence de sinistres sous convention IRSA ou IRCA et de coûts d'ouvertures. En effet, ces sinistres représentent 64% de la fréquence et 26% de la charge. Finalement, la modélisation sans contrainte opérationnelle montre la difficulté d'obtenir un gain de performance significatif sur la garantie Responsabilité Civile à cause

des conventions IRSA et IRCA. En revanche, en soustrayant ces forfaits, les variables de la base SRA, notamment, ont fortement participé à l'amélioration des performances du modèle. Par rapport à la rentabilité du produit, il n'y a donc pas d'intérêt d'adopter le modèle sans contrainte pour tarifer les affaires nouvelles, par contre il peut être utilisé pour la majoration des contrats lors des renouvellements.

Les améliorations futures viseront dans un premier temps à reconsidérer l'étude sur le véhiculier. L'idée de créer un véhiculier après avoir défini un effet lissé de l'information sur les véhicules se révèle être trop peu efficace. En effet, la crédibilité est appliquée sur un nombre trop faible de voisins. Une idée pour augmenter le voisinage serait d'exploiter la distance euclidienne, au lieu d'utiliser la triangulation de Delaunay. Enfin, le gain de performance sur la prime modélisée de la Responsabilité Civile grâce aux nouvelles informations, nous incite à réitérer l'étude sur la garantie Dommage. La garantie Dommage est la seconde garantie la plus importante en matière de prime du portefeuille. Elle couvre les dommages matériels subis par le véhicule assuré. Cette garantie est privée de conventions ce qui en fait un candidat idéal pour tester les nouvelles données disponibles. De plus, la base SRA est composée également de véhicules de catégorie 6, les véhicules à deux roues additionnées des quads, ce qui peut faire l'effet d'un examen approfondi du modèle et de l'impact des nouvelles données véhicules.

Mots clés : Assurance automobile, Responsabilité Civile, tarification sans contrainte, gestion de bases de données, seuil des graves, théorie des valeurs extrêmes, modèle linéaire généralisé, segmentation, classification des véhicules, machine learning, méthode de regroupement.

Executive summary

Context

In France, all motor vehicles are required to be insured for third party liability to circulate legally. Civil liability covers bodily injury and material damage caused by the insured, or suffered by third parties. The coverage compensates for all losses, regardless of their severity. As a minimum guarantee, it has a high exposure. Competition in the car insurance market has significantly decreased prices. The ongoing revision of the bids aims to have a better description of the risk and a more attractive product. Several products are offered to companies depending on their needs. The product of 'Parc dénommé' is the largest share of AXA's turnover in auto-enterprise. This is a contract offered to companies to ensure a fleet of 5 to 50 vehicles.

The current pricing model for 'Parc dénommé' is subject to operational or commercial constraints and is updated every three years. The revision does not concern risk segmentation. In 2021, a major increase in premium has been carried out to update the tariff and take into account the high inflation resulting from the cost of car parts. The profitability of the product appeared to be declining and no longer met the goals set. Due to IT implementation difficulties, the introduction of new variables was not considered, as it would involve a significant cost for a profitability that would be long and gradual.

The subject of the study is to model without operational constraint the Civil Liability guarantee of the product of 'Parc dénommé' on category 1 vehicles : four-wheeled motor land vehicles of less than 3.5 tons. Indeed, the loss ratio is impacted differently depending on the category of the vehicle, thus, for each category, a calibration is done. The first category of vehicles represents 60% in number and 50% in premium in 2019. This is a significant scope in the business car sector. The goal is to compare the performance of the current model with the performance of the unconstrained model so as to determine the impact of a segmentation based on new data, especially vehicle or financial data.

In fact, the new tariff version only applies to new business. For a vehicle added to a renewed contract, the tariff version used is the one that was set at the time the contract took effect. In this way, the vehicles in the fleet are priced consistently and the customer has a better understanding of the evolution of his or her tariff. In addition, when contracts are renewed, a mark-up is applied taking into account the expert judgement of insurance intermediaries and a ratio. The ratio influence the tariff and it take into account possible changes in risk segmentation. Currently, this ratio is the ratio between the cost of claims and premiums. In the near future, it could become the ratio between the modelled pure premium and the annual premium.

A radical change in segmentation results in a significant IT cost for pricing new business. This cost is even greater when implementing a tariff with a new segmentation in the pricing tool for old contracts due to the many dependencies of the tools. New business represents more than 10% of contracts, so implementing a new tariff is slow and gradual. The profitability of implementing a new tariff is not immediate. This is why, the results must show a real difference in performance to be implemented, between the two tariffs without and with constraints for the unconstrained model.

The idea of the unconstrained tariff is to better segment the risk. Such a tariff can allow more realistic mark-ups to be applied at renewal, on the portfolio without additional IT costs. Thus, a small performance gain from the unconstrained tariff is accepted.

Data processing

The estimation of the pure premium of a risk is performed on a dataset made of the history of the loss ratio and all available information that could explain the risk. More specifically, these data come from :

— a claim's database,
— a vehicle's database,
— a database on contracts and insurance distributors (brokers and agents),
— a database on companies.

All this information must be coherent, of good quality and as exhaustive as possible. To ensure consistency and completeness, the dataset is made up of information from internal sources and then completed by external sources.

The description of the vehicles is detailed in two external sources : the SIV file (vehicle registration system) and the SRA database (automobile repair safety). The SIV file is developed by the Ministry of the Interior and it lists all the information recorded on the registration card of vehicles registered in France. The data provided in this database has been used to develop the tariff currently in place. Moreover, the database of SRA is a new source of information. It is made up of the technical and commercial characteristics of land motor vehicles of less than 4 wheels and less than 3.5 tons. The study will therefore focus on the advantages of this new data.

The database of SRA relates to the vehicle type level, whereas the SIV file and the internal database are at a finer level : the vehicle registration. An algorithm has been improved to recover information from the SRA database. This algorithm makes it possible to complete the information relating to the SRA database of +268%. It juxtaposes the vehicle information from the SIV file and from the internal vehicle database with the information assigned from the SRA database. The hypothesis is that the most relevant SRA codes are those already present in the database. In fact, the juxtaposition of the data does not make it possible to define a unique SRA code in all cases. So, the code chosen will be the one with the highest frequency in the data. Indeed, it is reasonable to think that the same types of vehicles from a company to another are selected. This can be seen from the high redundancy of the SRA codes.

Finally, a second algorithm is implemented in order to complete the data. The link between the claims database and the vehicle database generates a loss of claims information due to data entry errors on vehicle registrations. Approximately five-sixths of the uncaptured claims, representing 5.5% of claims, are interpreted or corrected. The algorithm corrects vehicle registrations from the claim database using different combinations by comparing them to the registrations from the vehicle database. Indeed, pieces of the registration number are modified to match the claim to the correct vehicle. The study is carried out at the vision and contract level on all categories of vehicles and on contracts composed only of vehicles of less than 3.5 tons. This makes it possible to focus more specifically on motor land vehicles of less than 3.5 tons, which essentially make up this type of fleet. Of these uncaptured claims, a quarter is explained by a registration that was not provided in relation to vehicle categories 4 and 5 (agricultural and construction machinery).

Without these two algorithms, data completion is classically produced by the consistency of the information output from the databases. The vehicle database and the contract database are assumed to be of better quality than the other databases. Therefore, their data will be used as a reference. Indeed, the contract and the vehicle databases are built by underwriting tools that greatly reduce the risk of error, moreover the format of the information follows a certain model defined by AXA.

Despite the completion stage, some data are missing. To overcome this problem, several options are available. When a tariff is set up, the data needed to calculate the pure premium must be completed. Additionally new modeling implies the use of new data whose risk-related information is not mandatory. This results in missing data. This loss of information can produce a bias in the risk modeling, as, by definition, these data may contain either important information or information already captured by the model. Thus, depending on the method used to model the risk, missing data will be treated differently. The dataset is therefore not directly impacted. Subsequently, GLM (Generalized Linear Model) type models will be used and the quantitative variables will be categorized. The unspecified data will be kept and will be modeled as a separated modality. This allows to model the missing data, to study their behavior and to carry out a possible clustering with data describing a common effect. However, variables with too little information are excluded from the study.

Finally, data collection and processing ensure the quality of the modeling. The dataset is split in two, 20% of the database will be used for performance testing and 80% for model calibration purposes.

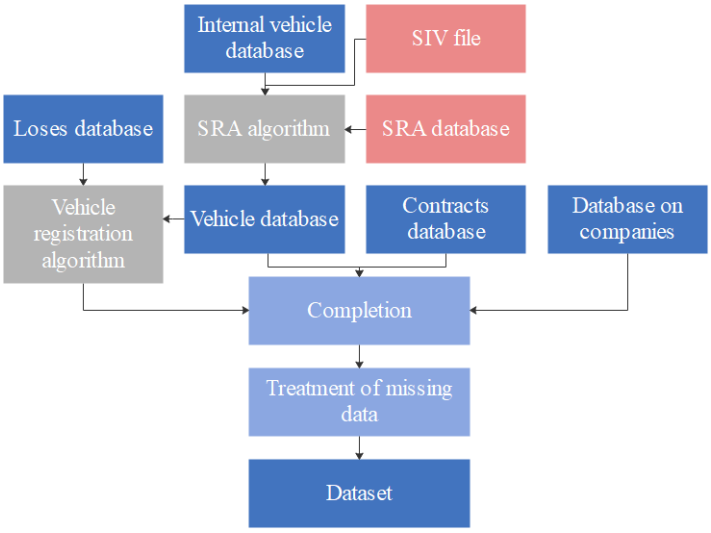


FIGURE 3 – Mapping of treatments on the dataset

Modelling

Insurance is a service that provides a benefit upon the occurrence of an uncertain and random event. This risk is insured in exchange for the collection of an insurance premium. A premium is defined as the sum of the pure premium, expense and profit. It is the pure premium that we wish to estimate, i.e. the average claim amount as a function of tariff variables. The loss ratio is random, however, part of the risk can be explained by the available data, particularly on vehicles, contracts and companies. Finally, we would like to model the pure premium linked to the Third party liability cover.

Pricing is performed without any operational constraints and therefore with all available data. Over 80 of variables can be priced. Creating a model made up of all the variables will tend to overfit the data, i.e. follow a variability similar to the claims data. However, this kind of model turns out to be very inefficient on a new dataset, especially for GLM (generalized linear model) or GAM (generalized additive model) type models. Thus, a parsimonious selection and a smoothing step are performed on the variables before each adjustment using a specific method of a calibration tool. In fact, several GAMs are built to answer calibration questions. GAM and GLM models are commonly used in insurance.

Civil Liability (CL) can be linked to bodily injury (BI CL) or material damage (MD CL). It is possible to distinguish CL according to the damage, as the behavior of the two sub-risks is different. However, the BI CL has a limited loss ratio to be modeled separately. Moreover, it is generally associated with severe load claims which will be discarded later. The CL modeling presents a performance similar to the aggregation of the two models for a smaller number of explanatory variables. Thus, the subsequent adjustment will be built on the loss ratio of the Third party liability regardless of the related damage.

A GAM or a GLM are parametric models using an exponential law to fit the data. The laws chosen that seem to fit the data best are, for the frequency and average cost model, the Poisson law and the Gamma law respectively. This is a collective model, i.e. the estimation of the pure premium is performed from the aggregation of a frequency and average cost model allowing a better description of the risk.

Finally, in the pre-calibration step, the charge is reprocessed. The opening costs, the recourse or the claims whose the cost is linked to the IRSA and IRCA conventions that allow the insured to be compensated more quickly, bias the model. Indeed, their cost does not correspond to the real amount compensated or it is negative. These claims are therefore excluded from the model and are subject to a lump-sum estimate of the pure premium. At the same time, claims with a severe burden, i.e. exceeding a relatively high amount, are pooled on a pro-rata basis of the premium of the claims premium (excluding flat-rate and opening costs) because of their atypical distribution. The threshold for severe claims was determined by the Extreme-value theory and more specifically by comparing the results of the Hill, Pickands and DEdH estimators. A threshold of 18,000€ will be used.

In order to improve the performance of the GLM, a deep variable selection step is produced after clustering the modalities of qualitative variables and after categorising the quantitative variables. The re-categorisation of the categorical variables is performed using five methods :

— Wald test judging the significance of the model coefficients,
— AIC or BIC, two penalisation criteria, to compare models,
— Lasso type penalization by a coefficients contraction effect,
— zone classification allowing to classify the geographic information by projection on a map,
— expert judgment, the modalities are grouped for their common operational meaning.

The last method is also used in the other four in order to maintain, beyond consistency at the risk level, coherence at the level of the operational significance of the grouping.

Finally, the quantitative variables are processed using a machine learning method : a decision tree known for its classification power. For a better robustness, it is the smoothed effect of the variable by functions called splines which is put at the input of the tree.

Categorisation makes it possible to create tariff boxes, to simplify the tariff and to make it easier to interpret. In insurance, working by tariff box is generally preferred and does not imply any loss of performance.

Subsequently, it is mainly the restated variables that were selected, proving the significance of the grouping methods. In addition, these methods made it possible to highlight certain variables which were strongly penalized by the selection linked to the large number of modalities.

New variables were retained in the model, in particular information on vehicles from the SRA database. One of the objectives of the study is to know the inherent performance gain of the SRA variables. Thus, a deep method on the effect of vehicles is exploited, i.e. the residuals of a GLM model in which the SRA variables have been voluntarily omitted. This method, called vehicle classification, makes it possible to summarise all the vehicle information in a single variable by means of a decision tree. Three different vehicles classifications are built. Each one deals with the vehicle effect in its own way before entering it into the tree. The first vehicle classification uses the raw effect, the second the vehicle effect removed from the data deemed not credible based on the definition of a threshold from credibility theory (Bühlmann-Straub model), and the last the smoothed effect. The preamble to obtain a smoothed effect is to spatialise the individuals at the SRA level by a Multiple Component Analysis (MCA). The projection is made from the correlations of the vehicle variables between them. Then, on this spatialization, the neighbourhood of each individual is defined by Delaunay triangulation (spatial geometry) to create a map of neighbours. In this way, the vehicle effect is smoothed from the Bühlmann-Straub model making the information credible using the neighbour map.

However, the performance gain of vehicle classifications is low, probably related to the application of clustering methods on the GLM. Moreover, it is more difficult to interpret the output of the trees from an operational point of view than interpreting trends on the coefficients predicted by the GLM. Thus, none of the vehicles are retained.

Throughout the study, the only method used to price risk is a GLM. Other learning methods exist. It is therefore interesting to compare the selected GLM model with machine learning models such as XGBoost or Random Forest. The best model in terms of performance criteria is the XGBoost. On the other hand, these adjustments turn out to be difficult to interpret because of their construction on a vast tree structure, and difficult to implement operationally. The definition of tariff boxes is less clear.

The entire calibration of the pricing model can be summarized by the following mapping :

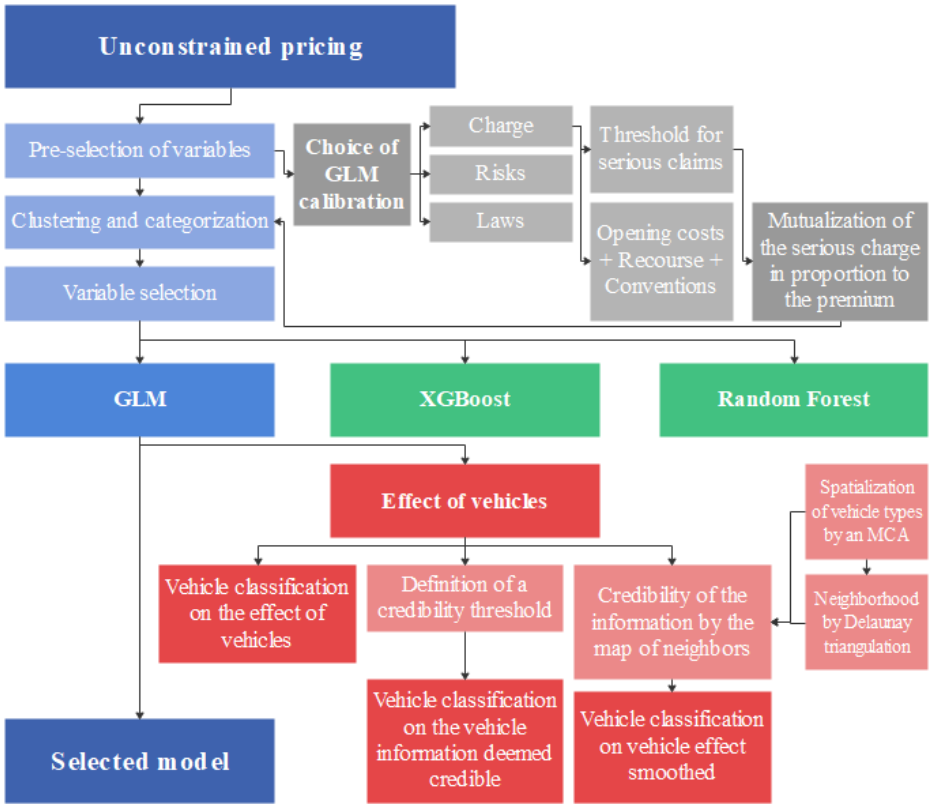


FIGURE 4 – Modelling map

Results

The tariff currently marketed in auto-entreprise responds to operational needs to facilitate subscription, but also to commercial strategies. The last major revision of the tariff took place in 2016 and introduced new data, especially vehicles with the SIV file. Since then, the tariff is updated every 3 years without any major change in the segmentation. The unconstrained model uses segmentation with new data, notably information on the financial health of companies, on the rate of referral of policyholders to partner garages by insurance intermediaries, and on vehicles (SRA database).

The current tariff is based on a similar treatment of the load. The goal is to juxtapose the two tariffs. A performance gain is effectively observed : increase of +5 points in the Gini of the frequency model and of +10 points on the average cost model.

When visualizing the price differences using a decision tree, the commercial variables bias the performance of the current price. In addition, the average premiums of firms engaged in freight or passenger transport seem to be overestimated. Indeed, these two classes were voluntarily increased during the last pricing update. Then, the new information on vehicles (bodywork, height, price at the origin of the vehicle) and the referral rate to partner garages turned out to be significant in the risk segmentation. These are not sufficiently captured by the constrained model.

Also, the tariff deviation reveals an average premium over the entire test dataset of -1%, i.e. the total prediction of the unconstrained model is lower than the prediction of the constrained one. This is partly due to the loss of claims information related to claims not captured by the vehicle database after the application of the registration algorithm.

Due to the numerous load adjustments, part of the study is devoted to reconciling the unconstrained model prediction to the gross load. The trends are similar, moreover, the performance gain is lesser than before, compared to the current price. This is due to the high presence of claims under IRSA or IRCA convention and opening costs. Indeed, these claims represent 64% of the frequency and 26% of the cost. Finally, the modeling without operational constraints shows the difficulty to obtain a significant performance gain on the Civil Liability cover because of the IRSA and IRCA conventions. Furthermore, by subtracting these lump sums, the variables of the SRA database, in particular, have strongly contributed to the improvement of the model's performance. In terms of product profitability, there is therefore no point in adopting the unconstrained model for pricing new business, but it can be used to mark up contracts on renewals.

Future improvements will first aim at reconsidering the study on the vehicle. The idea of creating a vehicle after having defined a smoothing effect on the information on the vehicles turns out to be too inefficient. Indeed, the credibility is applied on too few neighbors. An idea to increase the neighborhood would be to exploit the Euclidean distance, instead of using the Delaunay triangulation. Finally, the performance gain on the modelled Civil Liability premium thanks to the new information, encourages us to repeat the study on the Damage cover. The Damage guarantee is the second most important guarantee in terms of premium in the portfolio. It covers material damage to the insured vehicle. This guarantee is deprived of conventions which makes it an ideal candidate to test the new data available. In addition, the SRA base is also composed of category 6 vehicles, the two-wheeled vehicles plus quads. Thus, a study on category 6 vehicles can be carried out.

Keywords : Car insurance, third party liability, unconstrained pricing, database management, severity threshold, Extreme-value theory, generalized linear model, segmentation, vehicle classification, machine learning, clustering method.

Remerciements

Je tiens tout d'abord à remercier mon tuteur en entreprise, Brice Decaux, pour son accompagnement et ses conseils pendant toute la durée de l'alternance. Il a su rendre cette expérience enrichissante et agréable. Je souhaiterais également remercier les deux managers qui ont participé au suivi de ce mémoire, Gérald Lucas et Abdlekrim Rebadj, pour leur expertise et leurs encouragements.

Par ailleurs, je veux également remercier la responsable de l'équipe Actuariat Produits IARD Entreprise, Véronique Marpillat, et tous les membres de l'équipe pour m'avoir accueillie. Je leur suis reconnaissant pour le temps conséquent qu'ils m'ont consacré. Leur sympathie, leur compétence et leur coopération professionnelle m'ont permis de mener cette étude dans les meilleures conditions. Je suis très fier de les avoir eus comme collègues.

Je remercie les professeurs et le responsable de la formation en actuariat de Strasbourg, notamment Jean Bérard, Areski Cousin et Laurent Gardes. Ils ont assuré la partie théorique de ma formation. Ainsi, ils m'ont apporté une aide technique précieuse.

Enfin, je tiens à témoigner ma reconnaissance à l'ensemble des personnes qui ont participé à la relecture de ce mémoire et qui sont un réel soutien permanent dans ma vie. Il s'agit de mes proches, mes parents et mes amis.

Table des matières

Introduction

L'assurance Incendie, Accidents et Risques Divers, appelé IARD, est un service qui fournit une prestation lors de la survenance d'un sinistre afin de couvrir les biens de l'assuré, ainsi que les conséquences engageant la responsabilité de l'assuré. Par définition, en assurance, la survenance d'un sinistre est un événement incertain et aléatoire, appelé aussi risque. En contrepartie de cette couverture, l'assuré paye une prime. La détermination du montant de ces primes est rendue complexe du fait de la nature même de l'activité d'assurance et de son cycle de production inversé. En effet, la tarification des contrats est réalisée avant même de connaître la survenance d'un sinistre et son coût de dédommagement le cas échéant.

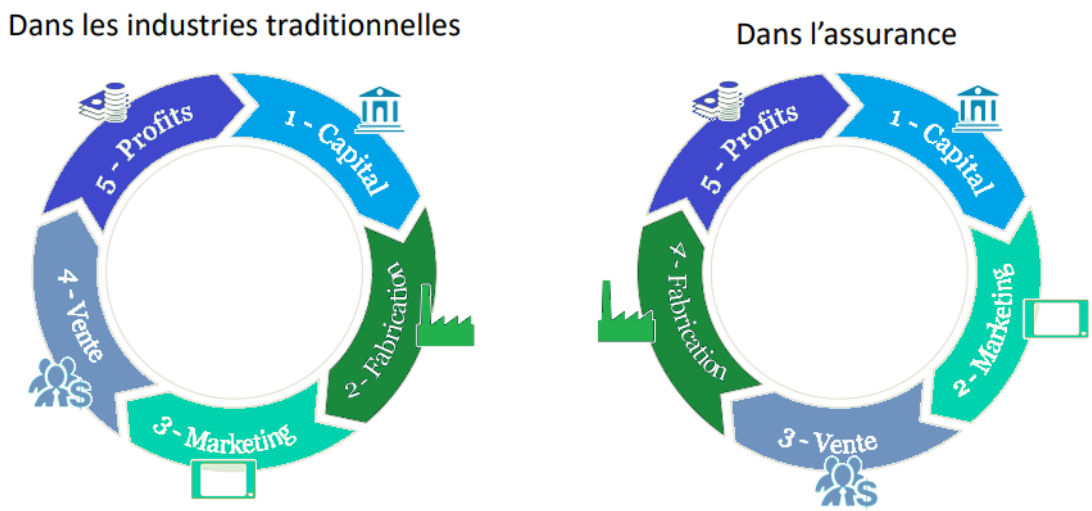


FIGURE 5 – Schémas du cycle de production

L'intérêt d'adhérer à une assurance est la mutualisation du risque, c'est-à-dire un principe d'indemnisation des préjudices d'un groupe homogène de personnes à partir des primes collectées pendant une période d'exposition au risque. C'est l'application directe de la loi des grands nombres [1]. Dans un portefeuille, l'objectif est de regrouper des risques indépendants dont la réalisation de l'un n'a pas d'impact sur la réalisation des autres et de permettre la compensation statistique. Ainsi, l'assuré préserve une pérennité financière, car il sera couvert en cas de survenance d'une catastrophe imprévue auquel il n'aurait peut-être pas eu les fonds. Un risque est assurable s'il est aléatoire, futur, licite, involontaire, réel et non rare ou peu fréquent [2].

Ce mémoire traitera la garantie responsabilité civile du produit *parcs dénommés* de la branche automobile entreprise. Cette garantie représente les charges et les primes les plus importantes du produit, expliquée en partie par l'obligation d'assurance. La réglementation de la responsabilité civile prend sa source d'un acte de bon sens et de solidarité. Elle a pour but d'indemniser tous les dommages causés totalement ou partiellement par l'assuré à un tiers. La part de responsabilité est déterminée à partir du constat amiable et/ou de toutes pièces fournies permettant de comprendre les circonstances du sinistre. L'assureur peut demander l'intervention d'un expert. La part de responsabilité impacte directement le

1. Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires réelles indépendantes et identiquement distribuées, la loi des grands nombres s'écrit,

$$\lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mathbb{E}[X].$$

2. ... exceptions faites sur certains risques dont les risques liés aux catastrophes naturelles.

montant de l'indemnisation.

De plus, la convention IRSA/IRCA [3], signée par la plupart des sociétés d'assurance en France, permet de faciliter et d'accélérer le traitement des dommages matériels/corporels en cas d'accident de la circulation en réduisant les frais de gestion. Pour bénéficier de cette convention, l'accident doit impliquer au moins deux véhicules terrestres à moteur, soumis à l'obligation d'assurance et assurés auprès de sociétés adhérentes. En cas d'accident responsable ou non responsable, faisant intervenir cette convention, l'assureur est en obligation d'indemniser l'assuré. Lorsque ce dernier n'est pas responsable, l'assureur effectue un recours auprès de(s) assureur(s) adverse(s). Cette indemnisation est forfaitaire ou réelle selon le plafond défini et la Responsabilité Civile (RC) mise en jeu. On distingue,

— la responsabilité civile liée à des dégâts matériels (RC MAT),
— la responsabilité civile liée à des dégâts corporels (RC CORP).

Pour la RC MAT, ce plafond est de l'ordre de 6 500€. Pour un recours inférieur à ce plafond, l'indemnisation est forfaitaire (1 678 € en 2021) et proportionnelle à la part de responsabilité. La convention IRSA/IRCA n'est pas opposable à l'assuré qui a le droit de refuser l'application des modalités de convention (cf. *article 1199 du Code Civil*).

Le tarif actuellement mis en place a été créé avec des contraintes informatiques sur les variables et les modalités de celles-ci. Il a été ajusté à partir d'une segmentation quasi-identique à la version tarifaire antérieure. Cette contrainte peut être problématique, car le risque assuré évolue. Une revue de ce tarif est alors nécessaire. Ainsi, l'objet de ce mémoire est de réaliser une refonte du tarif sans contrainte sur les variables tarifaires. Les données collectées sont à la fois internes et externes à l'entreprise. L'enjeu est d'optimiser la segmentation du risque afin d'assurer une meilleure rentabilité. Pour cela, une analyse des variables tarifaires à disposition sera détaillée. Cette analyse permettra de vérifier si la pertinence du modèle est encore d'actualité et d'en déduire si des modifications sont à apporter au détriment du coût informatique que cela impliquerait.

Étant donné que la RC n'est pas heurtée par une franchise qui viendrait rogner les sinistres, la modélisation de cette garantie sera en partie focalisée sur les recours et les coûts d'ouverture. En effet, ces derniers, ne représentant pas le coût réel du risque, peuvent apporter un biais aux résultats.

L'optimisation de la segmentation est assurée à partir d'une sélection pertinente de variables et d'un modèle type GLM (régression linéaire généralisée). Ensuite, des méthodes innovantes de traitements de variables telles que le véhiculier [4] ou des méthodes de regroupement seront utilisées à des fins d'amélioration de la performance de prédiction du modèle.

Afin de répondre à l'ensemble de ces problématiques, ce mémoire s'articulera en plusieurs parties. Dans un premier temps, une mise en contexte des enjeux sera détaillée pour mieux comprendre l'intérêt de ce projet. Dans un deuxième temps, une base de données sera créée à partir de différentes sources de données. Ensuite, les modélisations retenues pour la Responsabilité Civile automobile du produit *parc dénommé* seront présentées. Finalement, des méthodes et des modèles (véhiculier, méthodes de regroupement et *machine learning*) seront utilisés en complément, afin d'ouvrir le sujet sur de nouveaux outils.

3. La convention d'Indemnisation directe de l'assuré et de Recours entre Sociétés d'assurance Automobile a été créée en 1968 sous l'appellation IDA, Convention d'Indemnisation Directe des Assurés. IRCA est la convention concernant les dommages corporels, contrairement à la convention IRSA qui réglemente les dommages matériels.

4. Un véhiculier permet de créer une variable décrivant l'ensemble de l'information sur les véhicules.

Chapitre 1

Mise en contexte

Cette première partie du mémoire a pour vocation de présenter le contexte et les objectifs de l'étude. Dans un premier temps, elle présentera les fondamentaux de l'assurance automobile et les garanties. Puis, dans un second temps, elle décrira les enjeux de l'entité. Enfin, elle exposera les problématiques qui mèneront à la construction du sujet de ce mémoire.

1.1 Présentation du périmètre de l'étude

L'étude s'intéresse aux moyens de transports à quatre roues de moins de $3,5$ tonnes inclus dans les contrats composés de 5 à 50 véhicules. La prime liée à la responsabilité civile sur ce périmètre représente une part importante de l'assurance automobile entreprise. La modélisation d'un risque sur un champ d'action large permet de mieux appréhender les différences entre le tarif qui sera modélisé et le tarif actuel. Ainsi, il s'agira du périmètre étudié.

1.1.1 L'assurance automobile entreprise

L'assurance se définit comme le *"contrat par lequel l'assureur s'engage à indemniser l'assuré, moyennant une prime d'assurance, de certains risques ou sinistres éventuels"*. Dans le cadre réglementaire d'un contrat, l'assurance limite les désagréments financiers ou matériels à la suite de la survenance d'un sinistre. Il s'agit d'une protection contre l'aléa, qui peut entraîner la ruine de l'assuré. La pérennité d'une société d'assurance est consolidée par la souscription d'une multitude de contrats, dans l'hypothèse que les assurés ne soient pas sinistrés en même temps. Ainsi, le concept de mutualisation des risques découle du fait que les uns participent à l'indemnisation des sinistres des autres.

Les objets sont aussi bien assurables que les personnes. Du point de vue d'une entreprise ayant un parc de véhicules, c'est le dirigeant qui souscrira à un contrat d'assurance afin d'assurer son parc. Quelle que soit l'activité de son entreprise (messagerie, transports de marchandises ...), l'assurance automobile sur les véhicules à moteur est obligatoire depuis la loi du 27 février 1958. L'assurance minimale automobile indemnise l'ensemble des dommages causés à des tiers. Il s'agit de l'assurance auto Responsabilité Civile (RC) et doit couvrir les dommages matériels (RC MAT) et les dommages corporels (RC CORP).

Les bénéficiaires concernés par le contrat d'assurance entreprise parc sont les conducteurs des véhicules et le souscripteur. Ce sont des contrats fractionnés annuellement, semestriellement, trimestriellement ou mensuellement. Le contrat a une durée limitée d'un an et prend effet à partir des jours et heures indiqués dans les conditions particulières. Celui-ci est résiliable par le souscripteur selon plusieurs raisons décrites dans les conditions générales.

Le souscripteur doit à la souscription du contrat, fournir l'état du parc automobile et répondre exactement aux questions posées par l'assureur, notamment dans le formulaire de déclaration du risque, sur les circonstances permettant à l'assureur d'apprécier le risque. En cours de contrat, il doit déclarer les circonstances nouvelles qui ont pour conséquence soit d'aggraver les risques, soit d'en créer de nouveaux, et rendent de ce fait inexactes ou caduques les réponses faites à l'assureur.

Cette déclaration doit être faite, par lettre recommandée avec accusé de réception ou par envoi électronique, dans un délai de 15 jours à partir du moment où le souscripteur a connaissance de ces circonstances.

L'omission ou la déclaration inexacte du souscripteur ou de l'un de ses représentants n'entraîne pas la nullité du contrat lorsque la mauvaise foi n'est pas établie (cf. article *L 113-9 du Code des assurances*). En revanche, cela peut entraîner soit à une augmentation de la cotisation, soit à une diminution des indemnités, soit à la résiliation du contrat.

L'assureur verse les prestations à partir du moment où l'assuré réclame et que le fait est dommageable, garanti et justifié. Selon les cas, un expert peut être désigné par l'assureur afin de constater et d'évaluer les dommages subis par le véhicule. En application de l'article *L 211-5-1 du Code des assurances*, l'assuré a la faculté de choisir le réparateur automobile professionnel.

Le montant de réparation du véhicule ne peut pas dépasser la valeur économique. En cas d'aggravation du risque, l'assureur peut proposer d'augmenter la cotisation ou de résilier le contrat précédé par un remboursement. De même, en cas de diminution du risque, l'assuré a droit à une diminution du montant de la cotisation et peut dénoncer le contrat en cas de non-consentement de l'assureur.

L'automobile entreprise propose divers produits à destination des entreprises qui souhaiteraient faire assurer leur parc. Ces produits se distinguent selon le volume du parc et le poids des véhicules,

— **pour un parc de 1 et 4 véhicules de plus de** $3,5$ **tonnes**, ou des véhicules légers dont la classe socio-professionnelle de l'assuré est chef d'entreprise, le souscripteur sera dirigé vers un contrat dit *mono*, souscrit pour chaque véhicule. Un contrat *mono* est composé d'un véhicule et éventuellement d'une remorque. La tarification de ce type de produit entreprise se rapproche de la tarification des produits côté Particuliers et Professionnels (PP). Les informations communiquées utilisées pour la détermination de la prime pure sont spécifiques au conducteur et au véhicule,
— **pour un parc de 5 à 50 véhicules**, le produit proposé sera celui du *parc dénommé*. La tarification se définit selon les caractéristiques des véhicules, mais l'information sur les conducteurs n'est plus nécessaire. Cela simplifie la gestion du contrat d'assurance par le souscripteur de l'entreprise.
— **pour un parc de plus de 50 véhicules**, on parlera de *flottes ouvertes*. Cela concerne des grandes entreprises ou des collectivités locales. Pour une plus grande simplification de la gestion de l'assurance automobile, les détails des véhicules sont absents. Le tarif est construit à la maille contrat et non véhicule. L'introduction du Fichier des Véhicules Assurés (FVA) réglemente cependant les assureurs afin que l'ensemble des immatriculations des véhicules assurés soient recensées. Cette convention permet de lutter contre la non-assurance et de faciliter l'identification du véhicule en cas de délits. C'est d'ailleurs avec ce fichier que l'information véhicule est récupérée.
— **pour les professionnels de l'automobile, les garages et concessions**, un produit spécifique est proposé.

Dans le cadre de ce mémoire, nous traiterons du produit *parc dénommé*, et plus précisément de la tarification des véhicules de catégorie 1 inscrit dans les contrats d'assurance. En effet, le tarif est segmenté par catégorie de véhicules afin de mieux différencier les risques liés à chacune d'elles. Les véhicules sont regroupés en 7 catégories,

— *Catégorie 1* : Les véhicules légers, utilitaires, voiturettes et véhicules de sports,
— *Catégorie 2* : Les camions, tracteurs, remorque de plus de $3,5$ tonnes et les semi-remorques,
— *Catégorie 3* : Les autobus et autocars,
— *Catégorie 4* : Les engins agricoles,
— *Catégorie 5* : Les engins de chantier,
— *Catégorie 6* : Les deux-roues (scooter, cyclomoteur, moto et quad),
— *Catégorie 7* : Les remorques de moins de $3,5$ tonnes.

1.1.2 Les garanties proposées du produit *parc dénommé*

Les garanties en automobile de l'entreprise s'adaptent à chaque véhicule de la flotte du souscripteur, ainsi qu'au type d'activité du souscripteur. Les agents ou les courtiers d'assurance proposent aux souscripteurs par défaut une liste des garanties optimales, c'est-à-dire une liste de garanties qui reflète les risques réellement portés sur le parc de véhicules.

De plus, les garanties sont regroupées en 5 catégories :

— les garanties sur la responsabilité civile forment un groupe de garanties minimales lors de la souscription dues à la légalisation de la garantie responsabilité civile automobile excepté dans de très rares cas,
— les garanties dommages au véhicule,
— les garanties annexes garantissent les matières transportées dans le véhicule (marchandises, effets et objets personnels ...),
— les garanties complémentaires ou diverses,
— les garanties d'assistance couvrent tous les besoins d'assistance à l'aide d'un service proposé par AXA Assistance.

Dans le tableau suivant, toutes les garanties induites par le produit *parc dénommé* sont décrites.

	Garanties	Description
Garantie minimale (Responsabilité Civile)	**Responsabilité civile automobile**	La responsabilité civile de la personne assurée satisfait l'obligation d'assurance prescrite par les articles *L 211-1* et *R 211-5 du Code des Assurances*.
	Responsabilité civile fonctionnement	Les conséquences pécuniaires de la responsabilité civile du souscripteur sont garanties pour les dommages causés aux tiers imputables à l'utilisation du véhicule assuré, lorsqu'il fonctionne comme outil pour le travail auquel il est normalement destiné.
	Responsabilité civile et les atteintes à l'environnement	Les conséquences pécuniaires de la responsabilité civile incombant à l'assuré des dommages résultant d'atteintes à l'environnement accidentelles sont garanties, étendues à la prise en charge du préjudice écologique et à la responsabilité environnementale.
	Mise en œuvre de la garantie responsabilité civile et recours	Défense, Recours de l'assuré en inclusion de la responsabilité civile et avance sur recours de l'assuré pour les dommages matériels du véhicule assuré.
Garantie dommage au véhicule	**Dommages tous accidents**	Les dommages matériels subis par le véhicule assuré sont couverts. La conduite sous l'empire de substances stupéfiantes et/ou d'un état alcoolique est néanmoins une exclusion de cette garantie.
	Dommages par collision	Ceci porte sur les dommages matériels subis par le véhicule assuré à la suite d'une collision avec un tiers identifié. Contrairement aux dommages tous accidents, cette garantie n'assure pas contre les actes de vandalisme.
	Incendie, explosion, attentats, grêle et tempêtes	Les dommages matériels directs garanties résultent des événements suivants : incendie, explosion, détérioration de l'équipement électrique par suite d'incendie ou d'excès de chaleur sans embrasement, foudre, grêle, tempêtes.
	Vol	Cette garantie concerne le vol ou la tentative de vol du véhicule ou de ses éléments.
	Bris de glace	Le bris accidentel et fortuit de glaces est indemnisé.
	Catastrophes naturelles	Elle garantit à l'assuré la réparation pécuniaire des dommages matériels directs non-assurables à l'ensemble des biens garantis par le contrat ayant eu pour cause déterminante l'intensité anormale d'un agent naturel (garantie légale, application des articles *L 125.1* et *L 125.2 du Code des assurances*).
	Indemnisation en valeur conventionnelle	Concerne un véhicule acheté neuf ou pendant les 12 mois suivant la première mise en circulation. L'indemnité sera égale à la valeur d'achat figurant sur la facture initiale d'achat du véhicule.
Garanties annexes	**Garantie des effets et objets personnels et garantie appareils radio et assimilés**	Rembourse des effets et objets personnels transportés dans le véhicule et les appareils électroniques intégrés. Elle est garantie dès la souscription de l'une des garanties suivantes : Incendie, Vol, Dommages.
	Transport de marchandises pour son propre compte	Indemnise des marchandises et du matériel transporté privé.
	Transport de marchandises agricoles	Elle garantit du matériel, marchandises, animaux et outillages transportés par le véhicule agricole lors de la rupture d'essieu ou d'un dommage garanti.
	L'absorption de corps étrangers	Elle garantit des véhicules agricoles contre tout dommage accidentel résultant directement de la pénétration de corps étrangers.
Garanties complémentaires	**La Sécurité du conducteur**	Elle garantit les dommages corporels subis par le conducteur responsable à la suite d'un accident de la circulation routière ou par les ayants droit en cas de décès du conducteur, déduction faite des indemnités versées par les tiers payeurs (dépenses de santé, les pertes de revenus professionnels, pertes de revenus, ...).
	Les pertes financières	Les garanties Incendie, Vol, Dommages, s'appliquent aux pertes financières. Cette garantie participe à l'indemnité de résiliation en cas de perte ou destruction totale du véhicule.
	Les peintures publicitaires ou décoratives	Elle garantit les dommages subit aux peintures et aux accessoires publicitaires du véhicule assuré. Les garanties Incendie, Vol, Dommages, s'appliquent aux accessoires et aux peintures publicitaires.
Garanties assistances	**L'assistance aux personnes ou assistance médicale**	Verse une prestation d'assistance à la suite d'une atteinte corporelle ou de décès de l'assuré survenu lors d'un déplacement avec le véhicule (rapatriement, frais de séjours, frais d'avocats, frais médicaux).
	L'assistance au véhicule	Assiste le remorquage ou le dépannage du véhicule en cas de sinistre. S'ajoute aussi une éventuelle indemnisation des frais de séjour, de transports ou de remplacement du véhicule en cas d'immobilisation.

Ce mémoire se concentrera sur la garantie responsabilité civile représentant l'assurance minimale de l'automobile en France. Tous dommages causés sur autrui induisant la responsabilité civile d'une personne, physique ou morale conduisant un véhicule terrestre à moteur, doivent être couverts. Toutefois, les dégâts causés par un véhicule volé ne sont pas remboursés. De plus, les dommages causés intentionnellement par l'assuré, le transport de matières dangereuses, la guerre ou encore la conduite en été d'ivresse, la présence de substance illicite ... sont exclus.

Il n'y a pas de franchise pour la responsabilité civile tant qu'elle concerne un dommage matériel ou corporel. Une limitation d'indemnisation peut être fixée par l'assureur à hauteur minimum de 1,29M€ [1] en 2021. Chez AXA, cette limite était de 10M€ par sinistre et de 7,6M€ par véhicule en 2019.

1.2 Enjeux de l'entité

AXA IARD Entreprises (EN) est leader sur le marché des assurances entreprises. Elle comptabilise 15% des parts de marché, cela représente environ 2,9 milliards de chiffres d'affaires (CA) à vision fin 2019. Le groupe AXA s'est focalisé ces dernières années sur le développement de son entité entreprise.

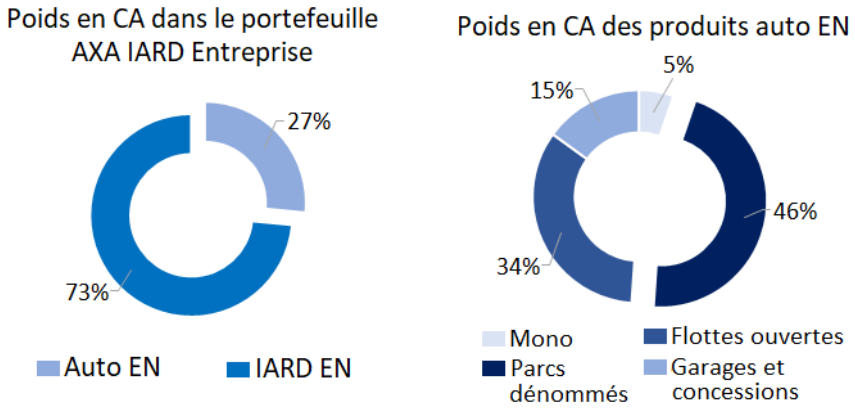


FIGURE 1.1 – Poids en CA du périmètre d'étude

L'assurance automobile occupe une place importante dans le portefeuille d'AXA IARD EN, ainsi que les *parcs dénommés*. En effet, le chiffre d'affaires des *parcs dénommés* représente environ 12% du chiffre d'affaire total EN. Cette répartition permet de mieux comprendre le positionnement du périmètre d'étude.

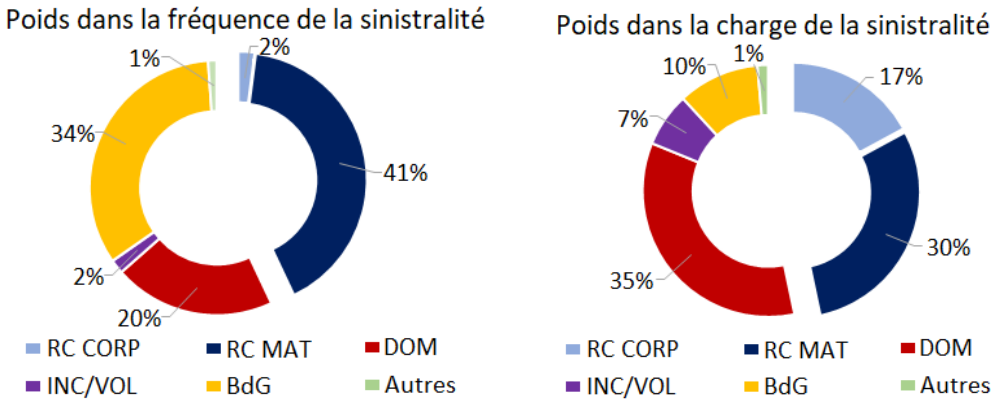


FIGURE 1.2 – Poids en CA des garanties périmètre d'étude

Pour les contrats en cours à fin 2019, la garantie RC automobile du *parc dénommé* représente en ordre de grandeur 43 000 contrats dont 430 000 véhicules de catégorie 1 assurés dans le cadre entreprise. Les véhicules de catégorie 1 en 2019 représentent 60% en nombre et 50% en prime. Les primes acquises quant à elles s'élèvent à 140 millions d'euros toutes catégories confondues en 2019. La prime acquise est la prime annuelle rapportée à la durée d'exposition au risque du contrat et à une année civile. Ainsi, pour un contrat dont la prise d'effet est au 1er juillet de l'année N et la date de résiliation est au 30 juin

1. Ce montant est révisable tous les 5 ans par l'État en fonction de l'inflation.

de l'année $N+1$, la prime acquise de l'année N sera égale à $\frac{1}{2} \times$ *prime émise*, où la *prime émise* couvre jusqu'à l'échéance du contrat.

A vision fin 2019, la fréquence de sinistralité de la RC CORP s'élève à $0,5\%$ pour un coût moyen de 13 000€. Quant à la RC DOM la fréquence de sinistralité est de 11% pour un coût moyen de 1 300€. Les deux sous-garanties RC CORP et RC MAT semblent avoir une sinistralité différente.

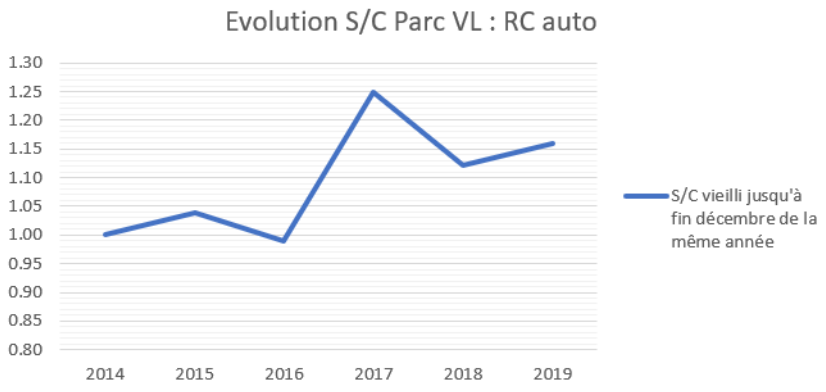


FIGURE 1.3 – Évolution S/C parc VL en RC

L'évolution du ratio S/C se dégrade. Les parcs VL correspondent aux contrats composés de véhicules de moins de $3,5$ tonnes (4 roues, 2 roues et engins agricoles). Ces contrats sont constitués majoritairement des véhicules de catégorie 1 (4 roues). Un réajustement de la prime a été donc réalisé en 2021 afin d'assurer la pérennité financière de l'entreprise.

À propos de la sinistralité, le seuil des sinistres graves ($\leqslant$ 30 000€) est fixe quel que soit le produit automobile pris en considération afin de faciliter leur pilotage. Il est toutefois révisable annuellement.

De plus, parmi ces sinistres, il existe des sinistres dits atypiques, c'est-à-dire dont la charge est supérieure à $2M$€. Les charges graves et les charges atypiques feront l'effet de retraitements, car pour un faible nombre, ils représentent une part importante de la charge ce qui peut biaiser les résultats.

Les chiffres précédemment inscrits, sont à vision fin d'année. En pratique, les charges évoluent même en cas de clôture du contrat, cela correspond à des sinistres qui n'ont pas été reportés alors qu'ils se sont déroulés pendant la période de couverture du contrat, ou des sinistres qui n'ont pas été clôturés et dont la charge évolue. Il sera donc nécessaire de faire vieillir les sinistres. Un vieillissement de 12 mois suffirait pour avoir une idée de la charge ultime, c'est-à-dire une estimation de la charge de sinistre une fois clôturée. Cependant, pour la RC CORP, ce délai d'un an ne stabilise pas suffisamment l'évolution des sinistres graves. En général, cela concerne les sinistres graves qui seront, lors de l'étude, retraités.

Un délai de 12 mois permet de concilier un historique de sinistralité récent et une vision de la sinistralité suffisamment vieillie. Des méthodes de calcul de la charge ultime existent, par exemple des méthodes utilisées en provisionnement, mais ne feront pas l'effet d'une étude approfondie pour des raisons qui seront expliquées par la suite.

De plus, les chiffres sur la sinistralité prennent en compte des montants forfaitaires (forfait IDA et coûts d'ouverture) qui devront également faire l'effet de retraitement ne reflétant pas la réalité du risque encouru. Ces charges sont liées à la convention IRSA/IRCA [2] qui permet aux assurés d'être indemnisés plus rapidement par leurs assureurs respectifs.

Dans le contexte de la crise sanitaire liée au COVID-19, la sinistralité a baissé en nombre et en charge de sinistre sur le parc auto. L'année 2020 ne sera pas prise en compte par la suite pour les raisons que cette année est atypique au niveau de la sinistralité et par un souci de vieillissement des charges des sinistres.

2. La convention d'Indemnisation directe de l'assuré et de Recours entre Sociétés d'assurance Automobile a été créée en 1968 sous l'appellation IDA, Convention d'Indemnisation Directe des Assurés. IRCA est la convention concernant les dommages corporels, contrairement à la convention IRSA qui réglemente les dommages matériels.

1.3 Objectifs de l'étude

Le tarif actuel est soumis à des contraintes informatiques. Le sujet de ce mémoire repose sur la construction d'un modèle sans contrainte s'appuyant sur l'ensemble des informations disponibles. De nouvelles données, notamment véhicules, sont introduites. Il s'agit de comparer la performance des modèles et de juger de l'importance de certaines variables décrites par la suite.

1.3.1 Mise en place de la problématique

La refonte du tarif du *parc dénommé* a été produite en 2021. Cette actualisation de tarif avait pour objectif principal de mettre à jour les coefficients associés à chaque modalité de variables tarifaires et d'ajouter une majoration. La rentabilité du produit était en déclin et ne répondait plus aux objectifs fixés. Un réajustement était nécessaire. Une analyse des segments NAF TPM<3T5 [3] et TPV<3T5 [4] a révélé une dérive d'environ +30% sur 5 ans, de 2016 à 2021. De plus, l'écart entre les objectifs de rabais de 2016 et l'observé 2020 implique un réajustement global de +10%, à partir de la constante du tarif.

La refonte du tarif a introduit quelques nouveautés sur la gestion des variables, notamment sur la variable *veh_energie* [5] et le zonier dans un but d'amélioration de la segmentation. En revanche, les actions pour améliorer le tarif sont limitées. Des contraintes informatiques empêchent l'introduction de nouvelles variables ou de nouvelles modalités. La modification d'une variable implique un coût qui peut dépasser le budget initial. D'ailleurs, le traitement du zonier par garantie et par catégorie a posé beaucoup de difficultés d'implémentation informatique. Ces limites proviennent d'outils dont les dépendances sont nombreuses et dont la construction repose sur des variables prédéfinies. L'introduction de nouvelles variables n'a pas été avisée. Ainsi, il est essentiel de réaliser une étude complète, sans contrainte opérationnelle, afin de vérifier que les variables sélectionnées ont toujours leur sens dans la tarification ou qu'elles soient toujours autant significatives.

Il y a trois différents tarifs. Les affaires nouvelles sont tarifées avec la dernière version tarifaire contrairement aux véhicules s'ajoutant à des contrats renouvelés où c'est la version tarifaire fixée lors de la date d'effet du contrat qui est utilisée. Cela permet de tarifer deux véhicules d'un même parc avec un même tarif afin d'éviter d'entrainer l'incompréhension de la part du client. Au moment du renouvellement, une majoration est appliquée dépendante du risque lié au contrat jugé par l'intermédiaire d'assurance et d'un ratio. Ce dernier permet d'inflater le tarif et de prendre en compte d'éventuels changements de segmentation du risque. Actuellement, ce ratio est le rapport entre la charge des sinistres et les primes. Dans un futur proche, il pourrait devenir le rapport entre la prime pure modélisée et la prime annuelle, appelé aussi ELR (*Economic Loss Ratio*). Un taux global va être décidé, pour piloter le produit ou la branche, et ensuite il sera distribué selon le niveau d'écart entre le tarif des affaires nouvelles contraint et le tarif sans contrainte. Si l'ELR est haut alors pas assez de primes n'a été capté par rapport au risque, il faut donc plus majorer qu'un ELR bas.

Un changement radical de segmentation a pour conséquence un coût informatique pour la tarification des affaires nouvelles important. Ce coût est encore plus important en implémentant un tarif avec une nouvelle segmentation dans l'outil de tarification des anciens contrats à cause d'outils dont les dépendances sont nombreuses. Les affaires nouvelles représentent plus de 10% des contrats, l'implémentation d'un nouveau tarif est donc lente et progressive. La rentabilité de la mise en place d'un nouveau tarif n'est pas immédiate. C'est pourquoi, les résultats doivent présenter une réelle différence de performance entre les deux tarifs sans et avec contrainte pour que le modèle sans contrainte soit mis en place.

L'idée du tarif sans contrainte est de mieux segmenter le risque. Un tel tarif peut permettre d'appliquer des majorations plus proches de la réalité lors du renouvellement sur l'ensemble du portefeuille sans coût informatique supplémentaire. Ainsi, un gain de performance faible du tarif sans contrainte est accepté.

Le tarif actuel pour la RC catégorie 1 du *parc dénommé* se présente sous la forme suivante,

3. TPM<3T5 concerne tous les contrats dont l'activité principale de l'entreprise prospecte est le transport de marchandises et dont la majorité des véhicules assurés sont inférieurs à 3.5 tonnes.

4. TPV<3T5 concerne tous les contrats dont l'activité principale de l'entreprise prospecte est le transport de voyageurs. De plus, les véhicules assurés sont inférieurs à 3.5 tonnes.

5. Source d'énergie du véhicule (essence, gasoil, électrique *dots*).

$PP_{RC} =$ prime RC (constante) $\times$
coefficient de la zone du risque $\times$
coefficient de l'ancienneté du véhicule $\times$
coefficient sur l'activité de l'entreprise (NAF) $\times$
coefficient sur la carrosserie du véhicule $\times$
coefficient sur le genre du véhicule $\times$
coefficient sur la puissance fiscale $\times$
coefficient de la marque du véhicule $\times$
coefficient en cas de véhicule appartenant au dirigeant [6] $\times$
coefficient pour les véhicules transportant des matières dangereuses [7] $\times$
coefficient sur l'énergie [8] à l'origine de la motorisation du véhicule $\times$
coefficient sur le mode d'achat [9] effectué pour acquérir le véhicule $\times$
coefficient top ME [10]

La prime pure est calculée sans les informations sur les antécédents sinistres et les sinistres atypiques dont la charge est supérieure à $2M€$. En effet, les antécédents sinistres de la RC et de la garantie dommage (DOM) ne sont considérés qu'après la modélisation à la maille contrat, c'est-à-dire quelle que soit la catégorie de véhicules considérés.

1.3.2 Les points d'attention

La tarification sans contrainte implique l'introduction de multiples variables dont plusieurs d'entre elles ont des résultats attendus. Il s'agit de variables qui ont déjà fait leur preuve sur d'autres produits auto ou d'informations que l'on souhaite étudier. Il y a :

— les variables véhicules,
— l'information géographique,
— le taux d'orientation garage partenaire,
— les variables à la maille entreprise dont l'effectif, le chiffre d'affaires et la note financière,
— le kilométrage à l'ouverture du sinistre.

Variables véhicules : fichier SIV et base SRA

Par rapport, aux premières versions tarifaires, l'accès à des données externes est possible. En effet, le fichier SIV, la base SRA ou la base INSEE sont trois bases apportant des informations supplémentaires sur le risque. D'ailleurs, l'automatisation de l'intégration de ces bases générerait un coût côté informatique. Le fichier SIV et la base SRA détaillent les données véhicules et ainsi, peuvent permettre une segmentation du risque plus pertinente. Ces deux bases seront plus amplement décrites par la suite.

La base SRA regroupe toutes les informations véhicules selon le type de véhicule caractérisé par un code, nommé code SRA. L'introduction de multiples variables dans la modélisation peut avoir comme conséquence une sur-segmentation et ainsi peut sous-performer sur de nouvelles données. Plusieurs méthodes seront utilisées pour limiter le sur-apprentissage tel que la cross-validation, la sélection de variables, des méthodes de regroupement et enfin une méthode de classification par caractéristique véhicule, appelé véhiculier. Plusieurs pistes seront explorées afin d'améliorer encore la performance du modèle au sens de certains critères.

6. Un véhicule assuré dont la carte grise est au nom du dirigeant de l'entreprise, ou appartenant au souscripteur mais réservé au seul usage du dirigeant, bénéficie d'une rémunération. Le coefficient avantageux appliqué sur le véhicule dirigeant est donc simplement commercial.

7. La classification des matières dangereuses dépend de la législation régissant le transport de ces marchandises, appelée *arrêté TMD* du 29/05/2009. Cet arrêté répartit en 12 classes les matières dangereuses. L'article *R 211-11 du code des assurances* fixe une limite d'autorisation de transport des matières dangereuses à 500kg ou 600 litres.

8. Le coefficient appliqué sur la source d'énergie du véhicule distingue les véhicules électriques et les véhicules à essence.

9. Le mode de financement d'un véhicule peut être de plusieurs types : au comptant, par crédit, par crédit-bail, par location longue durée ou par d'autres modes d'achat.

10. Un véhicule ME est un véhicule de moins de $3,5$ tonnes à usage interne de l'entreprise utilisé pour des missions ponctuelles à des fins de fonctionnement interne ou de service.

Information géographique

Pour l'information géographique, retraité à l'aide d'une méthode de zonage (zonier), l'objectif serait d'expliquer à partir de bases externes, la sur ou sous sinistralité selon la zone géographique. Par exemple,

- *est-ce que les zones à risques sont caractérisées par des zones où les routes sont en mauvais état ?*
- *est-ce que la météo aurait un impact sur la sinistralité ?*
- *est-ce que ce zonier se limite à notre étude ou en est-il de même pour des données sinistres nationales ?*

Une meilleure compréhension du zonier pourrait nous inspirer de nouvelles façons de traiter l'information géographique.

Taux d'orientation garage partenaire

Ensuite, AXA a des partenariats avec plusieurs garages afin d'obtenir des prix intéressants sur les réparations. AXA incite donc les intermédiaires d'assurances, courtiers et agents, à orienter leur client vers ces garages. D'ailleurs, les agents ont tendance à plus orienter leur client vers les garages partenaires que les courtiers. Un taux d'orientation élevé est intéressant, car en supposant une constance dans le comportement des clients, cela impliquerait des charges moins élevées, ce qui revient à baisser la prime et à devenir plus concurrentiel. Une étude sera menée pour mesurer la véracité de cette variable dans le modèle. Sa présence peut être liée à un effet comportemental. Par exemple, les agents qui orientent le plus vers les garages partenaires sont peut-être suffisamment proches de leur client pour les accompagner et ainsi réaliser une campagne de prévention contre les accidents de la route.

Information entreprise

La base INSEE contient plusieurs informations sur les entreprises dont l'effectif, le chiffre d'affaires et la note financière. Connaître la santé financière d'un prospect est un élément important pour l'appréciation du risque. Elle permet d'évaluer les capacités de l'entreprise, à engager des actions de prévention, à entretenir correctement son outil de travail et à disposer d'une main d'œuvre motivée et qualifiée. Subsidiairement, elle nous donne une indication quant à la probabilité de défaillance d'une entreprise d'un secteur donné et ainsi de voir les primes impayées. Compte tenu de toutes les incertitudes évoquées, il n'est pas possible d'attendre de cette note un grand degré de précision. Une note basse doit être comprise comme un signal d'alerte nécessitant des investigations plus poussées. La règle transversale de souscription est la suivante :

- il convient d'avoir à l'esprit que les entreprises dont la note de santé financière affichée est inférieure à la cible de 10/20 doivent faire l'objet d'une attention particulière avant toute décision de souscription,
- les entreprises dont la note de santé financière affichée est strictement inférieure à 5/20 doit faire l'objet d'un refus de souscription.

La santé financière d'un prospect est un élément important pour l'appréciation du risque. C'est pourquoi, AXA fait appel à une société externe privée spécialisée. Cette société fournit un score en fonction des éléments propres à l'entreprise et au secteur d'activité dans laquelle elle évolue.

Kilométrage à l'ouverture du sinistre

Enfin, une nouvelle variable est introduite dans l'étude, il s'agit du nombre de kilomètres parcourus à l'ouverture d'un sinistre. Aucune information sur le nombre de kilomètres parcourus d'un véhicule du parc à la date d'effet du contrat ou à la clôture n'est disponible. Il n'est donc pas envisageable de quantifier la distance parcourue d'un véhicule depuis la date d'effet du contrat. Plus le kilométrage est élevé, plus l'exposition au risque est accrue. Cette donnée aurait pu être une pondération au même titre que la durée d'exposition dans la modélisation. Cependant, il est possible que cette information soit captée par d'autres variables telles que l'activité de l'entreprise ou l'ancienneté du véhicule. Le kilométrage à l'ouverture du sinistre est une donnée sinistre, donc non-tarifaire. Par contre, une étude sur la corrélation de cette variable avec les autres sera réalisée afin de vérifier que l'information soit captée. Une seconde étude, par rapport à la variable réponse (la charge), sera approfondie afin de constater la significativité de la variable.

Chapitre 2

Présentation de la base de données

La base de données est un support à la tarification et aux différents travaux réalisés. Toutes les études qui suivent reposent sur cette base, il est donc essentiel de créer une base complète et fiable. Ainsi, une première étape sera d'expliquer la construction de cette base et de prendre connaissance des informations sous-jacentes. Une seconde étape consistera à fiabiliser nos données en contrôlant la qualité des données. Enfin, une dernière étape proposera une introduction de différentes notions afin de comprendre ce qu'il sera modélisé par la suite. L'ensemble des traitements réalisés sur la base de données, est réalisé sous l'utilisation du langage **SAS**.

Les variables ne seront pas décrites de façon univariées dans cette partie à cause de leur nombre trop important. Seules les variables retenues par nos modèles seront évoqués (sous-section 3.3.4, page 72).

2.1 Description des bases de données brutes

Les données sur le produit *parc dénommé* sont réparties dans différentes bases de données. Un projet de rassemblement de toutes les bases de données en une seule, nommée *data lake*, est en cours. Pour l'instant, la séparation des données selon leur maille spécifique (véhicule, contrat, SIREN ...) permet d'utiliser un minimum d'espace de stockage.

Ces données sont collectées à partir d'un outil de souscription propre au produit, d'un outil de gestion de sinistre alimenté par les courtiers et de données externes.

2.1.1 Sources internes

Les bases de données hébergées par les serveurs d'AXA France sont sécurisées, classées et versionnées permettant d'accéder à de nombreuses informations dont seules les données concernant l'étude seront agrégées à une même maille, la maille véhicule × année de survenance. Dans un contrat *parc dénommé*, le nombre de véhicules est compris entre cinq et cinquante. Les véhicules sont assurés de façon individuelle. Ils n'ont pas les mêmes garanties associées et ils n'ont pas le même impact sur le risque. L'année de survenance, nous permettra d'ajouter l'ancienneté du véhicule et de crédibiliser l'information lors de la modélisation.

Bases contrats

Les contrats sont isolés sur trois années de vision, en cours entre 2017 et 2019. Cet intervalle de temps a été fixé afin d'obtenir au minimum un an de vieillissement sur les données, notamment les sinistres. Les antécédents aux contrats ne seront pas étudiés lors de ce mémoire, car les données passées sont prises en compte dans un coefficient qui a été défini hors modélisation. Les données liées à la base contrat sont vieillies jusqu'en avril 2021 afin d'acquérir les informations les plus récentes sur celles-ci.

L'objectif est d'obtenir toutes les informations qui auraient un lien avec la sinistralité (variables tarifaires). En pratique, plusieurs bases sont utilisées,

— une base composée des contrats en cours entre 2016 et 2020. Cette base est à la maille contrat × année de vision. Les informations d'un même contrat d'une année à l'autre sont identiques. Elle est composée d'informations telles que la position géographique de l'entreprise, le fractionnement de la prime, l'activité de l'entreprise .... Seuls les contrats en cours concerné par le produit *parc dénommé* et dont la date de résiliation est supérieure à 2017 et la date d'affaire nouvelle est inférieure à 2019 ont été conservés,
— une base composée d'indices distributeurs dont un toutes branches et un autre sur la branche auto. Elle est fusionnée avec la clé numéro de contrat × vision. Chez AXA, les distributeurs sont les

agents généraux ou les courtiers. À chaque distributeur, il est attribué une note ou indice allant de 1 à 5 qui est le reflet de leur participation et de leur sélection des clients quels que soient les risques couverts. L'indice distributeur est l'agrégat de chaque indice donné par branche dont l'indice auto que l'on choisit également de conserver,
— une seconde base donnant un libellé sur l'expérience du distributeur,
— une base concaténée de trois années de visions consécutives recensant le nombre de réparations dans un garage partenaire. AXA a des partenariats avec des garages partenaires aux prix négociés sur les réparations. L'idée est de regarder si toutes choses égales par ailleurs, le risque est meilleur côté agent qui oriente plus vers les garages partenaires. Le taux d'orientation vers les garages partenaires est le rapport entre le nombre de réparations dans les garages partenaires sur le nombre de réparations total. Il y a trop peu de données à la maille contrat, c'est pourquoi ce taux est construit à la maille distributeur.

Les informations des trois dernières bases sont à la maille distributeur.

Bases véhicules

La base véhicule communique des informations telles que la marque, la puissance fiscale, ... Chaque véhicule est associé à un numéro de contrat. En effet, ces véhicules sont pris en charge par l'assurance par la constitution d'un contrat liant le souscripteur à l'assureur. Il existe une base recensant tous les véhicules associés aux contrats conservés précédemment. Cette base est en fait la concaténation de plusieurs bases véhicules qui ont chacune une année de vision différente. La date d'entrée et la date de sortie du véhicule doivent être cohérentes avec l'année de vision, ainsi qu'avec la date de résiliation du contrat ou la date d'affaire nouvelle.

Deux informations sont construites à partir de ces dates d'entrée et de sortie : la durée d'exposition au risque, appelé le taux de présence, et l'ancienneté du véhicule faisant intervenir également la date de mise en circulation du véhicule. Un contrat débutant au 1er janvier N et se terminant au 1er juillet N a une durée d'exposition au risque de $0,5$, soit une demi-année.

Cette base contient également toutes les informations sur les garanties, les primes, les franchises, la valeur assurée du véhicule, ... L'immatriculation du véhicule sera la clé pour lier la base véhicule à d'autres bases, par exemple les bases sinistres.

L'étude est réalisée seulement sur les véhicules utilitaires légers (VUL de moins de $3,5$ tonnes), les véhicules légers (VL), les voiturettes (VT) et les véhicules de sport (VS), assurés en cas de dommages liés à la garantie responsabilité civile.

Bases sinistres

Les bases composées des variables tarifaires ont été définies. Les variables à prédire dans la partie modélisation sont les charges et le nombre de sinistres répertorié dans la base sinistre. Les données sinistres sont agrégées à la maille véhicule $\times$ année de survenance (l'année de survenance correspond exactement à l'année de vision). C'est ici, qu'un historique de la sinistralité peut être avancé, mais ne sera pas réalisé dans le cadre du mémoire, car, comme écrit précédemment, les antécédents sont pris en compte par un coefficient à la maille contrat hors modélisation.

Afin de contrôler la cohérence entre la base sinistre et la base véhicule fusionnée avec la base contrat, les informations de la base sinistre sont ajoutées à partir des clés années de vision, numéro de contrat et immatriculation du véhicule.

Par la suite, quatre différentes bases sinistres seront exploitées. Elles sont issues des différents outils de gestion, afin de compléter et de contrôler la qualité des données. La base sinistre de référence est, comme la base véhicule, une concaténation de bases chacune liée à une année de survenance. La période d'étude est de trois ans (2017 à fin 2020). Pour un vieillissement de la sinistralité, c'est la dernière date d'observation du sinistre qui est prise en compte. Ainsi, les sinistres qui ont eu lieu fin 2019 sont vieillis d'un an. La prestation peut être versée tardivement, car le sinistre peut avoir évolué ou celui-ci pouvait ne pas avoir encore été déclaré.

Plusieurs informations ont été retenues en plus de la charge, du nombre de sinistres et de l'année de survenance telles que :

— la garantie impactée (seuls les sinistres liés à la RC sont intéressants dans le cadre de l'étude),

— la part de responsabilité de l'assuré pour les sinistres de type RC Automobile. Il y aura une intervention différente dans les recours ce qui permettra d'expliquer la charge finale du sinistre,
— le kilométrage à l'ouverture du sinistre.

Les sinistres dont la charge est supérieure à 2 millions d'euros sont supprimés, car d'après la sous-section 2.3, page 39, ces charges dites atypiques sont prises en compte dans le coefficient PLR hors modélisation.

2.1.2 Sources externes

Les assureurs ont accès à des bases externes qui leur permettent de compléter les informations à leur disposition. En effet, afin de faciliter la souscription à un contrat d'assurance, les assureurs demandent seulement de compléter les informations utiles à la tarification. En revanche, cette limitation a pour conséquence de réduire les informations internes à disposition de l'assureur et ainsi, il devient difficile d'introduire de nouvelles variables dans la tarification.

Bases externes à la maille contrat

Plusieurs informations provenant de trois bases externes sont ajoutées à la base contrat.

Pour recueillir ces informations, une base à la maille SIRET de l'INSEE est fusionnée avec la base contrat sur le numéro de SIRET. La différence entre le numéro SIREN (Système d'Identification du Répertoire des Entreprises) et le numéro SIRET (Système d'Identification du Répertoire des Établissements) est que le numéro SIRET est une maille plus fine que le numéro SIREN, car il identifie chaque établissement d'une entreprise au lieu de l'entreprise seule. Pour un même numéro de SIREN, il peut y avoir plusieurs numéros SIRET, en revanche pour un même numéro SIRET, il n'y a qu'un seul numéro SIREN. La base contrat est donc complétée par des informations liées à l'activité de chaque établissement, filialesx de l'entreprise. Le chiffre d'affaires ou l'effectif de l'entreprise sont des données très variables d'une année sur l'autre, c'est pourquoi, la base SIRET est également fusionnée selon l'année de vision grâce à un historique. Une information supplémentaire intéressante est une note financière sur 20, communiquée par *creditsafe*, une entreprise privée.

Deux bases externes composées d'informations géographiques vont également venir alimenter la base contrat. Lors de la modélisation, on a voulu expliquer les zones à risques (Section 3.5, page 86) avec des données telles que l'état des routes, ou des données sur les accidents de la route recensés par le fichier BAAC. Ce sont des bases mises à disposition au grand public. Les données sont agrégées à la maille département.

Bases externes à la maille véhicule

À la maille véhicule, il existe deux bases fournies aux assureurs composées d'information sur le véhicule,

— le fichier SIV,
— la base SRA.

Le fichier SIV (Système d'Immatriculation des Véhicules) est une base de données mise en place par le ministère de l'Intérieur pour simplifier la gestion des documents relatifs aux démarches d'immatriculation. Un véhicule reçoit un numéro d'immatriculation qu'il conservera à vie, jusqu'à son exportation ou sa destruction. Ce fichier contient les informations inscrites sur la carte grise du véhicule hors donnée à caractère personnel pour les professionnels d'assurances. La durée de conservation d'une immatriculation à compter de la date de la destruction physique du véhicule est de 5 ans. Elle est régulièrement mise à jour, afin d'ajouter de nouveaux véhicules et d'apporter des éléments de correction sur certains véhicules déjà recensés (erreur de manipulation, de déclaration, ...).

Les informations du fichier SIV sont ajoutées à la base véhicule à partir de l'immatriculation du véhicule et de l'année de mise en circulation pour une meilleure cohérence.

SRA (Sécurité et Réparation Automobiles) est une association fondée en 1977, à laquelle toutes les entreprises d'assurance automobile en France sont adhérentes. Une des missions de la SRA est la création d'une base de données recensant toutes les caractéristiques techniques et commerciales des véhicules terrestres à moteur de 4, 3 et 2 roues de moins de $3,5$ tonnes, destinées au marché français et commercialisées par un constructeur ou un importateur officiel. Elle est mise à jour tous les mois en fonction des nouveaux

véhicules apparaissant sur le marché automobile français. L'objectif de cet organisme est de mettre en œuvre toute action qui contribue à limiter le nombre et le coût des sinistres, dans l'intérêt des assurés.

Le code d'identification SRA permet de récupérer les informations détaillées d'un véhicule. Celui-ci est composé de deux lettres pour identifier la marque du véhicule suivi de 5 chiffres pour reconnaître la version et le modèle du véhicule. Ainsi, à partir du code SRA, des données telles que l'énergie, le nombre de places, la longueur, le poids à vide sont accessibles. Ce code est la clé pour joindre cette base à la base véhicule.

2.1.3 Agrégation des bases

Jusqu'à la constitution de la base finale à la maille véhicule × vision, la structure de fusion de ces bases peut se synthétiser par le schéma suivant :

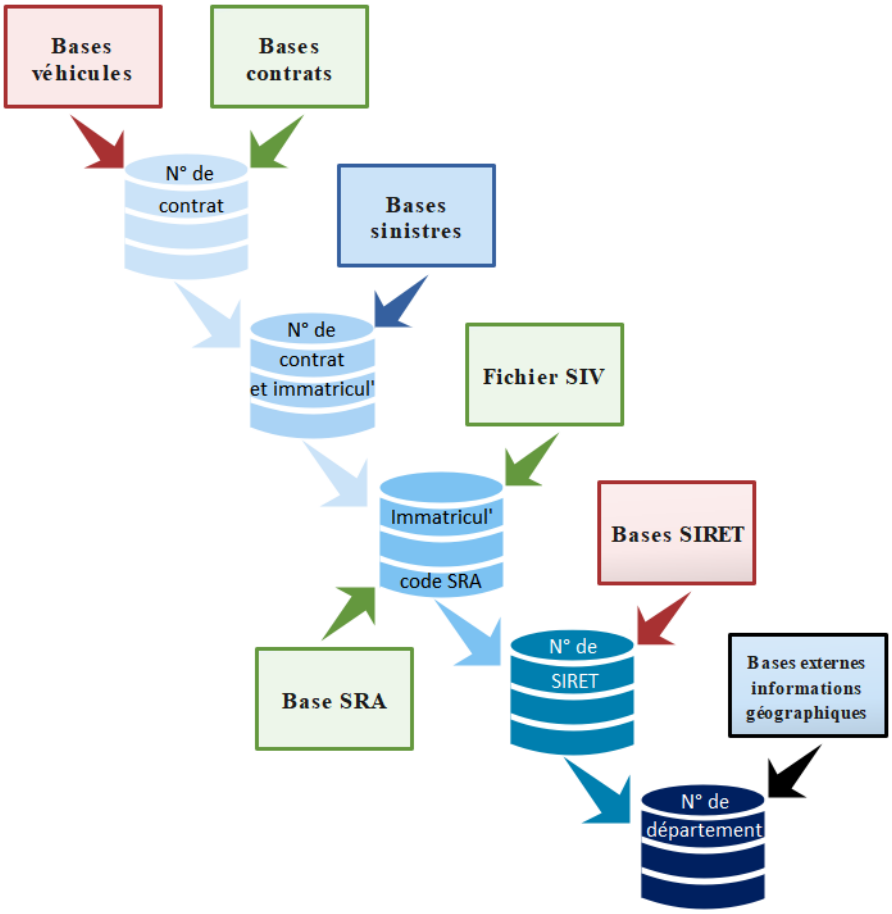


FIGURE 2.1 – Schéma d'agrégation des bases

2.2 Contrôle de qualité des données

La base de données est la principale matière utilisée par les algorithmes de tarification. Il s'agit donc de vérifier la cohérence et la qualité de l'ensemble des informations inscrites. De plus, une modélisation nécessite des informations les plus complètes possible. L'ensemble de ces traitements permettra de gagner en exhaustivité et en pertinence.

2.2.1 Traitement des données manquantes

Les données manquantes se traduisent par des valeurs non renseignées menant à un biais dans la modélisation. En effet, ne sachant pas *a priori* la réelle information cachée derrière la donnée manquante, par défaut, toutes les non-informations sont traitées comme une seule et même modalité. Plusieurs options sont possibles :

— ne pas traiter et conserver les valeurs manquantes ce qui revient à créer une modalité spécifique pour les variables catégorielles,
— supprimer ces observations, ce qui revient à supposer que l'ensemble des données manquantes a le même effet que le reste des données ou a un faible impact sur les résultats finaux. Cette méthode est en général déconseillée, car supprimer une fraction non-négligeable des données est problématique. De plus, même une faible fraction des données est susceptible de représenter une part essentielle de l'information qu'elle contient,
— remplacer les valeurs nulles par une estimation comme la médiane, la moyenne ou encore la modalité la plus représentée.
— utiliser d'autres bases ou d'autres variables qui cacheraient l'information manquante.

Compléter les données en cherchant l'information à partir d'autres variables ou d'autres bases reste la méthode la plus pertinente, mais ne concerne que très peu de données. Par exemple, la variable sur le NAF (Nomenclature d'Activités Française) est complétée par trois bases contrats non citées précédemment et une variable recensant l'activité de l'entreprise sous une autre maille à laquelle une table de correspondance est associée. De même, des variables véhicules sont redondantes et donc complétées entre la base SRA, la base véhicules et le fichier SIV (et d'autres bases véhicules non évoquées ici).

L'analyse des corrélations et l'influence des valeurs manquantes sur le modèle permettent une bonne utilisation des données manquantes restantes. Les données manquantes ne sont traitées qu'une fois la modélisation réalisée.

Ainsi, une première approche sur les données manquantes a été de les considérer comme une modalité à part entière pour chaque variable catégorielle. Lors du lissage, un effet de *grouping* est ajouté pour les variables catégorielles non-ordinales. En effet, le modèle peut attribuer aux données manquantes le même coefficient attribué à une autre modalité et ainsi les considère proches au sens de l'effet sur la variable réponse. Lors de la sélection du modèle optimal, cette approche permet de faire abstraction de l'influence des données manquantes sur la prédiction.

Les variables composées d'un nombre de données manquantes trop importantes ne seront pas conservées lors de la modélisation. La théorie de la crédibilité sera employée afin de minimiser l'impact des données manquantes sur la modélisation. Pour une exposition faible de la modalité *NA*, l'impact des valeurs non renseignées sur le tarif sera faible, voire nul.

Les données manquantes n'ont pas de sens dans une tarification. En effet, une fois le tarif mis en place, toutes les données tarifaires doivent être renseignés. En associant un coefficient à une modalité *NA*, l'intercept du modèle est réévalué. Il s'agit tout de même de regarder si les données manquantes cachent une information importante dans la tarification.

Chapitre 3, page 42, pour les méthodes de type GLM, seules des variables catégorielles ont été implémentées en entrée de l'algorithme.

Lors de la modélisation par des méthodes d'arbre de régression, nous ne nous limiterons plus à des variables catégorielles, car ce type de méthodes permet de créer directement des classes pour les variables quantitatives. Par conséquent, des tests sur un éventuel remplacement des données manquantes ont été réalisés. Par exemple, un remplacement par la médiane ou par la valeur moyenne de la classe ayant un coefficient équivalent, dans l'hypothèse où il existe une telle classe dans le GLM, a été envisagé.

Les algorithmes CART ont un système intégré fonctionnant par variable-substitut pour traiter les données manquantes. Il s'agit d'une méthode qui considère une seconde variable pour prédire l'effet d'une variable sur la donnée manquante. Cette approche sera décrite plus précisément par la suite.

D'autres méthodes ont été utilisées au cours de ce mémoire afin de compléter au mieux les données manquantes.

2.2.2 Cohérence et retraitement des données

La base de données a été explorée afin de remplacer ou de supprimer les anomalies. Plusieurs retraitements ont été effectués à l'aide de corrections, regroupements ou exclusions pour fiabiliser les données de l'étude. Voici quelques exemples d'ajustements et de nettoyages réalisés sur cette base :

— remplacement des anomalies par des données supposées plus fiables, car provenant de bases plus propres,
— suppression des anomalies,
— regroupement de modalités pour certaines variables catégorielles,

— discrétisation des variables quantitatives tarifaires par la création de quantiles,
— création de nouvelles variables comme l'ancienneté du véhicule, l'exposition, la charge RC, ...,
— suppression des informations redondantes (la marque du véhicule est renseignée à la fois dans le fichier SIV , base SRA et base véhicule interne).
— réconciliation de certaines variables comme le code NAF avec l'activité de l'entreprise,
— définition des bases les plus fiables.

Les anomalies détectées dans les bases de données sont généralement liées à des erreurs de saisies manuelles (ou à de mauvais renseignements). Il s'agit donc de contrôler la qualité des bases de données. Une mesure est transcrite à partir des écarts d'inscription entre les différentes bases de données. Les informations redondantes sont comparées en définissant une base de référence.

Les bases de données internes sont considérées comme fiables grâce à de multiples retraitements. Ce sont les bases de référence pour l'étude. Le fichier SIV et la base SRA, deux sources externes comportant des informations véhicules, ont été évaluées par rapport à la base de référence afin de juger une bonne intégration des données. Ces bases ont été utilisées par le passé sans avoir réalisé d'étude sur la cohérence des données.

Dans le cas du fichier SIV, en reliant par la clé immatriculation du véhicule × date de première mise en circulation, on constate un faible pourcentage d'erreur entre les différentes variables, notamment sur des informations telle que la marque, le nom commercial, la puissance fiscale, l'énergie, ... ($< 1\%$ d'erreurs). Plusieurs retraitements ont été nécessaires avant d'obtenir ce résultat.

Dans le cas de la base SRA, en reliant par le code SRA, on constate également un faible pourcentage d'erreur. Ces erreurs sont quasi-identiques au fichier SIV. En effet, le code SRA de la base interne est renseigné à l'aide d'un algorithme faisant appel au fichier SIV, sous-section 2.2.3, page 35. Peu de codes SRA sont renseignés (environ 25%), c'est pourquoi cet algorithme est utilisée pour compléter les informations véhicules. La plupart des erreurs seront rectifiées dans le même temps.

L'algorithme, précédemment cité, émet l'hypothèse que le fichier SIV et la base SRA ont des données cohérentes, mais *a priori* ce n'est pas nécessairement le cas. Une étude à soupçonné une divergence d'informations sur la modalité véhicule à motorisation électrique. Afin de challenger l'information, le fichier SIV lié à la base véhicule et la base SRA ont été fusionnés par le couple marque du véhicule × nom commercial. Le code SRA ou l'immatriculation du véhicule ne sont pas des clés viables, car le code SRA est construit à partir du fichier SIV. Trois critères d'évaluation sont étudiés pour vérifier la pertinence de la variable *veh_ energie* dans la base SRA :

— 100% **véhicules électriques**. En principe, le couple marque et nom commercial est composé de 100% de véhicules électriques dans les bases SIV et SRA. En contrôlant l'énergie moteur du véhicule sur ce couple pour les deux bases, des faux négatifs peuvent être détectés. En effet, dans le cas où la règle n'est pas respectée, cela signifierait qu'au moins un véhicule de la base SRA est identifié à tort comme essence alors qu'il s'agirait d'un véhicule électrique au sens du critère. L'inverse est possible mais ne sera pas étudié. Le but est de challenger la base SRA à partir du fichier SIV, en supposant que le fichier SIV soit plus exhaustif que la base SRA,
— $]0\%, 100\%[$ **véhicule électriques**. Lorsque le couple marque et nom commercial de la base SIV a au moins un véhicule électrique alors cela doit également être le cas dans la base SRA. Ainsi, cette méthode repère également les faux négatifs. Effectivement, si le critère n'est pas respecté alors l'un des véhicules de la base SRA est indiqué comme essence, alors qu'il devrait être considéré comme électrique. Cela présuppose l'exhaustivité de la base SRA.
— 0% **véhicules électriques**. Lorsque le couple marque et nom commercial de la base SIV contient $0\% + -$ de véhicule électrique alors la même observation doit être constatée dans la base SRA. Cette méthode distingue les faux positifs. Assurément, en supposant que le critère ne soit pas honoré alors les véhicules électriques sont mal renseignés, ils devraient être à essence.

Finalement, une seule erreur a été signalée, hormis les anomalies qui s'expliquent par 2 raisons :

— la base SRA n'est pas exhaustive. Certains couples marque nom commercial version électrique sont manquants tel que la *RENAULT MASTER Z.E.* ou la *VOLKSWAGEN E-CRAFTER*,
— des véhicules non commercialisés (prototypes) sont renseignés dans le fichier SIV,
— mauvais référencement des énergies hybrides par le fichier SIV.

À ce stade, l'ensemble des données a été étudié et traité afin d'assurer la cohérence et la qualité

des données. Finalement, les informations de la base SRA sont jugées plus fiables que la base véhicule elle-même malgré sa non-exhausitivié. En effet, elles ont été conçues pour les assureurs dans un schéma encadré et contrôlé. Aucun traitement a été réalisé sur les données de la base SRA. En revanche, des problèmes de complétion des données ont été relevés.

2.2.3 Algorithme code SRA

La base véhicule est complétée par les informations de la base SRA. Cependant, seulement 25% des codes SRA sont renseignés dans la base véhicule. Ainsi, pour résoudre ce problème, trois étapes ont été nécessaires :

— complétion des codes SRA par les informations déjà présentes sur une année de vision antérieure,
— application de l'algorithme SRA se basant sur des données véhicules externes et internes,
— sélection du code SRA jugé le plus pertinent.

Complétion des codes SRA de la base véhicule par les informations passées

La première étape consiste à enrichir les codes d'identification des véhicules par l'information qui est déjà présente sur une année de vision antérieure ou un contrat différent. Un véhicule sera identifié par son immatriculation, sa date de première mise en circulation et sa marque. L'immatriculation n'est pas une information suffisante, car des erreurs ont été découvertes sur certains véhicules (sous-section 2.2.4, page 36). De plus, il existe des cas rares où un même véhicule possède plusieurs codes SRA, et où le code référencé ne renvoie pas à des informations en cohérence avec les données de la base véhicule. Ainsi, ces codes seront corrigés dans l'étape suivante.

Définition des codes SRA voisins à partir des informations véhicules

Après l'enrichissement des codes SRA à partir des informations passées, il en suit la problématique suivante : *Comment récupérer le code SRA d'un véhicule ?*

Un algorithme a été conçu afin de répondre à ce besoin en 2017. L'entrée de cet algorithme est une liste d'immatriculation et la sortie est une liste de codes SRA. L'algorithme, codé en **Python**, a été révisé afin de s'adapter à différents cas d'erreurs, de simplifier la compréhension du code et d'améliorer la définition du voisinage. Les bases appelées sont la base SIV et la base SRA à partir d'un système de gestion de bases de données, *postgreSQL*. L'algorithme se résume en 2 étapes :

— rechercher l'immatriculation du véhicule dans le fichier SIV dont les données ont été préalablement retraitées par les informations de la base véhicule,
— corréler les informations du véhicule présent dans le fichier SIV aux données de la base SRA.

La marque du véhicule est la variable clé pour récupérer le code SRA. Cependant, il s'agit d'une donnée peu propre dans le fichier SIV. En préambule, un traitement des caractères spéciaux est réalisé, puis une correction de la marque a été faite grâce notamment au CNIT (Code National d'Identification du Type) indiqué dans le fichier SIV et en case D.2.1 de la carte grise. Ce code est associé à chaque type, variante, version de toute réception communautaire de véhicules, anciennement appelé *Type Mines*. Ce code est constitué soit de 12, soit de 15 caractères. Dans la version à 12 caractères, ce sont les 2 lettres situées en deuxième position qui indique la marque du véhicule, dans la version à 15 caractères, ce sont les 3 lettres situés en quatrième position.

Ainsi, pour chaque code marque est associé une marque de véhicule avec la plus grande fréquence de présence. De façon générale, une marque sans erreur de caractère est référencée plus de fois qu'une marque erronée.

De même, l'opération est répétée avec le VDS du numéro VIN (*Vehicle Identification Number*). Le VIN a été normalisé en 1981 sous un code alphanumérique unique de 17 caractères indiqué en case E de la carte grise. Le VIN est divisé en trois parties,

— WMI (*World Manufacturer Identifier*) est le code constructeur,
— VDS (*Vehicle Description Sector*) englobe les caractères de 4 à 9 et est le code descripteur du véhicule au même titre que le CNIT ou le code GTA,
— VIS (*Vehicle Indicator Sector*) exprime un code indicateur.

Ainsi, l'algorithme corrige de façon intelligente la marque du véhicule. Ensuite, le fichier SIV est retraité avec les informations de la base véhicule afin de corriger les erreurs et d'ajouter des informations.

Enfin, le fichier SIV est corrélé aux informations véhicules de la base SRA par une approche ordonnée progressive. Les informations véhicules présentes à la fois dans le fichier SIV et la base SRA sont notamment le numéro CNIT (présent dans la base SRA sous une version abrégée de 10 caractères), l'énergie, la puissance fiscale, ainsi que neuf autres variables.

La carrosserie, le nombre de places assises et la cylindrée sont trois informations qui ont été ajoutées lors de la révision de l'algorithme. La carrosserie est une information présente dans la base véhicule néanmoins un regroupement des modalités a été nécessaire, pour obtenir une correspondance exacte avec la variable carrosserie, indiquée dans la base SRA.

Sélection du code SRA

Finalement, cet algorithme cherche à approximer au mieux la version réelle du véhicule étudiée. Cependant, malgré toutes les informations à disposition, le manque de données et de précision sur les variables conduit à avoir une immatriculation associée à plusieurs codes SRA (dans environ 80% des cas). En effet, l'algorithme définit un voisinage, c'est-à-dire qu'il renvoie en sortie tous les codes SRA respectant les caractéristiques techniques d'un véhicule donné. Or, théoriquement, à un véhicule est associé un unique code SRA.

Il en résulte donc la problématique suivante : *Quel est le code SRA le mieux adapté parmi le voisinage ?*

En fait, les codes SRA présents dans la base véhicule sont, pour la plupart, redondants. D'une entreprise à une autre, les mêmes types de véhicules sont choisis. En moyenne, un code SRA se compose de 100 véhicules. Ainsi, une idée pour sélectionner au mieux le code SRA après application de l'algorithme est d'étudier la fréquence d'apparition des codes SRA sur un groupe de variables.

Illustrons avec un exemple simple : supposons que pour une immatriculation, il y ait 2 codes SRA associés notés A et B. Pour ces 2 codes SRA, les informations véhicules, (marque, nom commercial), sont respectivement (C, E) et (C, F). Dans la base véhicule, les informations complémentaires suivantes sont données :

Groupe de variables	Modalité	Fréquence
Marque	C	60%
	D	40%
(Marque, Nom commercial)	(C, E)	20%
	(C, F)	40%
	(D, G)	40%

La marque des deux codes SRA est identique (C), il n'est donc pas possible de déterminer lequel des deux codes a la marque la plus représentée. Par contre, en regardant à une maille plus fine (marque, nom commercial), le code SRA B sera sélectionné, car le couple (C, F) a la fréquence d'apparition la plus élevée ($40\% > 20\%$).

Par conséquent, les codes SRA sélectionnés seront aux plus proches de la base véhicule. Enfin, après sélection, pour 10% des immatriculations, plusieurs codes SRA pour une même immatriculation subsistent. Ces codes font référence à des types de véhicules très similaires. Le voisinage a été restreint par la première méthode de sélection. La deuxième étape de sélection consiste à choisir aléatoirement un des codes du voisinage. Le jeu du hasard a pour intérêt de moyenner les écarts sur un grand nombre de données.

Par conséquent, 94% des véhicules de la base ont désormais un code SRA, contre 25% avant l'application de l'algorithme.

2.2.4 Algorithme de correction des immatriculations

Les informations véhicules ou sinistres sont inscrites sous forme d'EDI (Échange de données informatisées) et envoyées par les courtiers, ce sont en général des fichiers **Excel**. Des outils de souscription pour les informations sur le parc et des outils de gestion de sinistres pour les informations sur les sinistres permettent quant à eux de sauvegarder chacune de ces informations dans respectivement une base véhicule et une base sinistre différentes selon l'outil utilisé.

En reliant la base véhicule à la base sinistre, certains sinistres ne sont pas captés par la base véhicule avec les clés immatriculation, numéro de contrat et année de vision. Le numéro de contrat est un numéro créé numériquement afin d'identifier un parc de véhicules. C'est donc à la maille contrat et année de vision qu'une correction sera réalisée. L'objectif est de récupérer les sinistres non captés afin de ne pas perdre une partie de l'information. L'étude qui suit a été réalisée sur la sinistralité de 2016 à 2020, et sur les contrats composés de véhicules de moins de $3,5$ tonnes, nommés par la suite parc VL. Pour rappel, le jeu de données est composé uniquement de véhicules terrestres à moteur de moins de 3,5 tonnes : véhicule de catégorie 1. Or, l'étude est réalisée à la maille contrat et vision, ainsi toutes les autres catégories de véhicules sont ajoutées. Se concentrer sur le parc VL permet de limiter le biais causé par les autres catégories de véhicule. De plus, ces contrats représentent 57% des individus et 75% des véhicules de catégorie 1. En effet, on retrouve ci-dessous la proportion des véhicules selon le type de contrat :

	Contrat M3T5	Parc
Véhicules léger de moins de 3,5 tonnes	80%	60%
Engins de chantier + Engins agricole	12%	15%
Autres	9%	25%

FIGURE 2.2 – Proportion des véhicules selon leur catégorie et leur famille de contrat

L'algorithme est réalisé sur un périmètre plus important que le sujet de ce mémoire, car l'immatriculation de véhicule d'un sinistre non capté peut ne pas correspondre à un véhicule de catégorie 1. Aucune information sur le véhicule hormis l'immatriculation n'est présente dans les bases sinistres. Il existe plusieurs explications sur la présence de tels sinistres :

— l'immatriculation est inconnue ou non renseignée. Un sinistre dont l'immatriculation est vide ne peut pas être identifiée par la base véhicule. En général, pour un véhicule dont l'immatriculation est non conventionnelle, un pseudonyme est inscrit à la place. C'est le cas par exemple de certains engins agricoles. Coté souscription et coté gestion de sinistres, ce pseudonyme est généralisé,
— l'immatriculation est connue, mais ne correspond pas au véhicule assuré. Par exemple, dans de rares cas, pour un sinistre lié à la RC, l'immatriculation inscrite correspond au véhicule de la personne non assuré chez AXA, acteur de l'accident avec un véhicule assuré. Il s'agit donc d'un véhicule hors périmètre. Pour récupérer l'identifiant du véhicule assuré chez AXA, il y a deux possibilités : utiliser une seconde base sinistre ou se référer au sinistre non vieilli, c'est-à-dire aux premières informations reçues lors de la survenance du sinistre (les incohérences sont moins fréquentes),
— l'immatriculation contient une erreur. Contrairement au numéro de contrat, il existe des erreurs de saisies sur les immatriculation de véhicule. Pour pallier à ce problème, un algorithme sur **Python** a été conçu. Celui-ci permet de rectifier au mieux les immatriculations afin de pouvoir capter les sinistres associés,
— le numéro de contrat est inconnu. Ce cas rare entraîne malheureusement une perte d'information sur les sinistres.

La correction des immatriculations afin de pouvoir capter les sinistres se déroule en deux étapes.

La première étape consiste à remplacer les immatriculations des sinistres non captés par d'autres immatriculations à date d'observation différente ou provenant d'une autre base sinistre. En effet, pour de très rares cas, dont une explication partielle a été donnée précédemment, les bases sinistres, pour un même numéro de sinistre, ne renvoient pas la même immatriculation. Il s'agit de corriger l'immatriculation par une autre immatriculation se trouvant dans la base véhicule afin de capter le sinistre. Il y a des bases sinistres pour toutes les branches par année de survenance et des bases sinistres dédiées au périmètre entreprise. Cette étape permet de récupérer 38% des sinistres précédemment non captés.

La deuxième étape consiste à corriger les immatriculations de la base sinistre qui sont jugées erronées. On suppose que la base véhicule comporte moins d'erreurs que la base sinistre, car elle est une base de référence. En réalité, la base véhicule contient également des erreurs. Néanmoins, en reliant les immatriculations avec le fichier SIV, la base véhicule introduit moins d'erreur que la base sinistre. Parmi les sinistres non captés, les immatriculations présentes dans le fichier SIV avec les clés date de mise en circulation et immatriculation sont retirées de l'étude ce qui entraîne une perte d'information. En fait, un véhicule se retrouvant dans le fichier SIV est considéré existant et donc il n'est pas souhaitable de

corriger son immatriculation.

Après avoir effectué la première étape de récupération des sinistres non captés, seulement 1% des véhicules sinistrés étudiés se retrouvent dans le fichier SIV. Cela semble confirmer l'hypothèse que des immatriculations sont erronées. En effet, le fichier renseigne théoriquement l'ensemble des immatriculations des véhicules assurés circulant en France.

Pour corriger les immatriculations plusieurs approches ont été utilisées :

1. les immatriculations de la base sinistre de longueur supérieure ou égale à quatre, se référant à une fraction d'immatriculations de la base véhicule, sont rectifiées.
   *Exemple : **V522T** comprise dans F**V522T**F.*
   Le choix de quatre comme longueur d'immatriculation est dû à la présence de véhicules monégasques qui sont précédés ou pas du sigle *MC* (par exemple, dans la base sinistre, figure l'immatriculation fictive *A123* alors que dans la base véhicule on trouve *A123MC*),
2. les immatriculations de la base sinistre de longueur supérieure ou égale à six, correspondant à un caractère près à des immatriculations de la base véhicule, sont remplacées.
   *Exemple : FV522**X**F sera reliée avec FV522**T**F, car seul un caractère est à changer et la position des caractères reste identique.*
   les immatriculations de longueur supérieure ou égale à six sont intéressantes à rectifier, car dans les immatriculations réglementées [1], hormis les immatriculations monégasques, une immatriculation comptabilise un minimum de six caractères,
3. les immatriculations de la base sinistre de longueur supérieure ou égale à six contenant une inversion de deux caractères par rapport à une immatriculation de la base véhicule, sont corrigées.
   *Exemple : il y a correspondance entre FV52**T**2F et FV522**T**F, car seulement deux caractères sont erronés et leur position est inversée*,
4. les immatriculations contenant deux caractères erronés, sont retouchées.
   *Exemple : FV5**Y**2**X**F ressemble à FV522**T**F, car les caractères différents entre les deux immatriculations sont dans la même position et sont de même longueur*,
5. les immatriculations sont segmentées en deux en supprimant un caractère dans une des fractions. Si les deux fragments associés à une immatriculation de la base sinistre sont compris dans une immatriculation de la base véhicule, alors l'immatriculation est convertie.
   *Exemple : on suppose le 4eme caractère de l'immatriculation 390**Y**N42 erroné. Il reste donc deux fragments **390** et **N42**. Ces deux éléments sont compris dans l'immatriculation **390**3YN**42***,
6. cette dernière approche consiste à reprendre les méthodes précédentes en considérant cette fois-ci que c'est l'immatriculation de la base véhicule qui est erronée. Il s'agit du travail inverse. Ainsi, l'hypothèse de départ sur la qualité de la base véhicule est contrôlée et les approches 1 et 5 de contenant/contenu (exemple : V522T contenu dans FV522TF) sont appliquées dans les deux sens.

La plupart de ces approches sont corrélées. Cela permet de corriger tout en contrôlant un éventuel effet de sur-apprentissage. En effet, une correction trop fine aurait pour effet de modifier trop d'immatriculations par erreur. À l'inverse, une correction grossière a pour conséquence de perdre une trop grande quantité d'information sinistre. Finalement, ce sont les approches 5 et 4, dans cet ordre, qui seront appliquées. En effet, la 4eme approche seule est trop grossière. Plusieurs immatriculations peuvent remplacer l'immatriculation erronée (par exemple, FV5**Y**2**X**F peut être rectifié à la fois par l'immatriculation FV5**Z**2**G**F et par FV5**A**2**A**F). Mais si l'on applique au préalable la 5eme approche, cette dérive est moindre. D'ailleurs, les cas où plusieurs immatriculations sont candidates à une correction, sont écartés. De plus, un traitement spécifique est réalisé pour les véhicules temporaires monégasques.

Chaque requête est considérée à la maille contrat et année de vision. Le numéro de contrat étant codé numériquement, il n'y a, *a priori*, aucune erreur. Ainsi, une immatriculation de la base sinistre considérée erronée sera approchée à une immatriculation de la base véhicule ayant le même numéro de contrat et la même année de vision. Pour rappel, un contrat *parc dénommé* contient au maximum 50 véhicules, ce qui réduit considérablement un éventuel effet de sur-apprentissage de l'algorithme. Finalement, après application des deux étapes de correction, c'est environ 50% des immatriculations de véhicules sinistrés qui sont rectifiées.

1. L'immatriculation est obligatoire pour tout véhicule à moteur circulant sur la voie publique. Le numéro d'immatriculation figure sur le certificat d'immatriculation du véhicule. Ce dernier est une série alphanumérique sous la forme AA-123-AA depuis la mise en place du Système d'Immatriculation des Véhicules (SIV) en avril 2009. Il est attribué à vie au véhicule jusqu'à sa destruction. L'ancien format FNI du numéro d'immatriculation est de la forme 1234 AA 00.

Avant correction

ANSURV	FREQ RC	CHARGE RC
2016	19,71%	19,46%
2017	6,02%	4,58%
2018	5,56%	5,09%
2019	4,88%	3,69%
2020	4,92%	5,26%
TOTAL	8,35%	7,95%

Après correction

ANSURV	FREQ RC	CHARGE RC
2016	2,09%	1,52%
2017	2,81%	2,35%
2018	3,51%	3,14%
2019	3,23%	2,49%
2020	3,19%	3,47%
TOTAL	2,96%	2,47%

FIGURE 2.3 – Récupération des sinistres non captés par la base véhicule sur le parc VL

L'année 2016 n'a pas été prise en compte dans la suite de l'étude, car elle présente une trop forte sinistralité non captée, malgré une correction possible satisfaisante. 20% de la fréquence n'était pas expliquée (Figure 2.4).

Au total, $2,47\%$ de la charge ne peut être introduite dans l'étude. Cette perte d'information peut être expliquée par des engins agricoles et de chantier qui représentent 12% du parc VL, et dont certains sont identifiés par une plaque constructeur et n'ont pas d'immatriculation. Les sinistres non captés pour des raisons d'immatriculations non renseignées représentent deux tiers de la sinistralité non expliquée après correction par l'algorithme et deux tiers de la charge. On peut émettre l'hypothèse donc que cette charge de sinistres est liée à des véhicules hors périmètres (engin agricole et engin de chantier). Néanmoins, il faudra garder en tête que le tarif qui sera modélisé par la suite devra être majoré d'au plus 1% ($2,47\% \times (1 - \frac{2}{3}) \approx 0.82\% < 1\%$).

À présent que l'on a jugé de l'efficacité de l'algorithme sur les contrats composés uniquement de véhicules de moins de $3,5$ tonnes, l'algorithme est appliqué sur l'ensemble des contrats EN *parc dénommé* sur les années 2017 à 2019, afin de capter l'ensemble des véhicules sinistrés de catégorie 1.

Avant correction

ANSURV	FREQ RC	CHARGE RC
2017	7,44%	5,60%
2018	6,85%	5,44%
2019	6,37%	4,50%
TOTAL	6,89%	5,20%

Après correction

ANSURV	FREQ RC	CHARGE RC
2017	3,92%	3,39%
2018	4,48%	3,13%
2019	4,33%	2,98%
TOTAL	4,24%	3,18%

FIGURE 2.4 – Récupération des sinistres non captés par la base véhicule sur l'ensemble du périmètre EN parcs dénommés

Remarque : 70% des sinistres non captés après correction sont liés à des immatriculations non renseignées.

2.3 Prime pure à la prime commerciale

Le ratio S/P [2] ou encore ratio combiné (CR) est un rapport entre les décaissements (frais généraux, commissions versées aux agents, provisions pour sinistres, et remboursement des sinistres) et celui des encaissements (primes versées par les assurés) sur un même contrat d'assurance. Il s'agit d'un ratio permettant à l'assureur de mesurer la rentabilité technique de l'assurance sur une période donnée. Il doit être supérieur à 100%, pour que les prestations versées par l'assureur soient supérieures à ses recettes.

Le ratio combiné est déterminé en sommant :

— *Loss ratio* : mesure le coût des sinistres par rapport au montant des primes encaissées. Il couvre les indemnités effectivement versées aux assurés, mais aussi les charges estimées correspondant aux sinistres en cours.

$$loss\ ratio = \frac{Coût\ sinistres}{primes\ acquises}$$

2. sinistre sur prime

— *Expense ratio* : prend en compte les coûts de commercialisation et de gestion, commissions versées aux intermédiaires, frais de gestion des sinistres.

$$expense\ ratio = \frac{frais\ de\ gestion,\ commercialisation}{primes\ acquises}$$

Il faut noter que les coûts de sinistres et les primes mentionnées sont nets de réassurance.

Cependant, il s'est avéré qu'avec des ratios combinés inférieurs à 100%, le résultat technique de la compagnie peut être négatif.

Le ratio combiné économique (ECR) fournit une version économique de la performance de souscription. L'ECR permet d'estimer la rentabilité d'un portefeuille ou d'une affaire du point de vue de l'actionnaire. Un ECR à 100% signifie que l'actionnaire aura une rémunération sur les risques techniques correspondant à ses attentes.

$$ECR = CR + CAT_{adj} + Escompte + Impot + Exg_{act}$$

— CAT_{adj} est l'ajustement catastrophe. Pour évaluer la rentabilité de la souscription pour une année donnée, il faut réajuster les réserves en prenant en compte la rareté des sinistres catastrophes,
— $Escompte$ correspond aux revenus issus du placement de la prime aux conditions du marché et dépend de la duration de la branche,
— $Impot$ est le taux d'impôts sur les sociétés,
— Exg_{act} est l'exigence des actionnaires au-delà du taux sans risque, aussi appelé coût du risque. Les actionnaires fournissent le capital de solvabilité requis à l'entreprise d'assurance et cette dernière l'investit dans des actifs non risqués. Les actionnaires exigent un retour sur capital (*return on equity*), en plus de la rentabilité espérée de l'investissement évaluée par la réglementation de Solvabilité II.

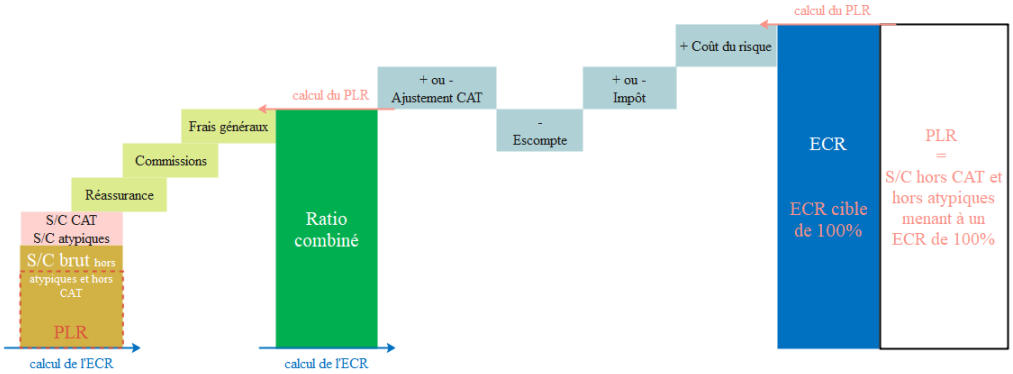


FIGURE 2.5 – De l'ECR au PLR

Le modèle interne d'AXA, nommé le STEC[3] comprend l'ECR (*Economic Combined Ratio*) comme un livrable officiel. Le PLR (*Permissible Loss Ratio*) est le loss ratio sinistres attritionnels à prime maximum pour avoir un ECR à 100%. Le PLR qui correspond simplement au S/C d'équilibre tel que $ECR = 1$ est décliné pour chaque branche par segment de produit (réseau, activité de l'entreprise et taille de la flotte) avec une approche économique des différents postes :

— affectation de la charge probabiliste climatique et atypique par segment,
— affectation des frais généraux selon la taille de l'affaire, le mode de délégation et le réseau,
— affectation des coûts de réassurance selon l'espérance de charge récupérée par segment.

Le PLR a pour objectif de cartographier la contribution de chaque segment aux résultats, d'identifier les cibles de développement, d'être au plus juste prix en fonction de la taille de l'affaire et d'assurer

3. *Short Term Economic Capital* : niveau de fonds propres requis pour une compagnie d'assurance afin qu'elle puisse honorer ses engagements envers les assurés quels que soient les événements qui pourraient survenir. Le niveau cible représente le capital nécessaire pour absorber un choc provoqué pour un risque bicentenaire (quantile à 99.5%).

l'homogénéité des différents outils.

Dans le cadre de l'étude, le PLR permet de passer de la prime pure, qui sera calculée dans le chapitre suivant, à la prime commerciale. Ce coefficient est calculé en comprenant les sinistres attritionnels (sinistres dont la charge est supérieure à $2M€$), c'est pourquoi ces sinistres seront exclus.

2.4 Préparation à la modélisation

La base de données est divisée en deux, une base d'apprentissage représentant 80% des données et une base de test indépendante composée des 20% données restantes.

De plus, pour ajuster un modèle, une base de validation est construite en appliquant la méthode de la validation croisée à k blocs (*k-fold cross-validation*). Une base de validation permet entre autres une estimation de la fiabilité d'un paramètre d'un modèle.

Le problème généralement rencontré en modélisation est le sur-apprentissage. En fait, l'objectif de la modélisation de la prime pure (ou cout moyen × fréquence) n'est pas de prédire l'apparition exacte d'un sinistre et son coût, mais de prédire la charge moyenne future d'un groupe homogène de risque pour un taux de présence d'un an. Il est impossible en assurance de prédire avec certitude la charge à venir à cause du caractère aléatoire du risque, aussi appelé bruit. Une modélisation de la charge sur l'ensemble des données ne généralise donc pas les caractéristiques des données et perd son pouvoir de prédiction sur un nouvel échantillon. En revanche, il est possible de capter un signal à partir de certaines variables pour expliquer des effets sur la variable à prédire. La validation croisée assure que seul ce signal soit modélisé.

La validation croisée consiste en un découpage du jeu de données en k échantillons de manière aléatoire pour calibrer le modèle sur $k-1$ échantillons et valider le modèle sur l'échantillon de validation restant. Cette séparation de la base d'apprentissage en un échantillon d'entraînement et de validation est répétée afin que l'ensemble de la base de données soit utilisé. Cette utilisation du jeu de données permet de tirer plusieurs ensembles de validation d'une même base et ainsi d'obtenir une estimation plus robuste de la performance de validation du modèle. La base de test permet de mesurer les performances de prédictions de chaque modèle pour une comparaison et une évaluation finale.

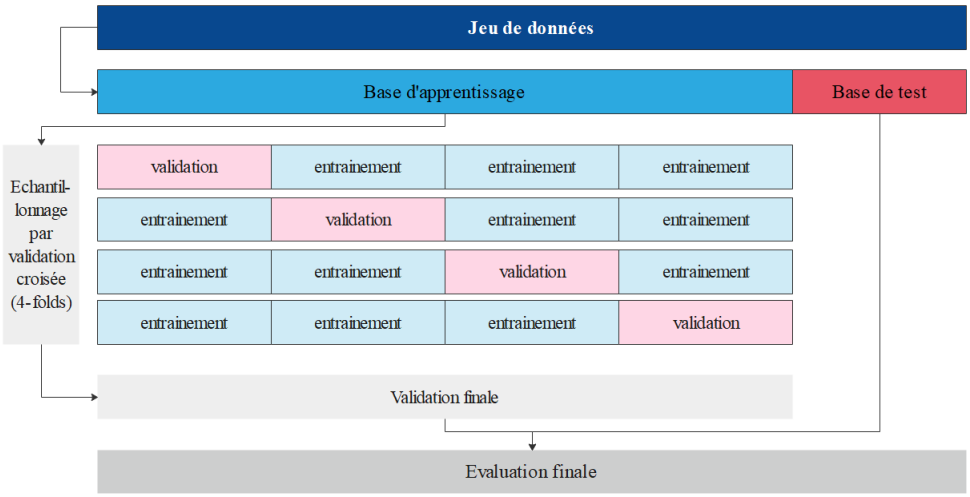


FIGURE 2.6 – Diagramme du découpage de la base de données

Le choix de k est arbitraire, c'est un compromis entre vitesse et variance des résultats. Pour $k = n$ [4], où n est le nombre de lignes de notre jeu de données, l'ajustement de $n-1$ modèles implique un temps de cryptage encore aujourd'hui irréalisable. De même, un k trop petit n'exploite pas assez la robustesse de la méthode. Dans les écrits, il est conseillé de choisir un k compris entre 4 et 9, pour un jeu de données séparé en une base d'apprentissage et une base de test. Dans la suite, une validation croisée à 8 échantillons est réalisée, exceptée pour la modélisation du coût moyen où $k = 4$ à cause de la quantité faible de sinistres et afin de mieux visualiser les résultats.

4. La *cross-validation k-folds* avec $k = n$ est le cas particulier de la validation croisée appelée *"leave-one-out"*. Cette méthode ne peut pas être utilisée sur une grande base de données. Une approximation utilisant les coefficients de levier et une généralisation existent, mais ne seront pas traitées au cours de ce mémoire.

Chapitre 3

Modélisation de la prime pure

La détermination du coût du risque est réalisée à partir de la modélisation de la prime pure. En assurance, ce sont des modèles paramétriques type GLM ou GAM qui sont communément employés. Ces ajustements sont faciles à mettre en œuvre grâce à la définition de cases tarifaires, des groupes homogènes de risque. La prime pure, c'est-à-dire le coût moyen du risque pour une case tarifaire est déterminé à l'aide de la sinistralité passée.

Dans le cadre de l'étude, nous utiliserons par la suite un outil réalisé sur *python* par une société privée externe. Cet outil réalise un ensemble de manipulations de données et d'entraînements de modèles type GAM. Cependant, il est limité et nécessite des interventions à jugement d'expert. Ainsi, plusieurs améliorations complémentaires seront apportées afin de segmenter au mieux le risque.

3.1 Modèles linéaires généralisés

L'objectif de la régression est de modéliser la valeur d'une variable y à expliquer, ou réponse, en fonction des valeurs de n variables explicatives $x^{(1)}, \ldots, x^{(n)}$.

En assurance, la variable à étudier est classiquement le nombre de sinistres ou le coût d'un sinistre et s'explique par des spécificités techniques d'un bien ou le descriptif d'un assuré ...

3.1.1 Théorie

Dans le cadre d'une Régression linéaire Généralisée, GLM en anglais, les valeurs de y et de $x = (x^{(1)}, \ldots, x^{(n)})$ se modélisent à l'aide d'un couple de variables aléatoires (Y, X) avec $X = (X^{(1)}, \ldots, X^{(n)})$. Ainsi, la régression vise à déterminer l'espérance conditionnelle $\mathbb{E}(Y|X = x)$ à partir de la loi conditionnelle $Loi(Y|X = x)$.

Les réponses $(y_i)_{i=1,\ldots,m}$ de $y = (y_1 \ldots y_m)$ sont issues des variables aléatoires indépendantes $Y_1, \ldots, Y_m$ telles que $Y_i \sim Loi(Y|X = x_i)$, où $x_i = (x_i^{(1)}, \ldots, x_i^{(n)})$

Un modèle linéaire généralisé est constitué de trois composantes essentielles,

- une composante systématique : fonction affine des variables explicatives x, noté $x\beta$ et appelé prédicteur linéaire. L'estimation du paramètre β est décrite par la suite,
- une composante aléatoire : spécification du type de $Loi(Y|X = x)$ au sein de la famille exponentielle,
- une fonction de lien : spécification de la relation entre $E(Y|X = x)$ et la composante systématique du modèle,

$$g(E(Y|X)) = \sum_{j=0}^{d} \beta_j x^{(j)},$$

où g une fonction bijective appelée fonction de lien, $\beta = (\beta_0, \ldots, \beta_d)$ les coefficients du modèle et $(x^{(0)}, \ldots, x^{(d)})$ l'encodage des variables explicatives $x^{(0)}, \ldots, x^{(m)}$, décrit par la suite.

Remarque : un GLM est une généralisation des modèles de régression linéaires usuels dont le modèle linéaire gaussien où $Loi(Y|X = x) = \mathcal{N}(\mu, \sigma^2)$ et où g est la fonction de lien canonique $g : x \to x$.

Un GLM a une composante aléatoire suivant une loi appartenant à la famille exponentielle. Une loi de probabilité sur $\mathbb{R}$, discrète ou continue, appartient à la famille exponentielle si elle possède une densité de la forme,

$$f(y|\theta,\Phi) = exp(\frac{y\theta - b(\theta)}{a(\Phi)} + c(y,\Phi)),$$

où $\Phi > 0$ est le paramètre de dispersion, $\theta \in I$ est le paramètre naturel, $b : I \subset \mathbb{R} \to \mathbb{R}$ est supposée régulière sur l'intervalle I, et $a()$ et $c()$ sont des fonctions.

En supposant l'espérance et la variance définies, une des propriétés fondamentales de cette famille de lois est $\mathbb{E}[Y] = b'(\theta) = \mu$ et $Var[Y] = b''(\theta).\Phi = \sigma^2$.

Exemple 1 : Loi Poisson

La loi de Poisson de moyenne λ, $\mathcal{P}(\lambda)$ appartient à cette famille,

$$f(y|\lambda) = exp(-\lambda)\frac{\lambda^y}{y!} = exp(y.log\lambda - \lambda - log(y!)),$$

où $y \in \mathbb{N}$, $\theta = log\lambda$, $\Phi = 1$, $a(\Phi) = 1$, $b(\theta) = exp(\theta) = \lambda$ et $c(y,\Phi) = -log(y!)$

Exemple 2 : Loi Binomiale Négative

La loi Binomiale Négative, de paramètre r et p, est définit sous la forme,

$$f(k|r,p) = exp(y.log(p) + r.log(1-p) + log\begin{pmatrix} y+r-1 \\ y \end{pmatrix}),$$

où $y \in \mathbb{N}$, $\theta = log(p)$, $a(\Phi) = 1$, $b(\theta) = -r.exp(\theta) = \lambda$

Exemple 3 : Loi Gamma

La loi Gamma de moyenne μ et de variance ν^{-1}, incluant la loi exponentielle, s'exprime sous la formule,

$$f(y|\mu,\nu) = exp(-\lambda)\frac{\lambda^y}{y!} = exp(y.log\lambda - \lambda - logy!),$$

où $y \in \mathbb{R}_+$, $\theta = -\frac{1}{\mu}$, $\Phi = \nu^{-1}$, $a(\Phi) = \Phi$, $b(\theta) = -log(-\theta)$ et $c(y,\Phi) = (\frac{1}{\Phi} - 1)log(y) - log(\Gamma(\frac{1}{\Phi}))$

L'estimation des paramètres β est obtenue en maximisant la log-vraisemblance du modèle. Soit un ensemble de variables aléatoires $(Y_i)_{i=1,\ldots n}$, la vraisemblance de β est donnée par :

$$l(\beta) = \sum_{i=1}^{n} log(f(Y_i;\beta,\Phi))$$

La même procédure peut s'appliquer pour le paramètre de dispersion Φ.

L'encodage des variables explicatives est effectué, car les $x^{(j)}$ prennent leurs valeurs dans $\mathbb{R}$. Un encodage numérique des variables qualitatives est donc indispensable.

Plusieurs choix d'encodage sont possibles, dans ceux retenus et couramment utilisés il y a l'encodage d'une variable catégorielle pure, distinction faite aux variables catégorielles ordinales, $x^{(l)}$ à K modalités par des variables indicatrices où chaque modalité $a_0,\ldots,a_{K-1}$ est topée hormis une. Cette exclusion d'une variable indicatrice provient de la relation de dépendance affine entre les indicatrices, $\sum_{k=0}^{K-1} \mathbb{1}(x^{(l)} = a_k) = 1$ qui rend le modèle non-identifiable.

Pour les variables catégorielles ordinales comme la classe de prix ou le groupe d'un véhicule, il est classiquement envisagé d'utiliser un codage emboité, $\mathbb{1}(x^{(l)} >= 1),\ldots,\mathbb{1}(x^{(l)} >= K-1)$. La méthode retenue par l'outil est une méthode bayésienne où il est utilisé comme *a priori* une distribution des coefficients des niveaux consécutifs très concentrée sur 0. Autrement dit, la différence entre une variable catégorielle et une variable ordinale réside dans la façon de faire passer le test statistique. Si la variable

est catégorielle, l'hypothèse nulle est "le coefficient de la modalité est nul" tandis que pour une variable ordinale, l'hypothèse nulle est "le coefficient de la modalité est égale à celui du voisin". Pour une variable quantitative, le modèle va tendre à rapprocher les coefficients voisins entre eux alors que pour une variable catégorielle, les coefficients vont tendre à se rapprocher indépendamment vers 0.

Remarque : l'outil utilise des fonctions en escalier par morceaux pour regrouper des modalités d'une variable ordinale. Pour des raisons évidentes de confidentialité, la méthode ne sera pas plus détaillée.

Enfin, les variables quantitatives peuvent être directement prises telles quelles dans un GLM, mais cela ne garantit pas *a priori* que l'effet de cette variable sur la réponse soit modélisé de manière satisfaisante. C'est pourquoi sur ces variables un changement d'échelle (passage au log), une expression dans une base de fonctions (polynômes, splines ...) de dimension D ou une catégorisation s'impose en général.

Une modélisation seulement sur des variables catégorielles est généralement utilisée pour des raisons d'interprétation, de traitement informatique et d'un point de vue commerciale. Cela permet de segmenter le risque en case tarifaire. Un outil externe d'AXA sera utilisé. Ce dernier n'accepte que des variables explicatives catégorielles et a une structure tarifaire multiplicative. Il définit alors une constante permettant d'accéder à l'effet global des variables explicatives sur la composante systématique du modèle. Le terme constant est communément noté β_a dans la régression et il est mis à jour régulièrement en interne afin de prendre en compte, notamment, l'effet de l'inflation des prix.

Les lois exponentielles les plus couramment utilisées en assurance pour modéliser la fréquence, une variable discrète, sont la loi de Poisson et la loi binomiale négative qui est une extension de la loi de Poisson utilisée pour modéliser une sur-dispersion. La loi Gamma ou son alternative la loi Gaussienne inverse prédisent une réponse continue et positive dont le coût moyen. Enfin, une dernière loi utilisée classiquement dans la modélisation en assurance que ce mémoire traitera est la loi de Tweedie. Elle modélise directement une loi de Poisson composée avec une loi Gamma, permettant de prédire[2] la charge de sinistre.

Une mesure d'exposition est ajoutée dans le modèle, en plus des variables explicatives, afin de prendre en compte la durée d'exposition au risque pour le modèle de prime pure et le modèle de fréquence, et de considérer le nombre de sinistres pour le modèle de coût moyen. On distingue ainsi la réponse brute Y et la réponse normalisée par l'exposition $W = \frac{Y}{w}$ où w est l'exposition.

3.1.2 Modèles et lois

Modèles

Le modèle individuel et le modèle collectif sont deux évaluations de la prime pure à dissocier. Le modèle individuel de risque se traduit par une suite finie $X_1, \ldots, X_n$ de variables aléatoires indépendantes et identiquement distribuées. Le montant cumulé des sinistres S est alors défini par $S = \sum_{i=1}^{n} X_i$.

Dans ce modèle, la variable aléatoire X_k représente le cumul des indemnisations allouées pour les sinistres affectant l'assuré k pendant la période d'observation. La prime pure est définie comme l'espérance $\mathbb{E}[S]$ du montant cumulé des sinistres S.

Une seconde approche pour déterminer la prime pure est la méthode fréquence $\times$ coût moyen. Dans le modèle collectif de risques, on définit une suite infinie $(Y_k)_{k>=1}$ de variables aléatoires indépendantes et identiquement distribuées, et une variable aléatoire N indépendante et à valeur entière. Ainsi, le montant cumulé des sinistres S s'écrit $S = \sum_{i=1}^{N} Y_i$.

L'hypothèse forte d'indépendance de N par rapport à la suite de variables aléatoires $(Y_k)_{k>=1}$ implique que la prime pure s'écrit comme $\mathbb{E}[S] = \mathbb{E}[N] + \mathbb{E}[Y_1]$ dans la condition où ces espérances sont définies, c'est-à-dire $\mathbb{E}[Y_1] < +\infty$ et $\mathbb{E}[N] < +\infty$.

Ainsi, dans une approche fréquence $\times$ coût moyen, le principe repose sur la distinction et l'indépendance entre la fréquence de survenance d'un sinistre, et le coût d'un sinistre. Cette hypothèse d'indépendance est forte, mais permet néanmoins de vérifier l'efficacité et la stabilité du modèle de prime pure. En pratique, elle donne généralement, à condition d'avoir beaucoup de données, de meilleurs résultats qu'une approche prime pure, modèle individuel. Il est également plus simple de prendre en compte la surdispersion dans un modèle collectif, car elle est facile à identifier et à traiter au travers d'une loi Poisson et donc réduire le bruit, l'effet aléatoire.

Lois

Les distributions Tweedie appartiennent à la classe des modèles de dispersion exponentielle. La famille de distribution de la loi Tweedie est une sous-classe de la famille exponentielle et est très utile notamment pour modéliser une distribution continue pour des valeurs supérieures ou égales à 0 avec une masse importante de 0. Elle apparaît comme le modèle théorique adéquat dans le cadre de notre étude. Elle fait abstraction *a priori* de l'hypothèse d'indépendance entre fréquence et coût moyen. Il s'agit d'une famille de loi utilisée dans le modèle individuel pour simuler S lorsque $p \in]1,2[$ Le paramètre p est défini ci-après. En 1984, Tweedie a suggéré la famille suivante,

$$f(y|\nu,\Phi) = A(y,\Phi).exp(\frac{1}{\Phi}[y\theta(\nu) - \kappa(\theta(\nu)]),$$

où $\theta(\mu) = \begin{cases} \frac{\mu^{1-p}}{1-p} & \text{si } p \neq 1 \\ log(\mu) & \text{si } p = 1 \end{cases}$ et $\kappa(\theta(\mu)) = \begin{cases} \frac{\mu^{2-p}}{2-p} & \text{si } p \neq 2 \\ log(\mu) & \text{si } p = 2 \end{cases}$.

Cet ensemble de lois se définit par une relation entre l'espérance et la variance, $V[E[Y]] = \Phi.\ [E[Y]]^p$ avec Φ le paramètre de dispersion et $p \in \mathbb{R}\backslash]0,1[$ un paramètre de puissance. Selon la valeur de p, la loi est définie comme,

Lois	p	α	M_p	S_p
Stables extrêmes	$p < 0$	$1 < \alpha < 2$	$]0,\infty[$	$\mathbb{R}$
Gaussienne	$p = 0$	$\alpha = 2$	$\mathbb{R}$	$\mathbb{R}$
(N'existe pas)	$0 < p < 1$	$2 < \alpha < \infty$		
Poisson	$p = 1$	$\alpha = -\infty$	$]0,\infty[$	$\mathbb{N}$
Poisson-composées	$1 < p < 2$	$\alpha < 0$	$]0,\infty[$	$]0,\infty[$
Gamma décentrée	$p = 3/2$	$\alpha = -1$	$]0,\infty[$	$]0,\infty[$
Gamma multivariée	$p = 2$	$\alpha = 0$	$]0,\infty[$	$]0,\infty[$
Stables positives	$p > 2$	$0 < \alpha < 1$	$]0,\infty[$	$]0,\infty[$
Inverse gaussienne	$p = 3$	$\alpha = 1/2$	$]0,\infty[$	$]0,\infty[$
Stable extrême	$p = \infty$	$\alpha = 1$	$\mathbb{R}$	$\mathbb{R}$

La loi de Y est alors une loi de Poisson composée, avec des sauts Gamma lorsque $p \in [1,2]$,

$$Y \sim \mathcal{CP}oi(\mu^{2-p}\Phi(2-p), \mathcal{G}(-\frac{2-p}{\Phi(1-p}, \Phi(2-p)\mu^{p-1}))$$

Remarque : la distribution de la loi est composée d'une mesure de Dirac en 0 avec une distribution continue définie sur $\mathbb{R}_+$ très asymétrique à droite.

Les hypothèses à considérer lors de l'utilisation d'une loi de Tweedie pour modéliser la charge des sinistres sont :

— la charge annuelle est composée d'un nombre aléatoire de sinistres. Ce nombre est supposé suivre une loi de Poisson,
— les montants des sinistres sont indépendants et identiquement distribués selon une loi Gamma.

Les lois Poisson et Binomiale négative sont couramment utilisées pour simuler N dans le modèle collectif. Quant aux lois, Gamma, Log-Normale et Inverse gaussienne sont communément exploitées pour modéliser Y dans le modèle collectif.

Ces lois ne seront pas explicitées, car elles ont déjà été définies à partir de la famille de loi Tweedie ou dans la sous-section 3.1.1, page 42.

3.1.3 Critère de validation

De manière générale, les résidus constituent une mesure de l'écart entre l'observé y_i et la moyenne des prédictions du modèle pour la réponse $\hat{\mu}_i$. Il faut alors préciser la formule définissant l'écart étudié, les données utilisées (données d'apprentissage ou de test) et la façon de regrouper et/ou d'organiser les

données pour calculer et présenter les écarts de prédiction. L'objectif est de vérifier visuellement l'adéquation du modèle aux données, de repérer des tendances ou des données aberrantes. Une donnée est dite aberrante s'il s'agit d'un point ne suivant pas la loi du modèle comme les sinistres graves qui ont une distribution spécifique.

Classiquement, lorsque le modèle est valide, les résidus $r_1, \ldots, r_N$ constituent des réalisations de variables aléatoires $R_1, \ldots, R_N$ centrées [1], de variance unité [2] et de loi Normale $\mathcal{N}(0,1)$ indépendantes. En revanche, aucune de ces propriétés ne sont couramment vérifiées, mais restent des références. D'autres difficultés pratiques sont constatées dont la présence d'un grand nombre de points rendant la représentation graphique illisible et l'importance inégale des écarts de prédictions, par exemple lorsque des points distincts correspondent à des expositions différentes.

Plusieurs résidus peuvent être étudiés, plus ou moins adaptés selon les données utilisées :

— les résidus additifs sont définis comme

$$r_i^B = y_i - \hat{\mu}_i$$

Ce sont les résidus les plus simples, en revanche non adaptés pour les modèles multiplicatifs,

— les résidus multiplicatifs,

$$r_i^{O/P} = \frac{y_i}{\hat{\mu}_i} - 1,$$

permettent de prendre en compte les modèles multiplicatifs qui vont donc être préférés dans le cas de l'étude,

— les résidus de Pearson, en supposant la validité du modèle, permettent de disposer de résidus approximativement centrés et de variance constante. Les résidus de Pearson se définissent comme

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{v(\hat{\mu}_i/\omega_i)}},$$

— les résidus de déviance

$$r_i^D = signe(y_i - \hat{\mu}_i).\sqrt{\omega_i.D(y_i, \hat{\mu}_i)},$$

où $D(y_i, \hat{\mu}_i)$ est la déviance [3].

Quels que soient les résidus définis, il n'est pas attendu d'obtenir une distribution normale ou symétrique. En pratique, une manière de se rapprocher des propriétés attendues est de calculer des résidus agrégés sur des groupes de données, appelés *crunched residuals*, correspondant à des expositions suffisamment grandes de la forme,

$$\frac{\sum_{i \in G} \omega_i y_i - \sum_{i \in G} \omega_i \hat{\mu}_i}{\sqrt{\sum_{i \in G} \omega_i v(\hat{\mu}_i)}}$$

Pour un modèle de fréquence ou de prime pure, cette méthode prend son sens par la présence élevée de 0 impliquant donc des résidus strictement négatifs, car les prédictions sont strictement positives (les assurés ne peuvent pas avoir une prime égale à 0 par les principes de mutualisation et d'événement aléatoire).

Les résidus quantiles normalisés sont définis par,

$$r_i^{QR} := F^{-1}_{\mathcal{N}(0,1)}(\hat{F}_i(y_i)),$$

où $\hat{F}_i$ est la fonction de répartition estimée pour la réponse, randomisée dans le cas d'une loi discrète.

Lorsque le modèle est valide, à l'approximation de μ_i par $\hat{\mu}_i$, les résidus quantiles suivent une loi $\mathcal{N}(0,1)$. Ces résidus sont randomisés dans le cas d'une réponse de loi discrète en introduisant un bruit aléatoire supplémentaire, non-présent dans les données.

Dans la suite de l'étude, les résidus qui seront retenus afin de valider les modèles, seront les résidus quantiles présentant une distribution attendue ; une loi normale centrée et réduite.

1. $\mathbb{E}(R_i) = 0$
2. $\mathbb{V}(R_i) = 1$
3. La déviance se définit comme $D(y, \hat{\mu}) = 2(L(y; y) - L(y, \hat{\mu}))$.

3.1.4 Critères de performance

Un modèle prédictif est jugé meilleur par rapport à un autre modèle, entrainé sur le même jeu de données, au sens de critères de performance. Les critères de performance sont des mesures comparant la réalité à la prédiction. Ces critères sont utilisés sur un jeu de données indépendants de la base d'apprentissage, la base de validation ou la base de test. Les 5 critères de performance suivants seront exploités :

— la courbe lift,
— la courbe de Lorenz,
— le Gini,
— le MSE,
— la déviance.

Courbe lift

La courbe de lift est une mesure de performance d'un modèle prédictif ou descriptif. Cette courbe vérifie que les valeurs prédites et les valeurs observées respectent un ordonnancement et un ordre de grandeur similaires. Cependant, étant donné que le nombre de données est élevé, il est difficile visuellement de donner un sens à l'ordonnancement. En pratique, les données sont partitionnées en plusieurs quantiles [4] après les avoir triées selon les valeurs des prédictions, c'est-à-dire des intervalles équirépartis dans lesquels la moyenne des prédictions et la moyenne des observées sont calculées.

Ainsi, l'observation d'une tendance similaire sur les données moyennes et d'une appréciation des valeurs prises équivalentes permet respectivement d'analyser l'ordonnancement et les ordres de grandeurs. Par construction, la représentation des valeurs prédites sera croissante, mais ce n'est pas une certitude pour les valeurs observées. Dans le cadre d'un modèle parfait, les deux courbes sont superposées, ce qui signifie que les deux valeurs sont égales en moyenne.

De plus, avec la méthode de cross validation k-fold, les courbes de lift sont construites sur les données de validation, afin qu'elles reflètent les performances hors échantillon du modèle. Il est donc possible de construire un intervalle de confiance pour chaque quantile dont la partie inférieure de l'intervalle correspond aux données test du fold avec la moyenne des observés la plus basse, et vice-versa pour la partie supérieure.

Courbe de Lorenz

La courbe de Lorenz décrit la qualité des prédictions du modèle. Cette courbe est construite par le cumul des observations non normalisées par l'exposition et par le cumul des expositions. Dans les deux cas, ordonnés par les prédictions normalisées par l'exposition.

Cette représentation permet de vérifier le bon ordonnancement des valeurs prédites, de donner une mesure de la segmentation du portefeuille et d'obtenir une interprétation graphique du lien entre prédiction et observation. k courbes de Lorenz sont construites sur les données de validation à partir des k-fold formés par la cross validation, afin qu'elles reflètent les performances hors échantillon du modèle.

Il existe deux extrêmes :

— la bissectrice qui correspond à la représentation graphique d'un modèle qui n'expliquerait pas le risque, c'est-à-dire qu'aucune variable ne serait discriminante,
— la courbe optimale est celle où la valeur observée est prédite, ainsi la qualité de discrimination du modèle est maximale.

Remarque : la courbe de Lorenz peut se rapprocher de la courbe ROC en analyse binaire. D'ailleurs, en normalisant les données, $Gini = 2 \times AUC - 1$ où AUC est l'aire sous la courbe ROC et l'indice de Gini, l'aire sous la courbe de Lorenz.

Gini

L'indice de Gini est représenté par la courbe de Lorenz. Il correspond à l'aire sous la courbe définit comme,

4. 20 quantiles seront utilisés pour le reste de l'étude afin d'obtenir 20 intervalles représentant chacun 5% de l'exposition.

$$Gini = 2 \times Aire\ entre\ la\ courbe\ de\ Lorenz\ et\ la\ bissectrice\ dans\ le\ cas\ où\ les\ axes\ sont\ normés$$

De plus, le Gini peut s'écrire sous une forme normalisée,

$$Normalized\_Gini = \frac{Gini}{Gini\_max},$$

où $Gini\_max$ correspond au $Gini$ dans le cas où la prédiction est exactement égale à l'observé.

Un modèle a une qualité de prédiction meilleure par rapport à un autre au sens du Gini, si le Gini est plus élevé que le précédent. L'indice de Gini étant une aire, mathématiquement, il faut définir une intégrale pour pouvoir le calculer.

Soit $f : [a; b] \to \mathbb{R}$ une fonction définie en tout point du segment $[a; b]$. Soit $\sigma = (a = x_0 < x_1 < \cdots < x_n = b)$ une subdivision de $[a; b]$. Soit $t_1, \ldots, t_n$ des réels tels que, pour chaque i, $t_i \in [x_{i-1}, x_i]$, la somme de Riemann de f sur $[a; b]$ se définit comme,

$$S(f, \sigma, \xi) = \sum_{i=1}^{n} (x_i - x_{i-1}) f(t_i)$$

Si le pas de la subdivision tend vers zéro, alors la somme de Riemann converge vers $\int_a^b f(t)dt$. L'aire de la courbe est approximée en calculant ainsi une somme de rectangles. La courbe ajustée sur les données a généralement une tendance croissante. Il est donc possible de déterminer une aire inférieure (méthode des rectangles à gauche) et une aire supérieure (méthode des rectangles à droite) pour calculer le Gini. Cela nous permettra de nous rassurer sur le calcul de l'aire retenue (méthode du trapèze).

MSE

L'erreur quadratique de prédiction de la moyenne s'écrit,

$$MSE = \frac{1}{N} \sum_i \omega_i.(y_i - \hat{y}_i)^2$$

Le RMSE [5] est la racine carré de la moyenne de la différence au carré entre observé (y_i) et prédiction ($\hat{y}_i$) pondéré par l'exposition ω_i. Il est équivalent au log de la vraisemblance d'un modèle gaussien. Cette métrique est faussée par des valeurs extrêmes ou des erreurs. Le modèle le mieux ajusté au sens de cet indicateur est celui qui minimise cette valeur, car l'objectif est que l'erreur d'estimation soit la plus faible possible.

Une alternative au critère MSE est le Mean Relative Error (MAE),

$$MAE = \frac{1}{N} \sum_i \omega_i.|y_i - \hat{y}_i|$$

Il s'agit d'une mesure qui limite l'impact de l'erreur sur les valeurs extrêmes.

Déviance

La déviance est une mesure d'erreur. Une déviance faible signifie, au sens de ce critère, un meilleur ajustement sur les données. Cette mesure de qualité d'ajustement s'écrit,

$$D(y, \hat{\mu}) = 2(L(y; y) - L(y, \hat{\mu})),$$

où y correspond au vecteur des observations, $\hat{\mu}$ aux prédictions, $L(y, y)$ la valeur maximum de la log-vraisemblance du modèle saturé et $L(y, \hat{\mu})$ la valeur maximale de la log-vraisemblance du modèle considéré.

$D_{poisson}(y, \hat{\mu}) = 2(y.ln\frac{y}{\hat{\mu}} - (y - \hat{\mu}))$
$D_{gamma}(y, \hat{\mu}) = 2(y.ln\frac{y}{\hat{\mu}} + \frac{(y-\hat{\mu})}{\hat{\mu}})$

5. Root Mean Squared Error

3.2 Sinistres

La charge d'un sinistre est définie à partir de plusieurs éléments dont des provisions permettant de prendre en compte le coût futur des règlements à payer.

En fait, une charge est caractérisée par trois éléments,

— les *Règlements = Coût du principal+ Coût des honoraires + Recours encaissés* constatant l'indemnisation de l'assuré,
— les *Réserves = Provision des règlements – Provisions des recours*, qui prévoient une indemnisation future des sinistres non clôturés. Ces provisions correspondent aux provisions dossiers-dossiers,
— les *Frais* dont les frais de gestion ou d'expertise.

Ainsi, la charge nette de frais est égale à la somme des trois éléments qui précèdent. La responsabilité civile est une branche longue. Par conséquent, les charges sont classiquement projetées à l'aide de méthodes de provisionnement afin d'estimer la charge ultime.

La provision la plus importante en assurance non-vie est la provision pour sinistre à payer (PSAP). Il s'agit du montant intégral des dépenses nécessaires aux règlements de tous les sinistres survenus et non payés. La PSAP concilie le décalage temporaire entre la survenance du fait, le règlement effectif du sinistre, et le principe de comptabilisation par exercice.

Elle est divisée en deux provisions :

— les provisions dossier/dossier se constituent après déclaration d'un sinistre afin de couvrir les règlements attendus par l'assureur. L'évaluation de cette provision peut être faite sur-mesure en fonction des rapports d'expertise et des éléments provisoires à disposition ou à l'aide de données historiques similaires à la situation.
— les provisions IBNER (*Incurred But Not Enough Reported*) établissent l'aggravation tardive. En effet, l'assureur estime dès la déclaration du sinistre son coût probable. Cependant, il peut y avoir une aggravation intervenant plusieurs mois après la survenance. Les provisions IBNYR (*Incurred But Not Yet Reported*) constatent la déclaration tardive. Tous les sinistres ne sont pas déclarés instantanément après leur survenance. Ces deux provisions sont regroupées sous le nom d'IBNR.

Chez AXA France IARD Entreprise, le modèle de vieillissement des sinistres consiste à multiplier l'ensemble des charges des sinistrés par un coefficient tiré de la méthode Chain Ladder sur une année de survenance. Cette méthode n'apporte pas de segmentation significative du risque, elle réconcilie seulement la charge totale. Une autre méthode consiste à vieillir la charge des sinistres en cours, les autres sinistres étant clos. Les sinistres sans suite et les sinistres annulés ont été retirés de l'étude. Enfin, une dernière méthode serait de modéliser uniquement les sinistres clos dans le modèle de coût moyen. Aucune des trois méthodes précédentes n'a été retenue. En effet, la première méthode n'implique pas de segmentation suffisamment matérielle. De plus, les sinistres en cours ne représentent que 7% des sinistres dont les 2/3 sont survenus en 2019, et 7% de ces sinistres sont sous forme de coûts d'ouverture. Le tarif actuel repose sur la première méthode en admettant l'hypothèse forte que les sinistres en cours hors coûts d'ouverture suivent la même distribution que le reste de la sinistralité. Les coûts d'ouverture sont écartés de la modélisation. Ainsi, les résultats seront comparables.

Une amélioration de l'étude serait d'appliquer une méthode de vieillissement des sinistres sans suite avec une segmentation adéquate pour obtenir une modélisation de la charge ultime des sinistres.

3.2.1 Petits sinistres, forfaits IDA et forfaits AXA

Au préalable de la modélisation, la première étape est de représenter graphiquement la distribution des sinistres. Il s'agit de s'assurer que les données sinistres représentent bien le risque. Les sinistres avec recours peuvent conduire à une somme négative, nulle ou faible. De même, certains sinistres ne sont pas garantis par le contrat impliquant une indemnisation nulle, mais des frais d'expertise peuvent s'y imputer et ainsi créer de faibles charges. La base de données peut contenir des sinistres :

— **négatifs** où une mutualisation de ces sinistres sur l'ensemble du portefeuille devient indispensable, car ils ne font pas partie intégrante de la réalité de la sinistralité et sont liés à des contraintes de modélisation. En effet, la modélisation du coût moyen à partir d'un GLM demande une distribution de réels strictement positifs avec une queue de distribution épaisse (gamma, log-normale).
— **nuls** où il suffit d'affecter au nombre de sinistres la valeur 0 afin de rectifier l'erreur d'indemnisation.

— **dont le montant est faible**. Ceux-ci devront faire partie d'une étude afin de vérifier que ces coûts sont bien présents dans la base et qu'ils n'impliquent pas un poids significatif (Dirac) autour de zéro qui biaiserait le modèle. Traditionnellement, ils sont mutualisés et retirés de la modélisation, charge nulle et fréquence mise à 0, pour éviter de biaiser l'estimation de la prime pure.

D'autres sinistres doivent être considérés et donner lieu à un traitement spécifique.

Pour rappel, la convention IRSA/IRCA a pour objectif d'accélérer le processus d'indemnisation des assurés et la gestion des recours entre les assureurs. Elle se traduit par l'indemnisation par un montant de recours forfaitaire, nommé forfait IDA, établie par le GPSA[6] annuellement. Chaque assureur règle les dommages matériels du véhicule terrestre à moteur de l'assuré et procède au recours auprès de la compagnie d'assurance de l'auteur de l'accident à hauteur du montant du forfait. Ce forfait dépend du taux de responsabilité de l'assuré. Pour un accident responsable 50/50, c'est-à-dire où la cause est due aux deux conducteurs, la moitié du forfait IDA est versée. Ces montants diffèrent également selon le type de dommage : matériel ou corporel.

Ces sinistres seront supprimés de la modélisation du coût moyen pour éviter de biaiser la prime pure. En revanche, ils feront l'effet d'une mutualisation sur la prime pure, modélisée par un montant forfaitaire en supposant que le coût moyen des sinistres à coût forfaitaire suivent la même distribution que le reste de la sinistralité.

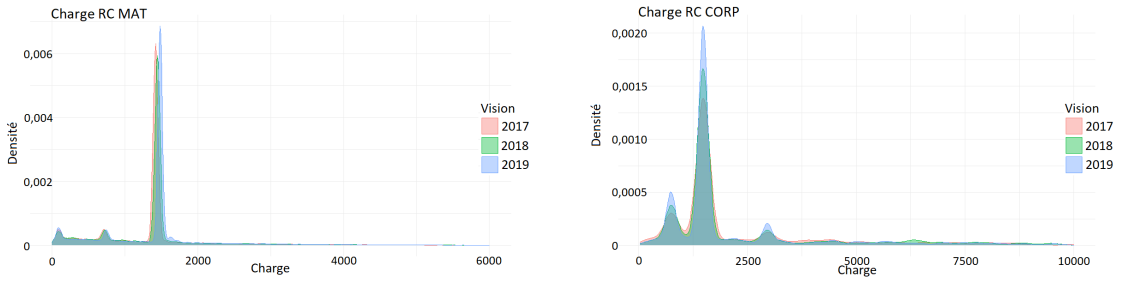


FIGURE 3.1 – Densité de la charge RC selon le type de dommage

Les forfaits IDA ne représentent pas le risque réel et créent des "pics" de charges quelque soit l'année de vision et le type de dommage. Cela explique la nécessité de mutualiser ces charges. De plus, des erreurs subsistent sur les forfaits IDA.

En effet, pour certains sinistres, le forfait IDA est versé deux fois voire de deux à quatre fois pour la RC CORP. Aucun forfait IDA avec une part de responsabilité 75% ou 25% n'est présent dans le jeu de données.

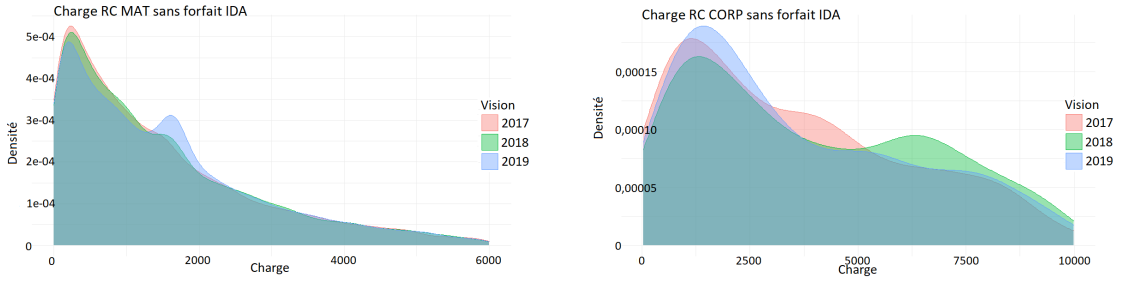


FIGURE 3.2 – Densité de la charge RC exclue des forfaits IDA selon le type de dommage

L'exclusion des forfaits IDA permet de nettoyer la distribution de la charge. En revanche, on remarque un "pic" de la charge RC MAT pour l'année 2019 liée aux coûts d'ouverture.

6. Gestion Professionnelle des Services de l'assurance

Les coûts d'ouverture correspondent à une provision au moment où il y a une connaissance partielle des circonstances du sinistre. Ces coûts d'ouvertures sont définis par l'assureur. AXA a défini un montant forfaitaire par garantie évoluant chaque année. Ces coûts d'ouverture sont utilisés à la discrétion du gestionnaire de sinistre qui évalue, au vu des informations dont il dispose, la gravité du sinistre et décide d'affecter ce forfait ou non.

Ainsi, ces coûts d'ouverture sont des éléments de preuve pour évaluer la maturité des sinistres. La distribution de ces sinistres résulte en une somme de plusieurs Dirac qui ne sont pas forcément représentatifs de la distribution finale.

Aucun coût d'ouverture n'est recensé pour la RC CORP. En revanche, pour la RC MAT, plusieurs coûts d'ouverture sont présents, cela est frappant pour l'année 2019. Il a été décidé de simplement supprimer ces coûts d'ouvertures de la modélisation et de les mutualiser au même titre que les charges négatives ou les forfaits IDA.

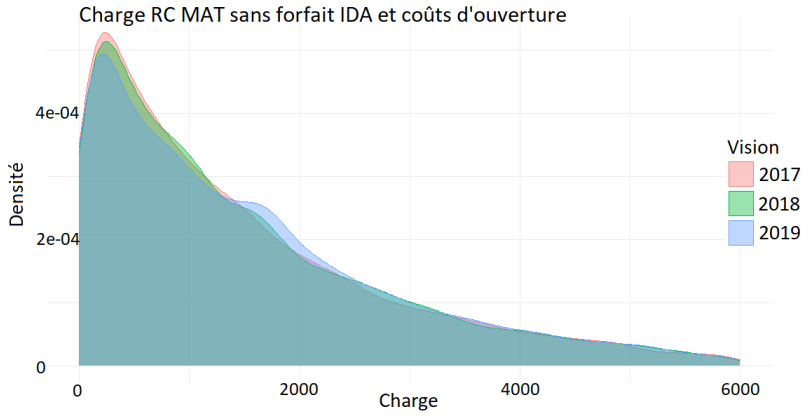


FIGURE 3.3 – Densité de la charge RC MAT exclue des forfaits IDA et AXA

Un dernier point qui n'a pas été abordé est la conséquence des faibles montants de sinistres. En zoomant sur ces faibles montants de sinistres, il est facile de constater la présence d'une sur-sinistralité due aux frais d'expertise. Il est difficile de faire un choix à ce stade sur la possibilité de suppression des sinistres de montant compris entre 0 et 120 euros (Figure 3.4). La décision sera faite en comparant les modèles dans le cas où ces sinistres sont retirés et dans le cas où ces sinistres sont conservés.

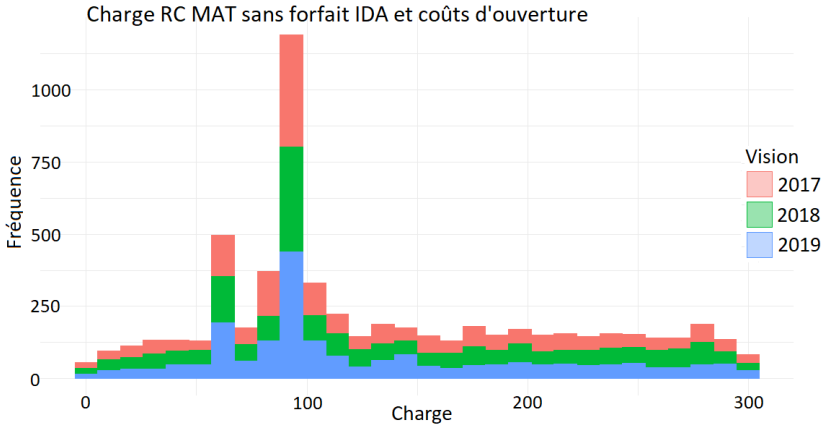


FIGURE 3.4 – Histogramme de la charge RC MAT exclut des forfaits IDA et AXA en zoomant sur les petites charges

3.2.2 Théorie des valeurs extrêmes

Au sein d'AXA France Entreprise, des seuils ont été prédéfinis afin de distinguer la sinistralité attritionnelle, la sinistralité grave et la sinistralité atypique. Ces seuils permettent une harmonisation sur les

communications au sein de l'entité. Pour la branche auto, ces seuils sont définis de la façon suivante,

— **sinistres attritionnels** : il s'agit des sinistres dont la charge est inférieure à 30 000€. Ils représentent la grande majorité de la sinistralité,
— **sinistres graves** : il s'agit des sinistres dont la charge dépasse 30 000€. Ces sinistres sont peu fréquents et ont une charge élevée. La particularité de ces montants présente une distribution spécifique qui les distingue des autres,
— **sinistres atypiques** : il s'agit de sinistres dont la charge dépasse $2M$ d'euros. Ces sinistres sont rares et ont une charge hors norme. Ainsi, il est difficile d'ajuster un modèle paramétrique sur ces sinistres. De plus, les sinistres atypiques ont une tendance à fortement biaiser la modélisation, entraînant des conséquences lourdes sur le montant de la prime. C'est la responsabilité civile dommages corporels qui est de façon générale fortement impactée par cette sinistralité.

Le seuil des graves a été défini indépendamment de la garantie auto concernée et est limité à un montant de 30 000€ quel que soit le type de dommages à l'origine de la responsabilité civile (corporel ou matériel). Ainsi, ce seuil n'est certainement pas le plus adapté à l'étude.

Les sinistres en-dessous d'un certain seuil suivent probablement une distribution connue telle qu'une distribution gamma pour le coût moyen. Cependant, au-dessus de ce seuil, les sinistres peuvent suivre une distribution spécifique, telle une loi de Pareto. Cet effet a pour conséquence de biaiser notre modélisation (résidus asymétriques).

C'est dans ce contexte que la théorie des valeurs extrêmes est utilisée. Celle-ci permet de définir un seuil à partir duquel les sinistres suivent une autre distribution. Il doit être suffisamment grand pour permettre une distinction optimale des distributions et d'éviter de mettre de côté une trop grande quantité de sinistres, mais suffisamment petit afin d'éviter un biais trop important dans la modélisation. Ce choix nécessite un arbitrage biais-variance. L'objectif final est de comparer les seuils obtenus par la théorie des valeurs extrêmes au seuil défini par AXA afin de valider la séparation des charges graves et des charges attritionnels.

Domaines d'attraction des valeurs extrêmes

Soit $X_1, \ldots, X_n$ un échantillon de n variables aléatoires indépendantes et identiquement distribuées de fonction de répartition commune F. On définit $X_{n,n} = max(X_1, \ldots, X_n)$ alors $\mathbb{P}(X_{n,n} \leqslant x) = F^n(x)$.

Si on note $x_F = inf\{x; F(x) = 1\}$ le point extrémal de F, on montre que $X_{n,n} \xrightarrow[n \to +\infty]{P} x_F$. Pour obtenir une convergence en loi, il faut donc normaliser le maximum.

Définition

Soit H une fonction de répartition non-dégénérée, c'est-à-dire une fonction de répartition qui n'est pas associée à une variable constante presque sûrement. On dit que F appartient au domaine d'attraction de H, noté $F \in \mathcal{DA}(H)$, si il existe deux suites $a_n > 0$ et b_n telles que en tout point de continuité x de H.

$$\lim_{n \to +\infty} F^n(a_n x + b_n) = H(x) \Leftrightarrow \frac{X_{n,n} - b_n}{a_n} \xrightarrow{\mathcal{L}} Y,$$

où la variable aléatoire Y admet H pour fonction de répartition.

Proposition

Soient H et G deux fonctions de répartition non-dégénérées.

Si $F \in \mathcal{DA}(H)$, $\exists\ a_n > 0$ et b_n telle que $F^n(a_n x + b_n) \to H(x)$, et si $F \in \mathcal{DA}(G)$, $\exists\ u_n > 0$ et v_n telle que $F^n(u_n x + v_n) \to G(x)$, alors les fonctions de répartition de H et G sont de même type[7]. De plus, on a $u_n/a_n \to a > 0$ et $(v_n - b_n)/a_n \to b$.

Cette proposition montre que le domaine d'attraction d'une fonction de répartition F est unique à un changement de paramètres d'échelle et de position près.

Théorème (Fisher-Tippett)

Si $F \in \mathcal{DA}(G)$ où G est une fonction non-dégénérée, alors G est de même type que l'une des trois distributions suivantes :

7. Soient H et G deux fonctions de répartition non-dégénérées. On dit que H et G sont de même type si $\exists\ a > 0$ et b telles que pour tout $x \in \mathbb{R}$, $H(x) = G(ax + b)$.

En d'autres termes, si $X \sim H$ et $Y \sim G$ alors, si H et G sont de même type, $Y = aX + b$ presque sûrement.

$$\textbf{Fréchet } (\alpha > 0) : \phi_\alpha(x) = \begin{cases} 0 & \text{si } x \leqslant 0 \\ exp(-x^{-\alpha}) & \text{si } x > 0 \end{cases}$$
$$\textbf{Weibull } (\alpha > 0) : \psi_\alpha(x) = \begin{cases} exp(-(-x)^{\alpha}) & \text{si } x \leqslant 0 \\ 1 & \text{si } x > 0 \end{cases}$$
$$\textbf{Gumbel } (\alpha = 0) : \Lambda_\alpha(x) = exp(-e^{-x}) \;\; x \in \mathbb{R}$$

Von Mises (1945) et Jenkinson (1955) ont proposé une famille paramétrique de distribution, appelée la distribution généralisée des valeurs extrêmes, noté $GEV(\mu, \sigma, \gamma)$. Cette famille permet d'unifier les trois types de distributions des valeurs extrêmes et en facilite également l'estimation. Le résultat fondamental de la théorie des valeurs extrêmes est à rapprocher du théorème central limite et permet de caractériser la loi de distribution des extrêmes.

Théorème fondamental de la théorie des valeurs extrêmes
Si $F \in \mathcal{DA}(G)$ où G est une fonction non-dégénérée, alors G est de même type que la fonction de répartition définie pour tout x tel que $1 + \gamma x > 0$ par :

$$H_\gamma(x) = \begin{cases} exp(-[1+\gamma(\frac{x-\mu}{\sigma})]^{-1/\gamma}) & \text{si } \gamma \neq 0 \\ exp(-exp[-(\frac{x-\mu}{\sigma})]) & \text{si } \gamma = 0 \end{cases},$$

où γ est le paramètre de forme, μ le paramètre de position et σ le paramètre d'échelle.

La loi limite du maximum dépend donc du seul paramètre γ appelé l'indice des valeurs extrêmes. La loi associée à la fonction de répartition $H_\gamma(x)$ est appelée *Generalized Extreme Values*, noté GEV. On définit trois types de domaines d'attraction selon le signe de γ :

— **Domaine de Fréchet** : Si $\gamma > 0$, la fonction de survie $1-F(x)$, converge vers 0 lorsque $x \to x_F^* = +\infty$ à une vitesse polynomiale (proportionnelle à une puissance de x). Ce domaine d'attraction regroupe les distributions *heavy-tailed* (à queue lourde).
— **Domaine de Gumbel** : Si $\gamma = 0$, la fonction de survie $1 - F(x)$, converge vers 0 lorsque $x \to x_F^* = +\infty$ à une vitesse exponentielle. On parle ici de distribution light-tailed (à queue légère).
— **Domaine de Weibull** : Si $\gamma < 0$, la fonction de survie $1 - F(x)$, converge vers 0 lorsque $x \to x_F^* = +\infty$ à une vitesse polynomiale. Ce domaine regroupe la plupart des distributions ayant un point terminal fini.

Remarque : il apparait que si deux fonctions de survie sont asymptotiquement proportionnelles, elles appartiennent au même domaine d'attraction avec le même indice des valeurs extrêmes.

Le tableau suivant (Figure 3.5) regroupe quelques lois usuelles classées en fonction de leur domaine d'attraction :

Gumbel	Fréchet	Weibull
Normale	Pareto	Uniforme
Exponentielle	Log-gamma	Beta
Log-Normale	Student	
Gamma		
Weibull		

FIGURE 3.5 – Lois usuelles par domaine d'attractions

Estimateurs du seuil

Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes et de même fonction de répartition F. Pour une suite (α_n) qui converge vers 0 lorsque n converge vers l'infini, on souhaite estimer le quantile extrême d'ordre α_n, c'est-à-dire la quantité $q(\alpha_n) = \bar{F}^{\leftarrow}(\alpha_n)$. [8]

La valeur $q(\alpha_n)$ est située dans la queue de distribution de F et ne peut pas être estimée par l'estimateur empirique classique $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \leqslant x}$. Le théorème fondamental des valeurs extrêmes et la loi

8. F est une fonction croissante et continue à droite sur $\mathbb{R}$. L'inverse généralisé de F est défini par $F^y := inf\{x | y \leqslant F(x)\}$

de Pareto généralisée permettent d'accéder à une expression approximée de $q(\alpha)$ de la même forme[9]. La différence entre ces deux approximations réside donc dans la méthode utilisée pour estimer les paramètres inconnus.

Remarque : l'utilisation de la loi GPD pour approximer γ se justifie par le théorème de Pickands qui annonce que pour un seuil u proche du point terminal de x_F de F. La fonction de répartition des excès définie pour $x > 0$ par $F_u(x) := \mathbb{P}(X - u \leqslant x | X > u) = 1 - \frac{\bar{F}(x+u)}{\bar{F}(u)}$ peut être approchée par la loi GPD de fonction de répartition $G_{\gamma,a(u)}(x)$ où a est une fonction positive.

La loi de Pareto Généralisée de paramètres $\gamma \in \mathbb{R}$ et $\sigma > 0$ est définie par sa fonction de répartition donnée par :

$$G_{\gamma,\sigma}(x) = 1 + logH_\gamma(\frac{x}{\sigma}) = \begin{cases} 1 - (1 + \gamma x/\sigma)^{-1/\gamma} & \text{si } \gamma \neq 0 \\ 1 - exp(-x/\sigma) & \text{si } \gamma = 0 \end{cases},$$

pour $x \in \{t \in \mathbb{R}; 1 + \gamma t/\sigma > 0\} \cap [0; \infty[$.

Par la suite, différents estimateurs de γ seront décrits, le paramètre de forme de la loi de Pareto Généralisée, car celui-ci dépend d'une suite intermédiaire que l'on veut déterminer. Cette suite permet de déterminer un seuil u distinguant les sinistres graves des sinistres hors graves.

Estimateur de Hill

Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes et de même fonction de répartition F appartenant au domaine d'attraction de Fréchet avec un indice des valeurs extrêmes $\gamma > 0$. L'estimateur de Hill (1975) est :

$$\hat{\gamma}_n^{(H)}(k_n) := \frac{1}{k_n} \sum_{i=1}^{k_n} log(X_{n-i+1,n}) - log(X_{n-k_n,n}),$$

où $1 \leqslant k_n \leqslant n$ est une valeur à choisir par l'utilisateur.

Cet estimateur repose sur le cas particulier du domaine d'attraction de Fréchet où $q(\alpha)$ peut être approximé par $q(\alpha) \approx q(t\alpha)t^\gamma$. Pour estimer $q(\alpha)$, il suffira donc d'estimer γ et $q(t\alpha)$. L'estimation de $q(\alpha)$ requiert l'approximation de $q(t\alpha)$ et il semble que le problème soit simplement déplacé. L'idée sera de prendre t suffisamment grand de telle sorte que $t\alpha$ soit un ordre de quantile classique que l'on pourra approximer par la fonction de répartition empirique $\hat{F}$. En pratique, on choisit $t = t_n = k_n/(n\alpha_n)$ où (k_n) est une suite intermédiaire. On a ainsi $q(t\alpha_n) = q(k_n/n)$ que l'on peut estimer en inversant la fonction de répartition empirique par la statistique d'ordre $X_{n-k_n,n}$. Il reste ensuite à proposer un estimateur $(\hat{\gamma}_n)$ pour estimer $\hat{q}(\alpha_n) = X_{n-k_n,n}(\frac{k_n}{n\alpha_n})^{\hat{\gamma}_n}$.

Un de ces estimateurs est l'estimateur de Hill dérivant de la méthode du maximum de vraisemblance[10], l'idée de départ étant de supposer que les $X_1, \ldots, X_n$ suivent une loi de Pareto de paramètres $c > 0$ et $\gamma > 0$. En prenant les k_n plus grandes observations de l'échantillon pour calculer l'estimateur $\hat{\gamma}$, on obtient l'expression de l'estimateur de Hill.

Le choix de k_n, c'est à dire le nombre d'observations considéré, est crucial. Si k_n est trop grand, alors l'approximation par une loi de Pareto sera mauvaise et l'estimateur de Hill aura un biais important. À l'inverse, si k_n est trop petit, il y a aura trop peu de d'observations et donc l'estimateur aura une variance importante. En théorie, $k_n \to \infty$ et $k_n/n \to 0$. En pratique, un *Hill Horror Plot* est utilisé, une méthode graphique où le meilleur choix de k_n se situe dans la zone de stabilité.

Le dAMSE est une alternative à l'*Hill Horror Plot* qui donne une estimation du seuil en minimisant le critère AMSE[11].

Remarque : c'est le Hill Horror Plot qui nous intéresse, car en déterminant le point de stabilité de k_n, le seuil u sera déterminé et donc le nombre d'observations à partir duquel on considère que l'estimateur

9. Dans le cas du théorème fondamental, on a $q(\alpha) \approx b_k + \frac{a_k}{\gamma}[(k\alpha)^{-\gamma} - 1]$ où (a_k) et (b_k) découlent du théorème. Dans le cas de la GPD, on a $q(\alpha) \approx u + \frac{a(u)}{\gamma}[(\frac{\alpha}{\bar{F}(u)})^{-\gamma} - 1]$ où $a(u)$ découle de la loi GPD.

10. Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes et identiquement distribuées suivant une loi de paramètre θ et $x_1, \ldots, x_n$ leurs réalisations. La vraisemblance se définit alors comme $L(\theta|x_1, \ldots, x_n) = \prod_{i=1}^n L(\theta|x_i)$ avec pour tout $i \in [\![1; n]\!]$, $L(\theta|x_i) = f_\theta(x)$ si la loi des X_i est continue de densité de probabilité f sinon si la loi est discrète alors $L(\theta|x_i) = \mathbb{P}_\theta(X_i = x_i)$. En général, on est amené à maximiser la log-vraisemblance en annulant sa dérivée lorsque cela est possible.

11. *Average Mean Squared Error* par rapport à k_n pour l'estimateur de Hill. Le critère AMSE correspond à une estimation de l'erreur quadratique moyenne.

de Hill estime le plus justement $q(\alpha_n)$. Ainsi, à ce nombre d'observations, on peut y associer un seuil qui sera le seuil des graves.

Estimateur de DEdH

L'estimateur a été introduit en 1989 par Dekkers, Einmahl et De Hann comme une généralisation de l'estimateur de Hill quelle que soit la valeur de γ. Cet estimateur est défini par la statistique :

$$\hat{\gamma}_n^{(DEdH)}(k_n) := \hat{\gamma}_n^{(H)(1)}(k_n) + 1 - \frac{1}{2}[1 - \frac{(\hat{\gamma}_n^{(H)(1)}(k_n))^2}{\hat{\gamma}_n^{(H)(2)}(k_n)}]^{-1},$$

avec pour $r = 1$ ou 2, $\hat{\gamma}_n^{(H)(r)}(k_n) = \frac{1}{k}\sum_{i=1}^{k_n}(lnX_{n-i+1,n} - lnXn - k_n, n)^r$

Estimateur de Pickands

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

$$\hat{\gamma}_n^{(P)}(k_n) := \frac{1}{log(2)} log \frac{X_{n-k_n/4,n} - X_{n-k_n/2,n}}{X_{n-k_n/2,n} - X_{n-k_n,n}}$$

Il est possible de montrer que sous certaines hypothèses l'estimateur de Pickands est un estimateur consistant de γ. En pratique, il est très sensible à la taille de l'échantillon, ce qui peut le rendre peu robuste, mais il peut être appliqué quel que soit le domaine d'attraction des extrêmes.

Cet estimateur repose sur l'approximation par la loi GPD [12]. On suppose que les excès $Y_1 = X_{n-k_n+1,n} - X_{n-k_n,n}, \ldots, Y_{k_n} = X_{n,n} - X_{n-k_n,n}$ sont indépendantes et de même loi de Pareto généralisée (ce qui n'est pas le cas). En notant, $\hat{\gamma}(k_n)$ et $\hat{a}(k_n)$ des estimateurs de γ et $a(k_n)$ obtenues à partir de l'échantillon des excès, on estimera donc $q(\alpha_n)$ par :

$$\hat{q}(\alpha_n) = X_{n-k_n+1,n} + \frac{\hat{a}(k_n)}{\hat{\gamma}(k_n)}[(\frac{n\alpha_n}{k_n})^{-\hat{\gamma}(k_n)} - 1]$$

L'estimateur de Pickands est un estimateur de γ utilisant la méthode percentile [13] fonctionnant quelle que soit la valeur de γ. Le choix de k_n se fait de la même façon que l'estimateur de Hill en cherchant le premier point de stabilité de l'*Hill Horror Plot* avec l'estimateur de Pickands.

3.2.3 Application de la TVE

Détermination du domaine d'attraction

La théorie des valeurs extrêmes suppose des sinistres indépendants et identiquement distribués. Un sinistre RC *a priori* respecte l'hypothèse d'indépendance même en cas de responsabilité partagée à 50/50, car un seul sinistre sera ouvert. Toutefois, l'hypothèse d'observations identiquement distribuées est davantage sujette à caution. Pour la responsabilité civile dommages matériels (RC MAT) et la responsabilité civile dommages corporels (RC CORP), il est raisonnable d'accepter l'hypothèse. En revanche cela est moins vrai pour la responsabilité civile (RC) qui est l'addition de la RC MAT et la RC CORP. En effet, la RC CORP *a priori* a une distribution différente de la RC MAT. Les sinistres RC CORP sont rares, le nombre d'échantillons n est faible, c'est pourquoi il est malgré tout préférable de considérer la RC. Dans l'étude qui suit, nous déterminerons un seuil pour la RC, la RC CORP et la RC MAT.

L'estimateur de Hill impose un domaine d'attraction de Fréchet. Ainsi, il est nécessaire de déterminer le paramètre de forme de la loi GEV ou de la loi GPD. De plus, la détermination du domaine d'attraction permet d'avoir une idée de la distribution spécifique des sinistres graves. Ainsi, la détermination d'un nouveau seuil est intéressante si les données appartiennent au domaine de Fréchet. En effet, pour une distribution à queue épaisse, le seuil distingue les sinistres graves, qui pour un faible nombre, représentent un poids important de la charge globale. Pour rappel, le paramètre de forme γ de la loi GEV doit être strictement supérieur à 0 pour considérer que les données appartiennent au domaine de Fréchet. Dans le cadre de l'étude, nous déterminerons un seuil pour la responsabilité civile sur les dommages matériels (RC MAT), la responsabilité civile des dommages corporels (RC CORP) et la responsabilité civile (RC) qui est simplement l'addition des deux sous-risques précédemment cités. Dans le tableau suivant, il est

12. Rappel : Pour un seuil u proche du point terminal x_F de F, on peut approcher $q(\alpha_n)$ avec $\alpha_n \to 0$ par $u + \frac{a(u)}{\gamma}[(\frac{\alpha}{\bar{F}(u)})^{-\gamma} - 1]$.

13. La méthode des percentiles est basé sur l'approximation $\bar{F}(Z_i, k) \approx \mathbb{E}[\bar{F}](Z_{I,k}] = \mathbb{U}_{\mathbb{k}-\mathbb{i}+\mathbb{1},\mathbb{k}} = 1 + \frac{i}{k+1}$.

indiqué l'estimation du paramètre de forme de la distribution EVD de la loi GEV par la méthode du maximum de vraisemblance.

	γ	Signe	Domaine
RC	0,82	γ > 0	Fréchet
RC CORP	1,11	γ > 0	Fréchet
RC MAT	0,71	γ > 0	Fréchet

FIGURE 3.6 – Détermination du paramètre de forme

Les résultats de l'estimation (Figure 3.6) soutiennent l'appartenance des données au domaine d'attraction de Fréchet. Ces estimations découlent de l'approximation :

$$\mathbb{P}[max(X_1, \ldots, X_k) \leqslant x] \approx H_y(\frac{x - b_k}{a_k}),$$

qui est d'autant meilleur que k grand ($k = 100$ ici). Les paramètres a_k, b_k et γ sont estimés à l'aide d'un échantillon de maxima de k variables aléatoires indépendantes de fonction de répartition F, dites méthodes des *maxima par blocs*. L'échantillon $X_1, \ldots, X_n$ est divisé en $m = [n/k]$ sous-échantillons de taille k assez grand et ainsi m maxima $M_1, \ldots, M_m$ seront utilisées pour déduire a_k, b_k et γ. Cette méthode requiert un grand nombre de données, ce qui est le cas.

Un moyen de valider ces résultats est de faire appel au Quantile plot généralisé. C'est une approche permettant de faire abstraction du choix *a priori* du domaine d'attraction. Il s'agit d'un graphique qui permet visuellement de déterminer le signe de γ. Dans la théorie des extrêmes, le *QQ-plot*, sous l'hypothèse d'une distribution exponentielle, est la représentation des quantiles empiriques de l'UH scores sur l'axe des Y contre les quantiles de la fonction de distribution exponentielle sur l'axe des X. Ainsi, le graphe est défini comme :

$$(log\frac{n+1}{j+1}, log\gamma_n^{UH}(j)),$$

où $\gamma_n^{UH}(k_n) = X_{n-k_n,n}\gamma_n^{(H)}(k_n)$, avec $\gamma_n^{(H)}(k_n)$ l'estimateur de Hill.

Selon le signe de γ [14], il est attendu visuellement :

En pratique, nous observons :

Quelles que soient les données représentées [15], le quantile plot généralisé soutient que le signe de γ est strictement positif et donc que les données étudiées appartiennent au domaine d'attraction de Fréchet. Au vu des résultats, il serait intéressant de déterminer un seuil de grave. Précédemment, on a vu que théoriquement le seuil k_n n'est pas fixé. En revanche, il est possible de déterminer ce seuil avec des méthodes graphiques par un arbitrage biais-variance (*Hill Horror Plot*). Le seuil doit être - ni trop petit - pour ne pas considérer à tort des graves comme des sinistres non graves et ne pas mutualiser de trop les données sinistres - ni trop grand - pour bien faire la distinction existante de la distribution des graves qui biaise la modélisation du coût moyen.

14. Pour rappel : $\gamma > 0$ (resp. < 0 et $= 0$) implique un domaine de Fréchet (resp. domaine de Weibull et domaine de Gumbel).
15. La RC CORP et la RC MAT ont été mises en annexe.

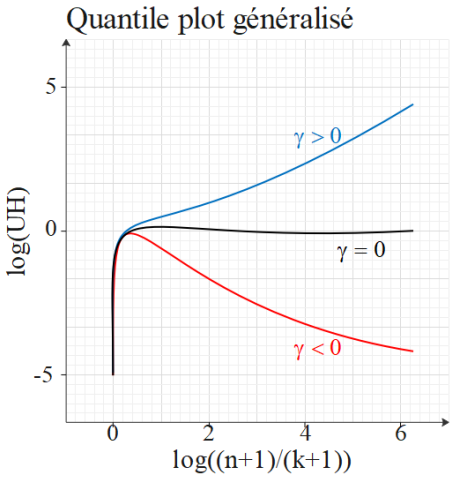


FIGURE 3.7 – Quantile plot généralisé attendu selon la valeur de γ

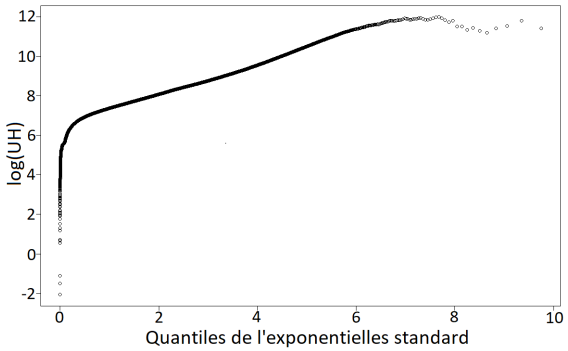


FIGURE 3.8 – Quantile plot généralisé de la RC

Remarque : l'estimateur de Pickands considère que les excès se comportent comme une distribution de Pareto Généralisé ainsi, il devient crucial de définir un seuil qui serait statistiquement convenable.

Application des estimateurs

Estimateur de Hill

Le graphique *Hill Horror Plot* représente l'évolution des estimateurs selon différentes valeurs de k_n, le nombre d'excès ou aussi communément appelé le nombre de statistiques d'ordre. L'estimateur de Hill est sensible à la taille de l'échantillon n et est en général volatile lorsque k est faible puis se stabilise. k_n est déterminé graphiquement lorsque l'estimateur de Hill se stabilise. Sous des hypothèses supplémentaires sur k_n et sur la fonction de répartition, l'estimateur de Hill est asymptotiquement gaussien, ce qui conduit à tracer un intervalle de confiance sur le graphe :

$$IC_{95\%}(\gamma) = [\gamma_n^{(H)}(k_n) - 1,96.\frac{\gamma_n^{(H)}(k_n)}{\sqrt{k_n}}; \gamma_n^{(H)}(k_n) + 1,96.\frac{\gamma_n^{(H)}(k_n)}{\sqrt{k_n}}],$$

c'est un outil visuel supplémentaire à la détection de la zone de stabilité. La représentation graphique de l'estimateur de Hill sur la RC est tracée ci-dessous (Figure 3.9), la RC CORP et la RC MAT seront mises en annexes.

Les traits verticaux représentent des seuils candidats, en rouge un intervalle où l'on repère une zone de stabilité, en bleu la valeur du dAMSE et en gris le seuil retenu pour l'estimateur. L'intérêt de déterminer plusieurs seuils candidats est de mesurer la sensibilité de la charge (indiquée sur le graphique) associée aux seuils k_n (indiqués en bas du graphique) et de vérifier par la suite que les trois estimateurs renvoient à peu près les mêmes résultats. Ainsi, un seuil à 20 000€ semblerait séparer les sinistres graves des sinistres attritionnels. De plus, l'estimateur de Hill pour $k_n = 400$ est environ égal à 1 ce qui est strictement supérieur à 0 et ainsi nous réconforte sur le domaine d'attraction précédemment relevé.

La zone de stabilité est déterminée à la main. En effet, il n'existe pas de réelles méthodes pour déterminer un seuil unique et son intervalle de confiance. Plusieurs approches ont été produites, mais celles-ci n'ont pas mené à des résultats probants à cause de la forte variabilité des résultats. Ainsi, la zone de stabilité est définie en respectant trois critères :

- elle doit être suffisamment dispersée pour connaître la sensibilité du seuil. Le dAMSE doit être inclut dans l'intervalle, car il s'agit d'une valeur en général utilisée par les réassureurs lors d'études sollicitant un minimum d'arbitrage,
- cette zone doit être relativement stable, c'est-à-dire que l'estimateur est à peu près constant,
- l'importance d'utiliser plusieurs estimateurs réside dans la définition d'une zone de stabilité qui soit similaire entre chacun d'entre eux.

Estimateur de DEdH

L'estimateur de DEdH (Figure 3.10) se stabilise à partir du 600e excès, soit à un seuil d'environ 17 000€.

Cet estimateur est plus stable que l'estimateur de Hill et est également proche de 1 pour l'intervalle de stabilité qui nous intéresse.

Estimateur de Pickands

L'estimateur de Pickands (Figure 3.11) semble être plus volatile que les deux précédents estimateurs lorsque k est faible. La stabilisation commence à partir de la 620e statistique d'ordre pour la RC, ce qui représente tous les sinistres ayant une charge strictement supérieure à 17 000€.

Remarque : l'estimateur de Pickands donne une estimation de γ relativement différente des deux autres estimateurs. En fait, pour un seuil à peu près équivalent, on retrouve une estimation de γ environ égale à 1 pour l'estimateur de Hill et de DEdH en revanche égale à $0,5$ pour l'estimateur de Pickands.

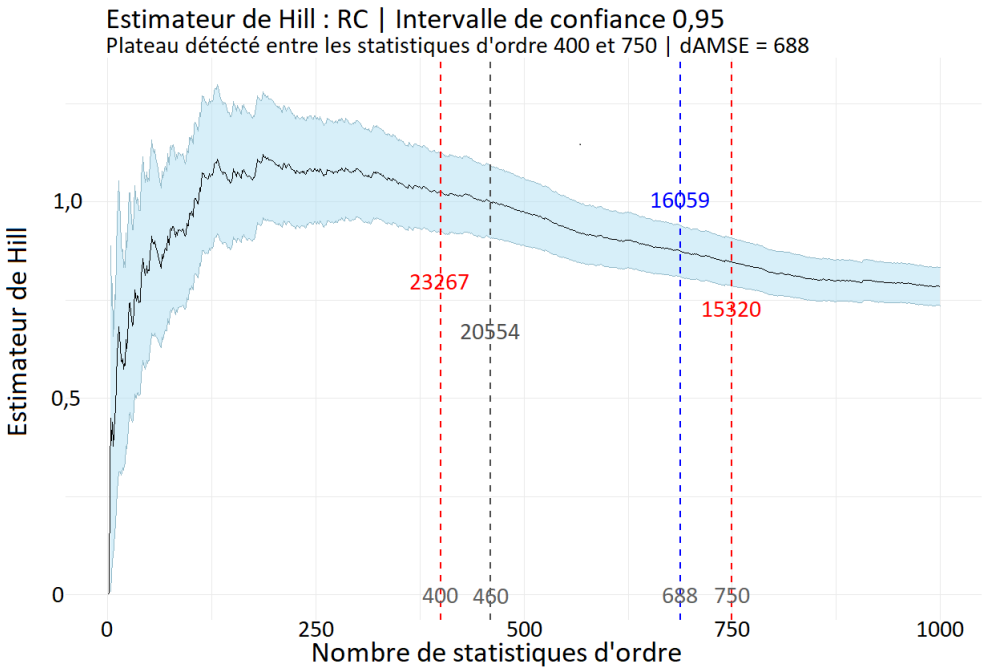


FIGURE 3.9 – Evolution de l'estimateur de Hill selon k_n

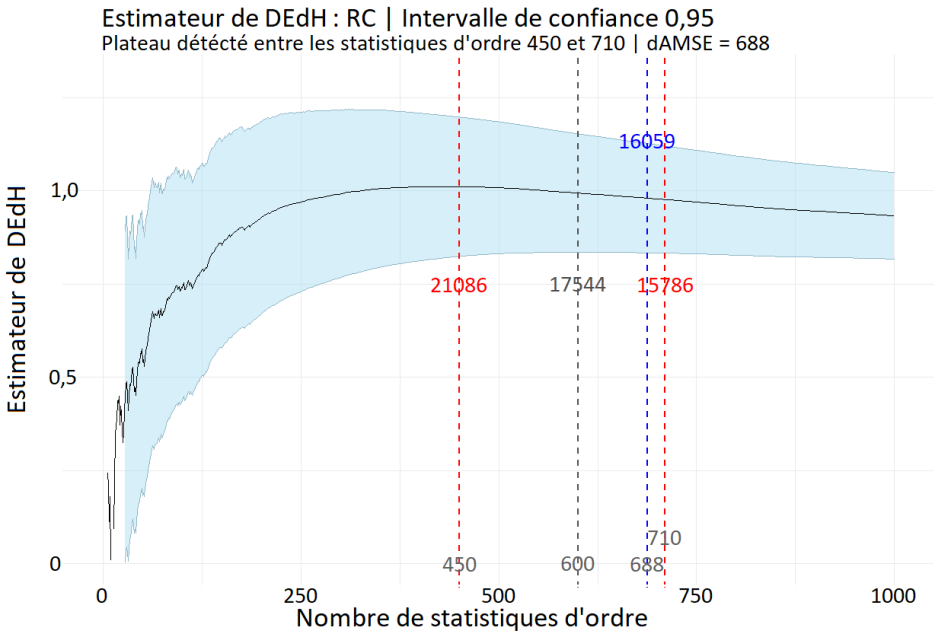


FIGURE 3.10 – Evolution de l'estimateur de DEdH selon k_n

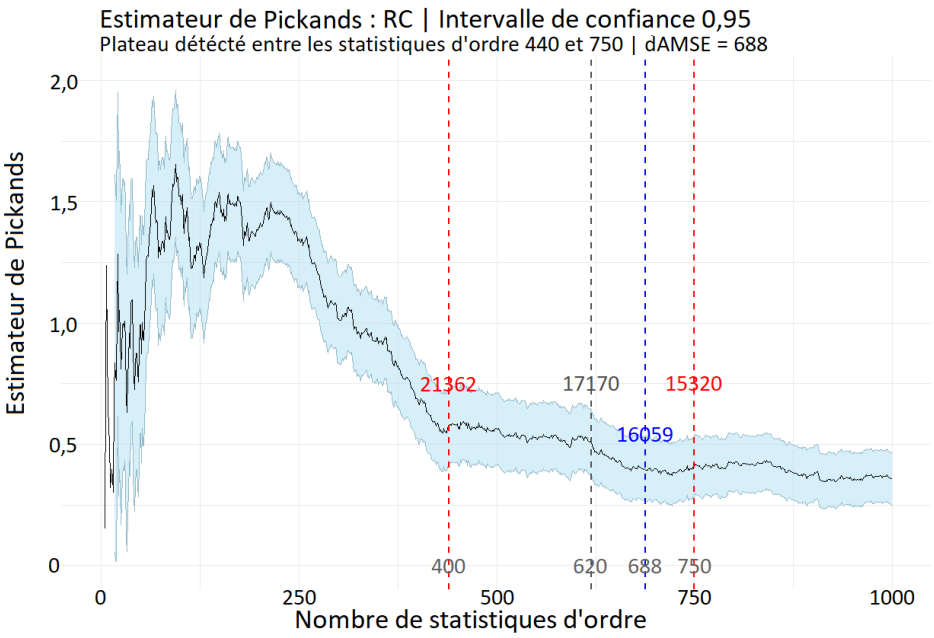


FIGURE 3.11 – Evolution de l'estimateur de Pickands selon k_n

Seuils des graves retenus

L'ensemble des seuils candidats (Figure 3.12) peut être recensé dans un même tableau :

	RC			RC CORP			RC MAT		
	Hill	Pickands	DEdH	Hill	Pickands	DEdH	Hill	Pickands	DEdH
Seuil 1	23 267	21 362	21 086	57 905	74 424	63 328	15 026	14 415	14 415
Seuil 2	20 654	17 170	17 544	44 898	44 898	44 898	14 415	11 862	13 163
Seuil 3	15 320	15 320	15 786	35 198	29 463	39 781	12 188	11 124	12 188
dAMSE	16 059			50 234			12 481		

Seuils retenus	18 000	50 000	13 000

FIGURE 3.12 – Tableau des seuils des graves retenus

On remarque une faible variation des seuils candidats pour la RC MAT, mais une forte variation pour la RC liée à la RC CORP. La variation importante des seuils peut s'expliquer par le faible nombre de sinistres de la RC CORP et la grande variation de la charge des sinistres, ainsi entre le 120e excès et le 250e excès de la RC CORP il y a une différence de charge d'environ 35 000€.

Ci-dessous (Figure 3.13), nous retrouvons le nombre de sinistres selon le risque étudié et la charge de la sur-crête ou sous-crête, ainsi que de la fréquence des seuils retenus.

	Seuils retenus	% Charge sous crête	% Charge sur crête	% Nombre sur crête	Nombre sinistres
RC	18 000	55%	45%	1.7%	34 397
RC CORP	50 000	22%	78%	9.4%	1 724
RC MAT	13 000	87%	13%	1.3%	33 281

FIGURE 3.13 – Tableau de la charge et la fréquence des seuils retenus

Pour la RC, les sinistres au-dessus des seuils définis représentent bien un faible pourcentage de l'ensemble des sinistres, environ 2%, mais un pourcentage important de l'ensemble de la charge, environ 45%.

Ces seuils sont donc efficaces et les sinistres graves ont une distribution spécifique du reste des sinistres, il serait donc à priori intéressant de mutualiser les graves.

Il est d'ailleurs intéressant de signaler que trois sinistres de la RC CORP représentent environ $6,7\%$ de la charge RC. Ces sinistres nommés atypiques dépassent les $2M$ d'euros, un seuil défini par AXA, et sont directement redistribués à l'aide du coefficient PLR [16] et donc mises de côté lors de la modélisation.

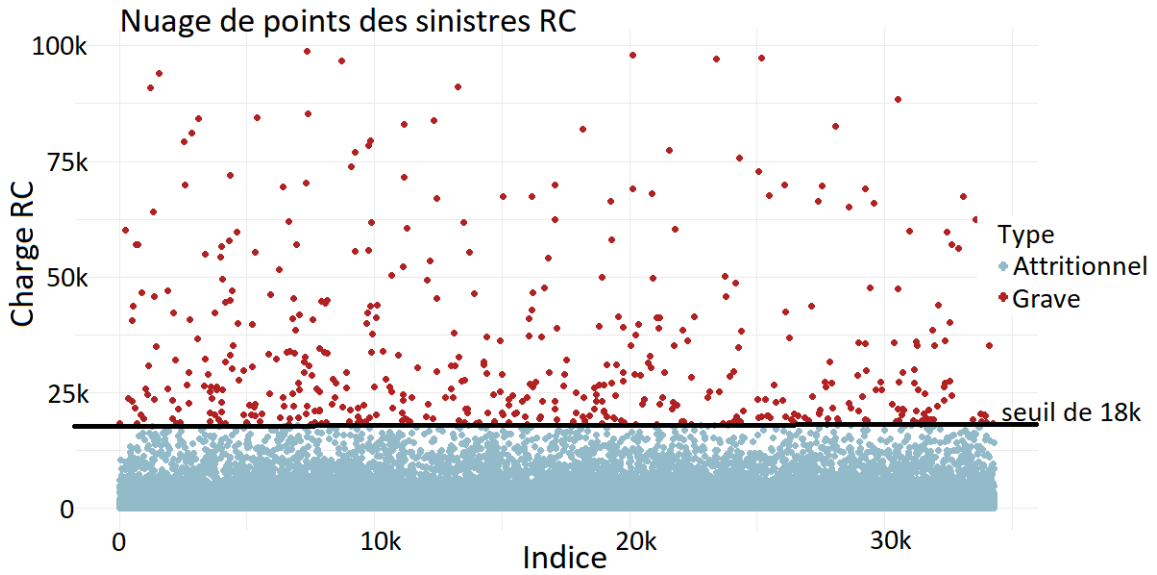


FIGURE 3.14 – Application d'un seuil de 18000€ sur les sinistres RC

Visuellement, ce seuil semble plutôt bien distinguer la charge grave de la charge attritionnelle. Attention, cette représentation graphique ne donne qu'un aperçu de la distribution des sinistres, mais il n'est pas possible de définir un seuil avec cette méthode.

Remarque : l'étude sur le seuil des graves a été réalisée sur l'ensemble du portefeuille et à la maille sinistre. En effet, la maille sinistre est plus fine que la maille véhicule utilisée pour la modélisation et permet ainsi de déterminer plus finement les sinistres graves. De plus, le choix d'utiliser l'ensemble des données et non, seules les données d'apprentissage, vient de la notion de sinistres graves qui est par définition rare, donc il serait plus intéressant et précis de prendre l'ensemble de l'information sinistre.

3.2.4 Exposition

Les sinistres dont la durée d'exposition au risque est égale à un jour ont été supprimés. La fréquence est par définition élevée lorsque les sinistres sont survenus sur des images de faible durée. Par exemple, lors de la mise en offset de la durée d'exposition dans la modélisation, on définit un véhicule ayant eu 1 sinistre pendant une année de survenance dont la durée d'exposition est de $0,2$ année. Alors, la fréquence de sinistre que l'on considère rapportée sur une année est de $\frac{1}{0,2}$, soit 5 sinistres. Cette estimation est simplement un produit en croix. Cependant, cela peut poser des problèmes sur une éventuelle sur-dispersion si la durée d'exposition est excessivement faible malgré la rectification par un poids. La mise en offset permet également d'accorder un nombre d'images dans la modélisation plus faible pour ces sinistres. Ainsi, les sinistres avec des images de durée faible sont moins influents dans la modélisation.

L'analyse de la répartition de la fréquence selon l'exposition annuelle montre la sur-représentation des images de faible exposition, plus précisément d'un jour. Ce phénomène s'explique par une clôture de la plupart des contrats au 1er janvier de l'année N, pour des contrats commençant au 1er janvier de l'année $N-1$. Ainsi, ces images ne correspondent pas a priori à un profil de risque particulier. Les images exclues représentent moins de $0,06\%$ des sinistres (Figure 3.15).

16. Pour rappel : le coefficient PLR est à la maille PLR et permet de passer de la prime pure à la prime commerciale. Il prend en compte les sinistres atypiques et climatiques.

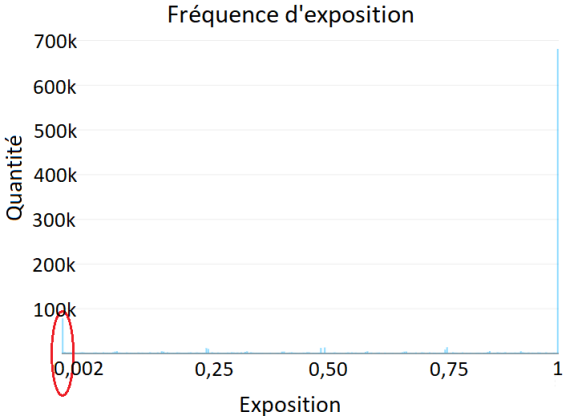


FIGURE 3.15 – Fréquence de la durée de l'exposition

3.2.5 Capitalisation

Pour la suite, les sinistres seront capitalisés à fin 2019 afin que la charge des années de survenance représente une même instance temporelle. La prise en compte de l'inflation des coûts est différente selon le type de dommage lié à la RC [17] (RC MAT [18] et RC CORP [19]).

En effet, la RC MAT augmente avec l'indice du coût total des réparations (pièces + main-d'œuvre), qu'il est donc intéressant d'appliquer. Cette donnée est disponible sur le site de la SRA.

L'inflation des coûts liée à la RC CORP est attachée à plusieurs facteurs dont l'inflation des frais médicaux, des augmentations de salaire, de la durée d'arrêt de travail ... Finalement, une étude menée par le département réassurance et modèle interne a finalement été conservée (pour la RC MAT et la RC CORP) dans le cadre de l'étude, car elle reflète l'inflation de la sinistralité propre à notre portefeuille.

3.3 Sélection de variables

La sélection de variables est une étape importante dans la modélisation. La base de données comprend plus de 90 variables tarifaires dont 36 variables quantitatives. Plus le nombre de variables est élevé, plus le modèle s'ajustera aux données grâce au principe de maximisation de la vraisemblance du GLM. Cependant, cela implique un généralement un sur-apprentissage. Il faut donc trouver un compromis entre parcimonies (données avec le plus petit nombre de paramètres possibles) et biais (qui diminue avec le nombre de paramètres). Afin de répondre à cette problématique, plusieurs critères sont utilisés afin de pénaliser les modèles en fonction du nombre de paramètres β :

— le Critère d'Information d'Akaïke (AIC) est défini par $-2ln(L) + 2k$ dont le principe repose sur la maximisation du log-vraisemblance du modèle noté $ln(L)$ [20] en pénalisant par deux fois le nombre de paramètres à estimer du modèle. L'AIC est utilisé dans un objectif de prédiction et non de décision vis-à-vis de la signification statistique des paramètres retenus dans le modèle (LANCELOT et LESNOFF, 2005). Le meilleur modèle au sens de ce critère serait celui avec l'AIC le plus faible,

— le Critère d'Information d'Akaïke corrigé (AICc) est défini par $AIC + \frac{2k(k+1)}{n-k-1}$ et est recommandé par rapport à l'AIC lorsque k est grand par rapport au nombre d'observations n, c'est-à-dire lorsque $n/k < 40$,

— le Critère d'Information Bayésien (BIC) est défini par $-2ln(L)+kln(n)$ dont le principe est similaire à l'AIC, mais dont la pénalisation est d'autant plus importante que le nombre de variables noté n est important (si $ln(n) > 2$ la pénalisation du BIC est plus importante que l'AIC). Le BIC a été initialement proposé par Schwartz en 1978 pour sélectionner les modèles dans le cas de grands échantillons. Il aboutissait à des modèles plus parcimonieux.

17. Responsabilité Civile
18. Responsabilité Civile des dommages matériels
19. Responsabilité Civile des dommages corporels
20. $deviance = -2ln(L)$

3.3.1 Sélection parcimonieuse

Classiquement, en assurance, on procède à une étape de sélection des variables en minimisant la valeur d'un des critères pour le modèle. Cependant, il est impossible numériquement de réaliser une recherche exhaustive des variables optimisant le modèle dont la complexité est de l'ordre de $\sum_{k=0}^{p} \frac{p!}{k!(p-k)!} = 2^p$, avec k le nombre de variables retenues dans le modèle et p le nombre de variables disponibles. Ainsi, l'objectif est de procéder pas-à-pas en partant du modèle de départ contenant les p variables et de supprimer l'une après l'autre les variables. À l'étape j, la variable supprimée est celle qui améliore le critère utilisé. Si le modèle sans suppression de la variable optimise le critère par rapport au modèle avec suppression alors l'algorithme s'arrête. Cette méthode est nommée méthode descendante. Une méthode ascendante existe, dans laquelle les variables sont ajoutées petit à petit, mais elle donne en général de moins bons résultats.

Une seconde méthode est utilisée couramment, dont le temps d'exécution est court, mais est non retenue, car moins robuste et moins adaptée à un jeu de données contenant beaucoup de variables ayant une forte corrélation entre elles. Elle ajuste un modèle à chaque étape en testant la significativité des variables.

La méthode retenue par l'outil opérationnel est une approche bayésienne qui prend *a priori* l'importance des variances dans les modèles. Ensuite, il est possible de jouer sur cet *a priori* d'importance pour implémenter une sélection automatique. Il s'agit en quelque sorte d'une généralisation de l'approche descendante qui est affranchie des problèmes de corrélation entre variables. De nombreux modèles sont créés afin d'évaluer lequel fonctionne le mieux. Chacun des modèles est ajusté grâce à une combinaison spécifique d'hyperparamètres, dont les plages sont définies par l'utilisateur lors de la configuration de la recherche de grille. Il existe deux hyperparamètres relevant du concept de parcimonie dont une plage comptant le nombre de variables à conserver (par exemple entre 5 et 22, Figure 3.16) et la granularité, en créant plusieurs modèles sur un nombre de variables compris dans la plage. Un dernier hyperparamètre permet d'évaluer la sensibilité d'un lissage. Un modèle lissé suivra principalement la tendance des données, mais ne permettra pas de capturer toutes les variations tandis qu'un modèle non lissé peut capturer des variations au détriment de la robustesse. Chaque point sur le graphique représente la performance moyenne du test K-Fold selon différentes métriques (ici le Gini) d'un modèle. Les barres d'erreurs affichent les valeurs minimales et maximales sur les folds. La position sur l'axe des abscisses indique le nombre de variables conservées dans le modèle et la position sur l'axe des ordonnées indique les performances hors échantillon.

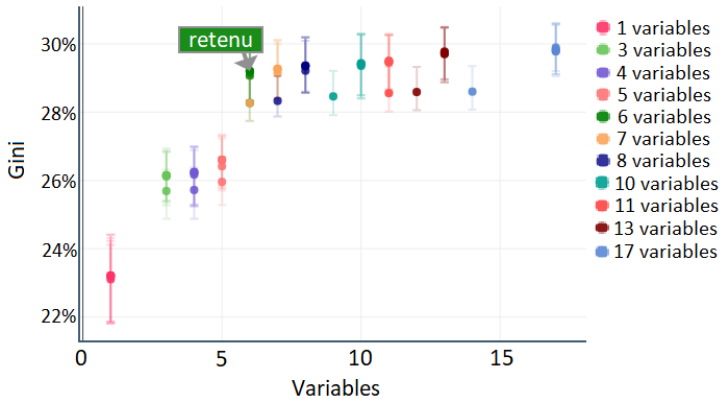


FIGURE 3.16 – Fréquence de la durée de l'exposition

Le modèle retenu est le modèle donnant un pouvoir prédictif proche de la meilleure performance au sens des différentes métriques utilisées en considérant le minimum de variables possibles. En effet, le compromis biais-variance est ainsi optimal.

3.3.2 Regroupement

Section 3.4, page 79, une recatégorisation des variables qualitatives et une catégorisation des variables quantitatives grossière seront produites. En effet, les variables quantitatives seront catégorisées en vingt quantiles et une partie des variables catégorielles sera dupliquée en regroupant des modalités sur des jugements historiques. D'ailleurs, des variables sont ajoutées comme des interactions complètes. Ce sont

des variables décrivant les valeurs conjointes de deux variables.

L'objectif de cette partie est d'améliorer la performance du modèle en retravaillant cette fois-ci les variables de façon judicieuse. L'impact est plus élevé quand il s'agit de catégorisation de variables quantitatives, car les modèles créés par l'outil ne sont composés que de variables catégorielles dont les coefficients sont lissés. Les variables qualitatives ont des modalités qui sont automatiquement regroupées (coefficients égaux). En revanche, il est intéressant de recatégoriser ces variables par d'autres méthodes que la pénalisation utilisée dans l'outil et de mesurer l'impact de cette modification. De plus, recatégoriser certaines variables qui auraient pu être sélectionnées dans l'étape de parcimonie pourrait confirmer la sélection des variables. En effet, certaines variables ont été peut-être trop fortement pénalisées à cause de la présence d'un nombre élevé de modalités. Ainsi, un regroupement judicieux des modalités a l'avantage de conserver seulement l'information jugée pertinente de chaque variable.

Certaines variables qui ont prouvé leur significativité historiquement, mais qui n'ont pas été sélectionnées par l'outil, seront retraitées dans le modèle de coût moyen et dans le modèle de fréquence. Par exemple, la marque du véhicule, ou encore les trois indicateurs de classements de véhicules proposés par SRA ; le groupe, la classe de prix et la classe de réparation.

Le groupe a pour finalité d'aider les assureurs à tarifer la garantie RC. C'est un indicateur qui est le reflet de la dangerosité intrinsèque des véhicules. Le calcul du groupe prend en compte plusieurs variables présentes dans le jeu de données :

— la puissance réelle du véhicule,
— le poids du véhicule (poids à vide et PTAC),
— la vitesse maximale du véhicule,
— un indicateur estimant la sécurité globale du véhicule dépendant de la conception et des équipements de sécurité du véhicule.

La classe de prix est un indicateur en cas de perte totale ou de vol du véhicule regroupant le prix des véhicules neufs en tranche de prix défini par une lettre. Il est réactualisé chaque année en fonction de l'évolution de l'indice INSEE du prix des véhicules neufs.

La classe de réparation est un indicateur estimant le coût de la réparation d'un véhicule par tranche, réactualisé chaque année en fonction de l'évolution de son indice.

Regroupement par méthode de type arbre

Une première méthode permet de regrouper les modalités d'une variable catégorielle non-ordinale de façon ascendante. Elle se concentre sur les principaux effets des prédicteurs catégoriels en utilisant des méthodes de type arbres de régression pour obtenir des clusters de catégories. Il s'agit de construire un GLM dans lequel certaines variables catégorielles sont recatégorisées via un découpage arborescent de leurs modalités. Cette méthode est introduite sous la forme d'une fonction simple *structree* sur **R** (TUTZ et BERGER, 2018a). Lorsque le prédicteur a plusieurs catégories, on veut savoir en particulier quelles catégories doivent être distinguées en fonction de leur effet sur la réponse. L'approche arborescente permet de détecter des clusters de catégories qui partagent le même effet tout en laissant d'autres prédicteurs, en particulier métriques, avoir un effet linéaire ou additif sur la réponse.

Un algorithme d'ajustement est proposé et différents critères d'évaluation sont calculés dont le BIC qui sera conservé pour cette méthode. De plus, la stabilité des clusters et la pertinence des prédicteurs est approfondie par des méthodes de bootstrap. Finalement, cette approche ne sera pas conservée, car elle ne se prête pas à un grand jeu de données et elle a une limite matérielle (mémoire vive).

Regroupement par test

Une seconde approche consiste à fusionner des catégories guidées par le test statistique de Wald. En supposant la validité du modèle GLM associé au jeu de coefficients $\beta = (\beta_0, \ldots, \beta_d)$, le test de Wald cherche à tester une hypothèse (H_0) sur les coefficients de la forme :

$(H_0) : h(\beta) = 0,$
où $h(\beta) = (h_1(\beta), \ldots, h_r(\beta))^T$ est une fonction régulière de β.

(H_1) : $h(\beta) \neq 0$.

Sous (H_0), $Loi({}^t h(\hat{\beta}) \times Q(\hat{\beta})^{-1} \times h(\hat{\beta})) \approx \chi^2(r)$,
avec $B(\beta) := [\frac{\partial h_l}{\partial \beta_j}]_{\substack{1 \leqslant l \leqslant r \\ 0 \leqslant j \leqslant d}}$ et $Q(\beta) = B(\beta) \times I(\beta)^{-1} \times^t B(\beta)$

Le test consiste à situer la valeur observée de $T_{Wald} =^t h(\hat{\beta}) \times Q(\hat{\beta})^{-1} \times h(\hat{\beta})$ vis-à-vis de la loi $\chi^2(r)$. La p-valeur du test est donnée par $1 - F_{\chi^2(r)}(T_{Wald}^{obs.})$. Une p-valeur inférieure au seuil α de 5% conduit à un rejet de l'hypothèse nulle pour un niveau de risque donnée et ainsi β_k apparait comme statistiquement significatif. Les limites de ce test sont :

— la définition du seuil à partir duquel on décide de rejeter ou pas l'hypothèse nulle,
— les valeurs plausibles pour β_k peuvent-elles conduire à un effet significatif en pratique ?
— les coefficients (β_i) encodant chaque modalité d'une variable catégorielle ayant une valeur non nulle traduit simplement une différence d'effet vis-à-vis de la modalité prise comme référence.

Les tests statistiques d'égalité de coefficients fournissent une manière d'aborder la fusion de catégories. Cette méthode partiellement automatique fusionne deux catégories pour lesquelles la p-valeur d'un test statistique est la plus élevée. Il est possible de visualiser le tracé des coefficients avec leurs intervalles de confiance afin de vérifier la bonne application du test. Cette procédure est répétée jusqu'à ce qu'aucun test d'égalité ne donne une p-valeur supérieure au seuil. Elle est utilisée pour fusionner une variable catégorielle quelle qu'elle soit. Cependant, cette méthode utilise une approche descendante au sens où on élimine progressivement des modalités en les regroupant avec d'autres. De plus, la multiplicité des tests pratiqués rendent peu claire la significativité statistique globale des résultats obtenus.

Remarque : il existe plusieurs tests statistiques portant sur les coefficients dont le test de Wald, le test du rapport de vraisemblance et le test du score. Le test de Wald est un test systématiquement implémenté dans la fonction GLM sur ***R*** *pour la nullité éventuelle de chacun des coefficients estimés.*

Regroupement par critère de pénalisation

Une troisième approche adaptée aux variables catégorielles ordinales consiste à un regroupement par minimisation d'un des trois critères de pénalisation ; AIC, AICc et BIC.

Pour une variable catégorielle ordinale, on veut en général regrouper seulement les modalités voisines entre elles. Cette approche utilise une méthode pas à pas, en ajustant un GLM par regroupement de deux modalités voisines. À chaque étape, si le GLM est plus performant au sens du critère de pénalisation alors le regroupement est conservé. Cette méthode peut demander un temps de calcul considérable, c'est pourquoi elle n'est réalisée que sur les variables catégorielles ordinales qui imposent un nombre moins élevé de combinaisons.

Remarque : cette méthode ne donne pas toujours un AIC ou BIC optimisé si le critère de pénalisation AIC respectivement BIC a été choisi, car le regroupement est optimisé seulement localement et utilise une méthode descendante.

Regroupement par pénalisation de type Lasso

Pour rappel, dans le cadre d'un modèle linéaire standard, les coefficients $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$ sont obtenus par minimisation de la somme des carrés des résidus : $min_{\beta_0,\ldots,\beta_p} \frac{1}{2} \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{i,j})^2$. En statistiques, *Robert Tibshirani* propose en 1996 d'ajouter une pénalisation dit Lasso [21]. Cette pénalisation permet de contracter les coefficients évitant ainsi un sur-apprentissage en lissant les coefficients. Cela se traduit par une détermination des coefficients :

$$\hat{\beta} = min_{\beta_0,\ldots,\beta_p} \frac{1}{2}||y - X\beta||_2^2 + \underbrace{\lambda||\beta||_1}_{Lasso} + \underbrace{\lambda \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|}_{Fused-Lasso} + \underbrace{\lambda \sum_{j=1}^{K} \omega_j ||\beta_{G_j}||_2}_{Group-Lasso},$$

avec :

— $\lambda \geqslant 0$ le paramètre de régularisation optimisé par un critère de performance pour le modèle ajusté,
— $\omega_j > 0$, un poids associé au groupe G_j, par défaut $\omega_j = \sqrt{Card(G_j)}$,

21. *Least Absolute Shrinkage and Selection Operator*

— $\Gamma = (G_1, \ldots, G_K)$, une partition des p variables en K groupes,
— $||x||_q = (\sum_{i=1}^{n} |x_i|^q)^{1/q}$ la norme l_q,
— y la variable à expliquer,
— X matrice des variables explicatives.

Dans cette formule, il y a deux extensions de la pénalisation Lasso :

— le fused Lasso est utilisé pour regrouper les coefficients des modalités voisines,
— le Group-Lasso fournit une sélection parcimonieuse de groupes *a priori*.

La fusion de catégories par pénalisation du type Lasso consiste à ajouter à la vraisemblance des termes qui pénalisent les différences d'effets entre catégories forçant certaines différences pénalisées à prendre une valeur nulle.

Ainsi, derrière la recatégorisation par pénalisation Lasso, l'idée est de regrouper les modalités qui ont des coefficients égaux en utilisant le fused-lasso pour les variables catégorielles ordinales, le Group-Lasso pour les variables catégorielles non-ordinales et le Lasso pour les variables quantitatives. Cette méthode n'est réalisable que pour une recatégorisation d'une variable catégorielle. La fonction *glmsmurf* sur *R* ajoute une sélection parcimonieuse de lambda par cross-validation.

Catégorisation par effet lisse expliqué par un arbre

Le modèle additif généralisé, GAM, est un modèle statistique dont la structure est analogue à celle d'un GLM, avec une composante systématique se décomposant sous la forme additive suivante :

$$\eta^{22} = \beta_0 + \underbrace{\sum_{j=1}^{q} f_j(x_{(j)})}_{effet\ des\ var.\ quantitatives} + \underbrace{\sum_{j=q+1}^{m} g_j(x^{(j)})}_{effet\ des\ var.\ catgorielles},$$

où f_j et g_j sont des fonctions s'écrivant comme combinaison linéaire de fonctions de bases. Typiquement, pour les variables quantitatives, les fonctions f_j correspondent à des splines. Une spline est une fonction polynomiale définie par morceaux. Elles sont utilisées dans des problèmes d'interpolation permettant de réaliser un compromis entre la régularité de la courbe et le degré des polynômes utilisés. Les polynômes sont reliés par une condition de continuité.

Une variable quantitative peut être catégorisée en relevant l'effet de la variable par lissage avec une fonction spline dans un modèle GAM. Ensuite, cet effet lissé peut être expliqué par la variable quantitative à partir d'une méthode de machine learning : un arbre de régression. En effet, le principal intérêt de cette méthode est d'utiliser un arbre de décision pour créer des classes. L'étape de lissage de l'effet de la variable sur le coût moyen ou sur la fréquence est une méthode de traitement du signal permettant d'obtenir un résultat plus robuste.

La fonction *evtree* sur *R* est utilisée afin de déduire les classes (HENCKAERTS et al., 2017). Il s'agit de l'apprentissage évolutif d'arbres globalement optimaux. La méthode d'arbre de classification et de régression couramment utilisée est l'algorithme CART [23] qui est une méthode de partitionnement récursive construisant le modèle à partir d'une recherche en avant. Cette méthode n'est efficace que localement, car les divisions sont choisies pour maximiser l'homogénéité à la prochaine étape seulement.
evtree implémente un algorithme évolutif. En revanche, cet algorithme ne sera pas utilisé au cours de ce mémoire bien que ce soit certainement le plus robuste des modèles d'arbres avec en supplément une pénalisation selon le nombre de feuilles avec l'AIC. En effet, la fonction est gourmande en mémoire et se révèle trop longue à faire tourner pour des variables continues.

Ainsi, deux autres méthodes ont été étudiées afin de répondre au mieux à nos objectifs, c'est-à-dire catégoriser la variable en expliquant au maximum l'effet tout en minimisant le nombre de classes. Ces deux méthodes reposent sur un arbre de régression utilisant un algorithme CART, mais la gestion des hyperparamètres est différente. Les deux méthodes consistent à développer un arbre avec l'effet lissé expliqué par la variable cible à catégoriser, en introduisant un poids selon l'exposition, selon la durée d'exposition au risque pour le modèle de fréquence, et selon le nombre de sinistres pour le modèle de

22. Pour rappel, $\eta = \sum_{j=0}^{d} \beta_j x(j) = g(\mu)$ avec g la fonction de lien et $\mu = \mathbb{E}(Y|X = x)$
23. L'algorithme *Classification And Regression Trees* (CART) construit des estimateurs constants par morceaux sur des partitions. Un découpage binaire récursif est réalisé en choisissant tout d'abord une variable sur l'ensemble des variables explicatives en maximisant la variance intergroupe. L'algorithme se finalise lorsque dans chaque nœud terminal il n'y a qu'une observation ou selon un critère d'arrêt à définir.

charge. Ce poids permet de ne pas accorder trop d'importance à de faibles expositions.

Il existe plusieurs hyperparamètres dans un arbre de régression. Seulement quelques-uns seront utilisés par la suite :

— la profondeur maximale de l'arbre. Ce paramètre *max_ depth* correspond au chemin le plus long entre le premier nœud (ou nœud racine) et le dernier nœud (ou nœud feuille),
— le nombre maximum de nœuds terminaux, c'est-à-dire le nombre maximum de modalités accordées à la variable dans notre cas,
— le nombre minimum d'observations requis pour être un nœud terminal, noté *min_ samples_ leaf*, c'est-à-dire l'exposition minimum que l'on juge bon d'admettre pour chaque modalité,
— le paramètre de complexité pour un élagage de complexité de coût minimal. Le sous-arbre avec la complexité de coût la plus élevée qui est inférieure à ce paramètre *cp* sera choisi. Son rôle est de gagner du temps de calcul en éliminant les scissions qui n'en valent pas la peine.

Chacun de ces hyperparamètres permet d'élaguer l'arbre de manière intelligente et ainsi catégoriser la variable en expliquant au maximum l'effet tout en minimisant le nombre de classes.

La première méthode, développée à l'aide de la fonction *DecisionTreeRegressor* sur **Python**, utilise 3 hyperparamètres ; la profondeur maximale de l'arbre, le nombre maximum de nœuds terminaux et le nombre minimum d'observations par nœud feuille.

Le nombre minimum d'observations par nœud feuille a été arbitrairement fixé à 2% de l'exposition totale afin d'avoir suffisamment de données par classe, et ainsi d'obtenir des classes plus crédibles. Cette contrainte, reprise pour la seconde méthode, signifie également que la variable continue ne peut être catégorisée avec plus de 50 modalités, ce qui est largement satisfaisant au vu des dernières expériences sur les variables catégorisées par quantiles. Dans la suite, les hyperparamètres sont optimisés par cross validation (5-fold) à l'aide de la fonction *RandomizedSearchCV* permettant de parcourir aléatoirement 200 choix possible de valeurs des hyperparamètres en délimitant un *max_ depth* à 8 et *min_ samples_ leaf* à 30 maximum.

Ensuite, le principe de la méthode pour déterminer quelle est la meilleure catégorisation de la variable quantitative repose sur le nombre de modalités que l'on fait varier entre 2 et 30. À la fin de l'algorithme, une sélection arbitraire est réalisée pour déterminer la meilleure catégorisation au sens du coefficient de détermination R^2 [24]. L'objectif est de trouver l'équilibre entre nombre de modalités retenues et l'optimisation de la performance du modèle par le biais du critère R^2.

La deuxième méthode consiste à utiliser un hyperparamètre d'élagage, noté *cp* [25] et à limiter le nombre maximum de nœuds terminaux par des feuilles ayant un poids minimum égale à 2% de l'exposition. De façon simple, plus ce paramètre de complexité est petit, plus l'arbre de régression est grand, et plus il est grand, plus l'arbre est petit, c'est-à-dire que le niveau d'élagage est élevé. Ainsi, un élegage judicieux correspond à une valeur du paramètre de complexité cp rendant petite l'erreur relative de la validation croisée [26]. La valeur optimale pour *cp* est celle qui minimise l'erreur relative. Cependant, l'intérêt n'est pas de minimiser le *cp*, mais de trouver un compromis entre prédiction et nombre de modalités.

Application

Pour le modèle du coût moyen ou le modèle de fréquence, chaque méthode est utilisée selon la nature de la variable :

— variable catégorielle non-ordinale : méthodes AIC, BIC, Lasso et test de Wald,
— variable ordinale : méthodes Lasso et test de Wald,

24. Le coefficient de détermination linéaire de Pearson noté R^2 est une mesure de la qualité de la prédiction d'une régression linéaire. Ce coefficient est défini comme,

$$R^2 = 1 - \frac{sum_{i-1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i-1}^{n}(y_i - \bar{y})^2} = corr(\hat{y}, y)^2,$$

où n est le nombre de mesures, y_i l'observé i, $\hat{y}_i$ la valeur prédite I et $\bar{y}$ la moyenne des mesures.

25. Pour de plus amples informations sur ce paramètre, se référer à l'article de *Breiman L.*, *Friedman J. H.*, *Olshen R. A.*, and *Stone C. J.* (1984) Classification and Regression Trees. Wadsworth.

26. L'erreur relative se définit comme,

$$\delta_i = \frac{\hat{y}_i - y_i}{|y_i|},$$

où y_i est l'observé i et $\hat{y}_i$ est la prédiction i.

— variable quantitative : méthode effet lissé par GAM + arbre élagué par plusieurs hyperparamètres (dont la profondeur et le nombre maximal de feuilles, ou dont le paramètre de complexité *cp*).

Chaque variable étudiée est introduite dans un modèle dont les variables explicatives sont les variables les plus significatives et la variable à catégoriser. La significativité d'une variable est déduite par la sélection répétitive de celle-ci dans l'outil et par son spread élevé.

Pour la fréquence, deux variables sont significatives ; l'ancienneté du véhicule et le segment NAF [27]. Un modèle composé de ces deux variables conduit à un Gini de 27% environ. Pour comparaison, un modèle précédé d'une sélection parcimonieuse et d'un lissage confère un Gini de 30% environ. De plus, ces deux variables possèdent le spread 100/0 et 95/0 le plus élevé. La notion de spread est décrite dans la sous-section 3.3.4, page 72.

Pour le modèle de cout moyen, il s'agit des variables *segment NAF* (l'activité de l'entreprise) et *PTAC* [28] du véhicule.

Ensuite, toutes les méthodes de regroupement ont été appliquées sur les données d'apprentissage afin de différencier les données utilisées pour l'ajustement, des données de test permettant de vérifier après modélisation la performance de la prédiction. Pour rappel, les données d'apprentissage représentent 80% des individus. Les données manquantes pour les variables quantitatives ou catégorielles ordinales ont été retirées afin de ne pas biaiser les méthodes de regroupements. Pour les variables catégorielles non-ordinales, une modalité spécifique pour les données manquantes a été créée afin de réaliser un regroupement de cette information, si possible.

Dans les méthodes de regroupement de modalités, un arbitrage supplémentaire est nécessaire, car les modalités sont regroupées sans prise en compte du sens opérationnel. De plus, certains groupes ont une faible exposition. Ainsi, le choix de la meilleure méthode de regroupement réunit les deux points définis précédemment. De plus, un score élevé, AIC ou BIC mesurant la performance du modèle avec la variable recatégorisée, est requis. De façon générale, le regroupement par le test de Wald donne de meilleurs résultats que les autres méthodes, viennent ensuite la méthode BIC, la méthode AIC et enfin la méthode avec la pénalisation Lasso.

Pour la catégorisation d'une variable quantitative, la méthode par élagage avec le paramètre de complexité est majoritairement conservée. Au sens des critères de performance AIC et BIC, elle donne souvent de meilleurs résultats. De manière générique, cette méthode retient un nombre inférieur de classes. De plus, il s'agit d'une méthode plus simple à mettre en place et à exécuter. En fait, les deux méthodes d'arbres renvoient des résultats fortement similaires.

Par exemple, prenons la variable définissant le taux d'orientation dans les garages partenaires, *garp_tx_orientP*.

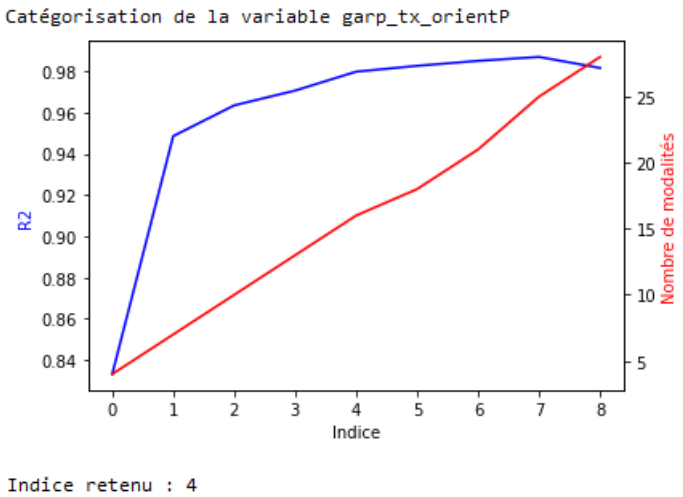


FIGURE 3.17 – Évolution du R^2 selon le nombre de modalités pour le taux d'orientation garage partenaire par arbre, élagué selon la profondeur et le nombre maximal de nœuds terminaux

27. Il s'agit de l'activité de l'entreprise.
28. Poids Total Autorisé en Charge

Au-delà de quinze nœuds terminaux conservés, la méthode réalisée sur **Python** (Figure 3.17) accorde une faible amélioration du R^2. Ainsi, dans un soucis de compromis biais-variance, une variable composée de 15 nœuds terminaux semble être un bon candidat. Le graphique suivant (Figure 3.18) représente la catégorisation retenue par la méthode.

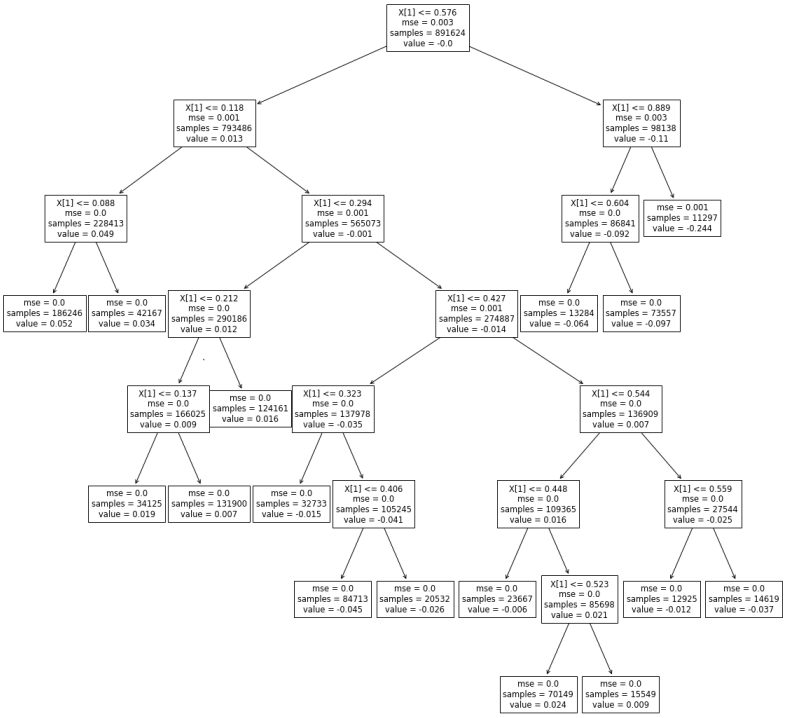


FIGURE 3.18 – Catégorisation du taux d'orientation garage partenaire par arbre après sélection des hyperparamètres

Pour la méthode sur R avec le paramètre de complexité, 15 modalités seront également retenues (Figure 3.20) avec $cp = 0.0014$ (Figure 3.19).

Les résultats sont regroupés dans un tableau où l'on note Méthode 1, la méthode utilisée figure 3.18, et Méthode 2, celle référencée figure 3.20.

Il apparaît que les deux méthodes donnent des résultats très similaires. Les seules différences détectées (en rouge, Figure 3.21) sont causées par un découpage plus fin sur certaines tranches. Ici, la méthode 2 sera préférée grâce au critère du score qui est plus élevé, impliquant une plus grande performance au sens de ce dernier. Une vérification de la bonne adéquation du regroupement peut être réalisée par une représentation graphique de l'effet lissé.

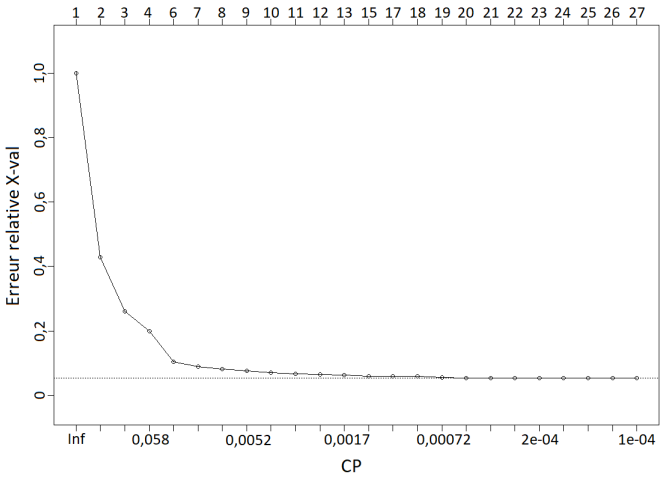


FIGURE 3.19 – Plot cp avec $cp = 0.0001$

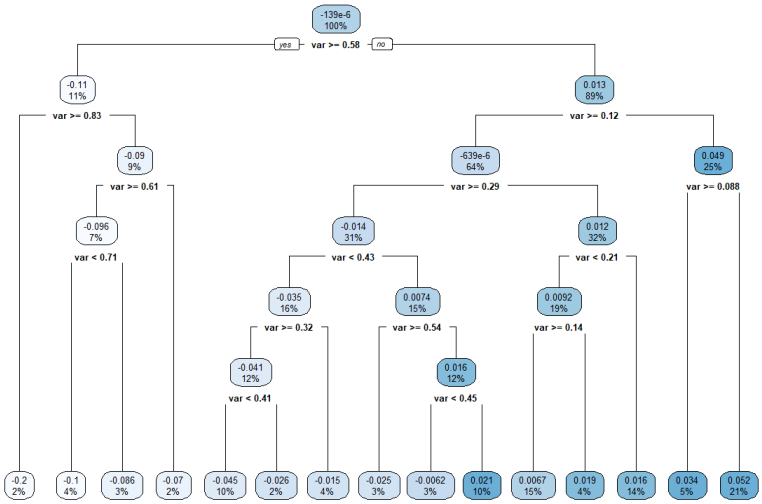


FIGURE 3.20 – Arbre avec élagage $cp = 0.0014$

	1	2	3	4	5	6	7	8	9
Méthode 1	[0;0.09[	[0.09;0.12[	[0.12;0.14[	[0.14;0.21[	[0.21;0.29[	[0.29;0.32[	[0.32;0.41[	[0.41;0.43[	[0.41;0.45[
Méthode 2	[0;0.09[	[0.09;0.12[	[0.12;0.14[	[0.14;0.21[	[0.21;0.29[	[0.29;0.32[	[0.32;0.41[	[0.41;0.43[	[0.41;0.45[

	10	11	12	13	14	15	16	17
Méthode 1	[0.45;0.52[	[0.52;0.54[	[0.54;0.56[	[0.56.;0.58[	[0.58;0.61[	[0.61;+[		
Méthode 2	[0.45;0.54[		[0.54;0.58[		[0.58;0.61[	[0.61;0.71[	[0.71;0.83[	[0.83;+[

FIGURE 3.21 – Classes construites par les deux méthodes

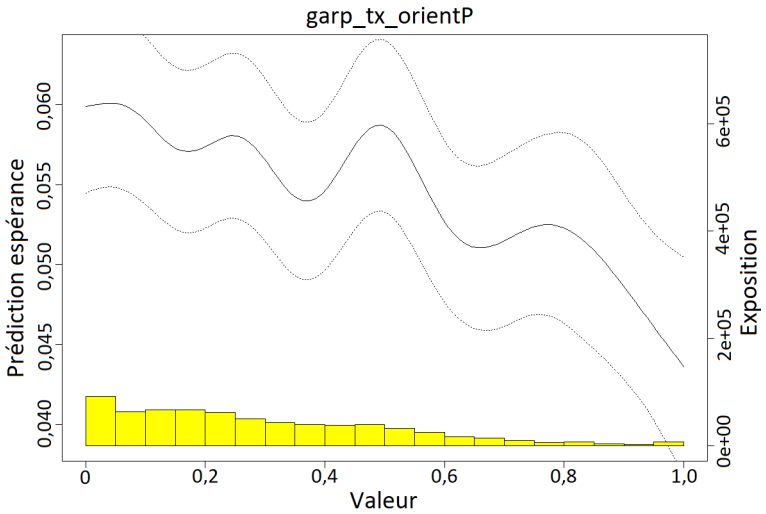


FIGURE 3.22 – Tracé de l'effet lissé

3.3.3 Sélection approfondie

Précédemment, certaines modalités de variables catégorielles ont été regroupées et les variables quantitatives ont été catégorisées finement afin d'améliorer le modèle au sens des critères de performance.

En parallèle à la sélection de variables faites par l'outil, une sélection parcimonieuse manuelle ciblée peut être exécutée à partir des variables les plus souvent sélectionnées par l'outil. Cette sélection manuelle permet d'ajouter deux métriques ; AIC et BIC. On souhaite minimiser ces deux critères de performances. L'AIC et le BIC pénalisent en fonction du nombre de cases tarifaires.

Une sélection de variables ne peut pas être réalisée sur l'ensemble des combinaisons possibles. Cependant, il est intéressant d'ajouter des combinaisons sur les variables qui *a priori* semblent significatives. Ensuite, une sélection de variables à partir des variables retenues sera implémentée dans l'outil. L'objectif de cette dernière sélection est de s'assurer que le modèle retenu soit plus performant au sens des métriques que les autres modèles candidats de l'outil. De plus, la méthode utilisée dans l'outil pour effectuer la régression n'a pas été effectuée à l'identique pour la sélection ciblée de variables.

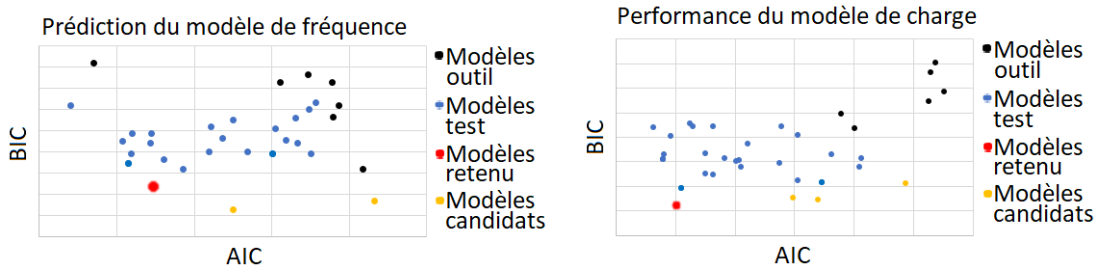


FIGURE 3.23 – AIC et BIC modèle de fréquence (à gauche) et modèle de coût moyen (à droite)

Les points en noir (Figure 3.23) correspondent aux modèles dont la sélection des variables a été réalisée par l'outil et un arbitrage parcimonieux. Les points en bleue correspondent aux modèles dont la sélection des variables a été ciblée. Le point en rouge correspond au modèle qui minimise au mieux l'AIC et le BIC (le plus en bas à gauche du graphique). C'est le modèle qui sera retenu. Les points orange sont les modèles dont la sélection des variables a été produite par l'outil sur les seules variables présentes dans le modèle retenu.

Le modèle ajusté par l'outil sur les variables explicatives ciblées présente de meilleures performances que tous les autres modèles jusqu'à présent entrainés. Pour rappel, l'objectif de l'étude est de sélectionner l'information la plus pertinente pour expliquer le risque. Toutefois, chacun des modèles présentés est valide et pertinent.

3.3.4 Analyse des variables retenues

Ainsi, les variables explicatives qui serviront à prédire la variable réponse ont été sélectionnées. Le spread fournit un classement des variables selon leur influence sur la variable cible.

Le spread est un indicateur calculé pour chaque variable catégorielle retenue dans le modèle. Il permet d'appréhender l'importance d'une variable dans la prédiction à travers la dispersion des coefficients. En effet, plus la dispersion des coefficients est élevée, plus la variable est significative. Un coefficient est accordé à chaque modalité d'une variable. Pour les modèles multiplicatifs et logistiques, le spread d'une variable correspond au rapport entre le coefficient le plus élevé et le coefficient le plus faible,

$$Spread\ d'une\ variable_{100/0} = \frac{max\ coefficient}{min\ coefficient} - 1$$

Le spread 95/5 est calculé de la même manière que le spread 100/0. En revanche, 5% de l'exposition la plus risquée et la moins risquée, autrement dit, les coefficients les plus élevés et les plus faibles sont retirés au préalable de l'application de la formule du spread. Cet indicateur permet une mesure de la sensibilité du spread et de la véracité de l'importance d'une variable.

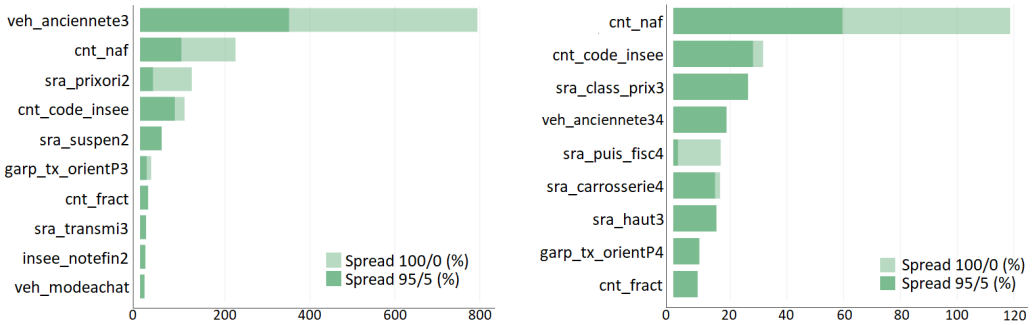


FIGURE 3.24 – Spread des variables du modèle de fréquence (à gauche) et du modèle de coût moyen (à droite)

Remarque : il est à noter que la variable cnt_code_insee est ici représentée après application d'un zonier (section 3.5, page 86).

Pour le modèle de fréquence, les variables sélectionnées sont :

— *veh_anciennete3* : l'ancienneté du véhicule en nombre d'années écrêtée à partir de 20 ans. Cette variable a été définie à partir de la date de première mise en circulation et de la date de sortie du véhicule du parc disponible dans les bases véhicules internes. Les coefficients sont décroissants lorsque l'ancienneté augmente. L'ancienneté du véhicule est une information corrélée à la sécurité du véhicule et à son prix,
— *cnt_naf* : code de 3 lettres définissant l'activité de l'entreprise défini par l'INSEE. C'est une variable à la maille contrat construite à partir de plusieurs bases internes afin d'avoir une meilleure cohérence et qualité de la donnée. L'activité de l'entreprise est une information corrélée à la carrosserie du véhicule,
— *sra_prixori2* : Prix à l'origine du véhicule catégorisé par un arbre à partir d'un effet lissé (sous-section 3.3.3, page 71). Plus le prix à l'origine est élevé plus le risque est élevé,
— *cnt_code_insee* : zone définie par un zonier (section 3.5, page 86),
— *sra_suspen2* : la suspension d'un véhicule est l'élément permettant de relier les masses non suspendues telles que les roues aux masses suspendues comme le châssis. Elle assure le contact des roues au sol et l'amortissement des chocs de la route sur le châssis. On distinguera deux systèmes qui semblent bien décrire le risque ; à roues indépendantes où sur un même essieu les roues droite et gauche ne sont pas liées et à essieu rigide ou semi-rigide où sur un même essieu les roues droite et gauche sont liées. Un dernier type de suspension disponible principalement pour les modèles de luxe est la suspension active qui s'adapte aux conditions de la route pour un meilleur confort. Cette modalité est proche, au sens opérationnel et au sens du risque, à la suspension avec roues

indépendantes, d'après un regroupement par pénalisation Lasso (sous-section 3.3.3, page 71). La suspension est étroitement liée au poids et à la vitesse max du véhicule,

— *garp_tx_orientP5* : le taux d'orientation garage partenaire est construit à partir d'une base interne et est agrégé à la maille portefeuille, c'est-à-dire courtier et agent. On remarque que plus le conseiller en assurance incite ses clients à rejoindre les garages partenaires AXA, plus le risque décroît. Cette variable est catégorisée à partir d'un arbre sur effet lissé. D'ailleurs, plus le niveau d'expertise du distributeur est élevé mieux il oriente ses clients,
— *cnt_fract* : le fractionnement de la prime à la maille contrat. Moins le fractionnement est élevé moins le risque est élevé (le fractionnement trimestriel a un risque significativement élevé),
— *sra_transmi3* : la transmission du véhicule est décrite en 4 modalités ; 4 roues débrayables, 4 roues permanentes, propulsion et traction. Les 4 roues débrayables sont souvent utilisées sous la forme de traction et ont un sens risque commun. Les deux modalités sont fusionnées (regroupement par pénalisation Lasso). Les véhicules ayant 4 roues motrices permanentes semblent plus risqués. La transmission est une information correlée à la carrosserie, à la puissance et au prix du véhicule,
— *insee_notefin2* : la note financière sur 20 de certaines entreprises à la maille SIRET (filiale). Cette note est fournie par une entreprise privée externe et est regroupée en seulement 3 modalités, regroupement par test de Wald. Plus la note financière est élevée, c'est-à-dire plus l'entreprise est pérenne financièrement, moins le risque est élevé. De plus, cette dernière est faiblement liée à l'effectif de l'entreprise,
— *veh_modeachat* : le mode d'achat du véhicule ; crédit, location longue durée, crédit-bail, partenaire financier ou autre financement. La location à longue durée et le crédit-bail semblent risqués contrairement à un crédit.

Pour le modèle de charge, les variables sélectionnées sont :

— *cnt_naf*,
— *cnt_code_insee*,
— *sra_class_prix3* : classe de prix regroupée en 4 modalités par test de Wald dont l'effet sur le modèle est croissant, c'est-à-dire que plus le prix actualisé du véhicule est élevé, plus le risque est élevé,
— *veh_anciennete34* : correspond à la variable *veh_anciennete3* vue précédemment regroupée en 3 modalités ($[0;1[$, $[1;10[$, $[10;+\infty[$),
— *sra_puis_fisc4* : l'effet de la puissance fiscale du véhicule est croissant. La puissance fiscale a été catégorisée par un arbre à partir d'un effet lissé,
— *sra_carrosserie4* : la carrosserie du véhicule type berline, break ... . Cette variable a donné lieu à un regroupement manuel en 11 modalités. Un sens opérationnel est conservé,
— *sra_haut3* : L'effet de la hauteur du véhicule est proportionnellement croissant avec la charge. C'est une variable catégorisée par un arbre à partir d'un effet lissé,
— *garp_tx_orientP4* : variable dont la structure est similaire au modèle de fréquence,
— *cnt_fract* : variable introduisant les mêmes constats que le modèle de fréquence.

Remarque : aucune information sur le conducteur n'est considérée, car il s'agit d'assurer des flottes de véhicules et non un conducteur.

Une description des variables permet de mieux comprendre leurs effets et ainsi, de mieux appréhender le risque. Pour chaque modalité de la variable étudiée, la moyenne de prédiction et la moyenne de l'observé sont tracées. Le but est d'observer des tendances qui influencerait la détermination de la prime pure. L'information pertinente à extraire est tirée de la variabilité des prédictions. Cette représentation graphique peut être distinguée selon les années de survenance en ajoutant une dimension *time consistency*. Cela vérifie la robustesse de la variable. Un point d'attention particulier est porté sur les variables dont la moyenne des prédictions ne présente pas une tendance similaire à la moyenne des observés, dont la seule modalité *NA* a un coefficient différent de 1, ou qui présente une forte variation *time consistency*. On ne souhaite pas que le modèle soit fortement influencé par les données manquantes. De même, une forte variabilité des résultats selon les années de survenance est à écarter, car cela signifie qu'une part de bruit est captée.

Plus le véhicule est ancien, moins il y a de sinistres (décroissance) proportionnellement à l'exposition quelle que soit l'année de survenance. De plus, les courbes de prédictions et d'observation quelle que soit l'année de survenance semble bien se superposer confirmant la robustesse de la variable.

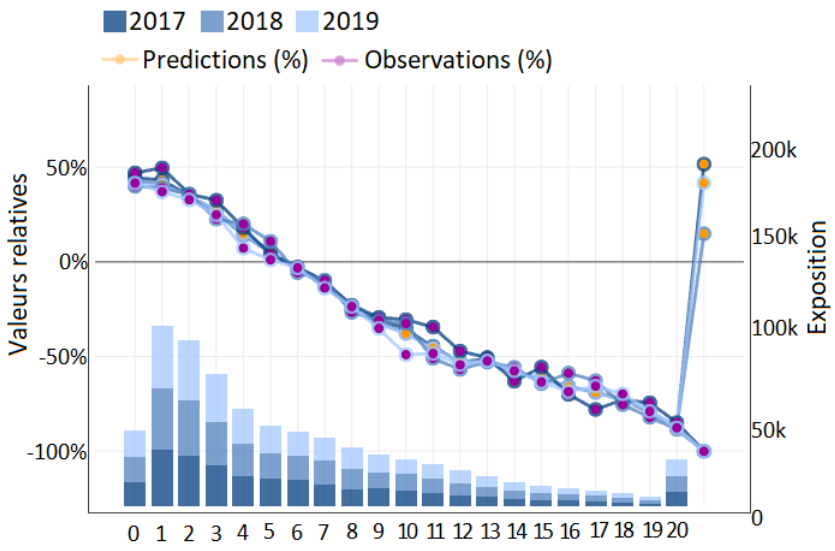


FIGURE 3.25 – Prédiction et observé par modalité et vision de l'ancienneté du véhicule sur le modèle de fréquence

Une étape avait été volontairement passée sous silence. En fait, la variable *garp_tx_orientP* du modèle de fréquence ou de charge a une faible consistance dans le temps (Figure 3.26).

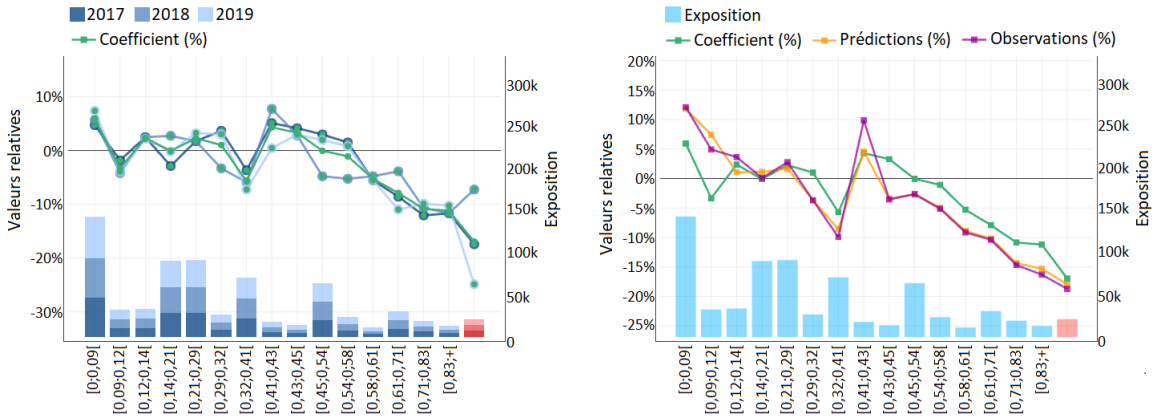


FIGURE 3.26 – Tracés de l'effet du taux d'orientation garage partenaire sur le modèle de fréquence

En raison de la grande variance des coefficients, certaines modalités sont regroupées afin de lisser l'effet. La recatégorisation est réalisée manuellement à l'aide des résultats graphiques. On souhaite conserver la décroissance à partir de 0.41 et regrouper les modalités qui ont une forte variation.

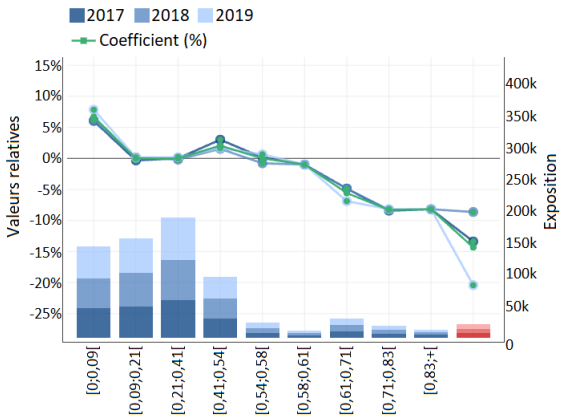


FIGURE 3.27 – Tracé de l'effet du taux d'orientation garage partenaire sur le modèle de fréquence après regroupement

Ce regroupement satisfait aux résultats escomptés. Les modèles retenus figure 3.23 ou figure 3.24, présentés précédemment, correspondent à des modèles dont les variables *garp_tx_orientP* ont déjà été retraitées.

3.3.5 Corrélation et colinéarité

Les GLM ainsi que les autres modèles multivariés, savent décorréler les variables entre elles, ce n'est donc pas un problème *a priori* d'utiliser deux variables corrélées dans un modèle. En revanche, une forte corrélation entre les variables est à l'origine d'un sur-apprentissage qu'il faut éviter. De plus, elle est la résultante d'une causalité qu'il est important de comprendre dans le processus d'évaluation de risque. Cela peut conduire peut-être à ajouter des termes d'interaction. Enfin, une forte corrélation ne permet pas d'interpréter directement l'impact d'une variable dans la modélisation.

Trois mesures de corrélations sont classiquement utilisées pour mesurer une relation entre deux variables quantitatives directement intégrées dans la fonction *cor* sur *R*.

Corrélations variables quantitatives

Coefficient de Pearson

Le coefficient de corrélation de Pearson entre deux variables X et Y mesure la relation linéaire. Il se présente comme la normalisation de la covariance des variables par leur écart-type :

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

où $COV(X,Y) = \mathbb{E}[(X - \mathbb{E}[X]).(Y - \mathbb{E}[Y])]$ et $\sigma_x = \sqrt{\mathbb{E}[(X - \mathbb{E}[X])^2]}$

Ce coefficient est ainsi compris entre -1 et 1 :

— $r_{xy} = 1$ si et seulement si $\exists$ a et b tels que $Y = aX + b$ avec $a > 0$. Si $r_{xy} > 0$, la relation entre les deux variables est linéaire et positive,
— $r_{xy} = -1$ si et seulement si $\exists$ a et b tels que $Y = aX + b$ avec $a < 0$. Si $r_{xy} < 0$, la relation entre les deux variables est linéaire et négative,
— $r_{xy} = 0$ implique une covariance nulle entre les deux variables. Cependant, cela ne permet pas de conclure sur l'indépendance hormis si le couple (X, Y) suit une loi normale bivariée, ce qui n'est pas le cas et donc ce ne sera pas une mesure retenue par la suite.

Le ρ de Spearman

Le ρ de Spearman se définit comme :

$$\rho = \frac{COV[rg(X), rg(Y)]}{\sigma_{rg(X)}.\sigma_{rg(Y)}}$$

où $rg(X)$ et $rg(Y)$ sont les variables de rangs respectivement de X et Y. Cette mesure est adaptée à des relations linéaires comme non-linéaire. En supplément, si $\rho = 0$ alors les variables X et Y sont indépendantes, sans hypothèse au préalable sur la distribution du couple (X, Y). Hormis ces avantages, le ρ de Spearman s'interprète de la même manière que le coefficient de Pearson.

Remarque : le ρ de Spearman s'apparente au coefficient de Pearson calculé à partir des rangs des réalisations.

Le τ de Kendall

Le τ de Kendall s'interprète comme une probabilité de correspondance de deux séries de données. Si $\tau > 0$, alors la probabilité de concordance des variables de rang est supérieure à celle de discordance et vice-versa lorsque $\tau < 0$, c'est-à-dire que la corrélation entre les variables est positive (respectivement négative). Enfin, lorsque $\tau = 0$, alors la probabilité de discordance est égale à celle de concordance, donc il n'y a pas de relation entre les variables de rang au sens de cette mesure.

C'est le ρ de Spearman qui sera retenu lors de cette étude, car il concilie interprétabilité de l'indépendance, corrélation à tout types de variables quantitatives et vitesse de calcul.

Ces trois mesures de corrélations sont limitées aux variables quantitatives ainsi, on introduit une troisième mesure, le V de Cramer, adapté pour évaluer les relations variables catégorielles.

Corrélation des variables qualitatives : V de Cramer

V de Cramer

Le V de Cramer est basé sur le test d'indépendance du χ^2 [29]. Soient U et V deux variables aléatoires indépendantes prenant respectivement leur valeur dans les ensembles $\{a_1, \ldots, a_K\}$ et $\{b_1, \ldots, b_L\}$. On suppose $\mathbb{P}(U = a_k) > 0$ pour tout $1 \leqslant k \leqslant K$ et $\mathbb{P}(V = b_l) > 0$ pour tout $1 \leqslant l \leqslant L$. On considère N couples indépendants et identiquement distribués (i.i.d) de variables aléatoires $(U_i, V_i)_{1 \leqslant i \leqslant N)}$ ayant pour loi commune celle du couple (U, V) et l'on défini :

$$S = N \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{(\frac{N_{kl}}{N} - \frac{N_{k.}}{N}\frac{N_{.l}}{N})^2}{\frac{N_{k.}}{N}\frac{N_{.l}}{N}},$$

où $N_{kl} = \sum_{i=1}^{N} \mathbb{1}_{\{U_i = a_k et V_i = b_l\}}$, $N_{k.} = \sum_{i=1}^{N} \mathbb{1}_{\{U_i = a_k\}}$ et $N_{.l} = \sum_{i=1}^{N} \mathbb{1}_{\{V_i = b_l\}}$

On a alors $S \xrightarrow[N \to \infty]{loi} \chi^2((K-1)(L-1))$.

Le test non-paramétrique du χ^2 d'indépendance repose sur l'hypothèse *H0 : "U et V sont indépendantes"*, les variables sont considérées comme indépendantes, et *H1 : "U et V ne sont pas indépendantes"*, donc les variables sont considérées comme dépendantes, avec une marge d'erreur de $\alpha = 5\%$ si les observations conjointes $(u_i, v_i)_{1 \leqslant i \leqslant N}$ ont le même comportement. Ces observations conjointes sont vues comme des réalisations des N couples i.i.d $(U_i, V_i)_{1 \leqslant i \leqslant N}$.

Pour un test d'ordre $(1 - \alpha)$, La zone de rejet du test correspond aux degrés de liberté [30] : $ddl = (K-1).(L-1)$. Ainsi, si la valeur $S^{obs.}$ [31] effectivement observée pour S est supérieure au seuil déterminé par le quantile d'ordre $(1 - \alpha)$ de la loi du χ^2 de $(K-1).(L-1)$ degrés de liberté, l'hypothèse *H0* est rejetée. Sinon, l'hypothèse nulle n'est pas rejetée. La p-valeur correspondante est donnée par $1 - F_{\chi^2((K-1)(L-1))}(S^{obs.})$.

Le V de Cramer est une version normalisée de $S^{obs.}$ permettant de comparer les valeurs ainsi obtenues entre des paires de variables possédant des valeurs différentes de K et L :

$$V = \sqrt{\frac{S^{obs.}/N}{min(K-1, L-1)}},$$

29. La loi du χ^2 est la loi engendrée par la somme au carré de k lois normales centrées et réduites indépendantes.

30. Les degrés de liberté correspondent au nombre maximum de valeurs du modèle telles qu'aucune d'entre elles n'est calculable à partir des autres.

31.

$$S^{obs.} = N \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{(\frac{n_{kl}}{N} - \frac{n_{k.}}{N}\frac{n_{.l}}{N})^2}{\frac{n_{k.}}{N}\frac{n_{.l}}{N}},$$

où $n_{kl} = \sum_{i=1}^{N} \mathbb{1}_{\{u_i = a_k \ \& \ v_i = b_l\}}$, $n_{k.} = \sum_{i=1}^{N} \mathbb{1}_{\{u_i = a_k\}}$ et $n_{.l} = \sum_{i=1}^{N} \mathbb{1}_{\{v_i = b_l\}}$.

où $N = \sum_{kl} N_{kl}$ est le nombre d'observations.

L'interprétation du V de Cramer est la suivante :

— plus V est proche de 0, plus les deux variables sont indépendantes.
— plus V est proche de 1, plus elles sont proches et corrélées.

Pour utiliser le test, la validité approchée de loi asymptotique $\chi^2((K-1)(L-1))$ suppose que chacune des catégories possibles a_k et b_l soit suffisamment représentée. Il est possible d'utiliser le V de Cramer sur les variables quantitatives une fois catégorisées.

Colinéarité

Dans une régression, la multicolinéarité est un problème qui survient lorsque certaines variables explicatives mesurent le même phénomène. Une multicolinéarité importante s'avère problématique, car elle rend instables les coefficients, avec ces conséquences :

— les coefficients peuvent sembler non significatifs,
— les coefficients des variables fortement corrélées peuvent varier considérablement d'un échantillon à un autre,
— les coefficients peuvent avoir un signe opposé à la réalité.

L'absence de multicolinéarité parfaite, c'est-à-dire lorsqu'une variable est la combinaison linéaire d'autres, est une condition requise pour pouvoir estimer un GLM. Une multicolinéarité approximative peut poser problème dans l'estimation et l'interprétation d'un modèle.

Dans les deux cas, on choisit simplement d'éliminer la variable soit la moins propre quand il y en a une, soit celle qui a un sens opérationnel moins important.

Remarque : deux variables colinéaires sont fortement corrélées, mais la réciproque est fausse. La corrélation permet simplement d'y porter un point d'attention.

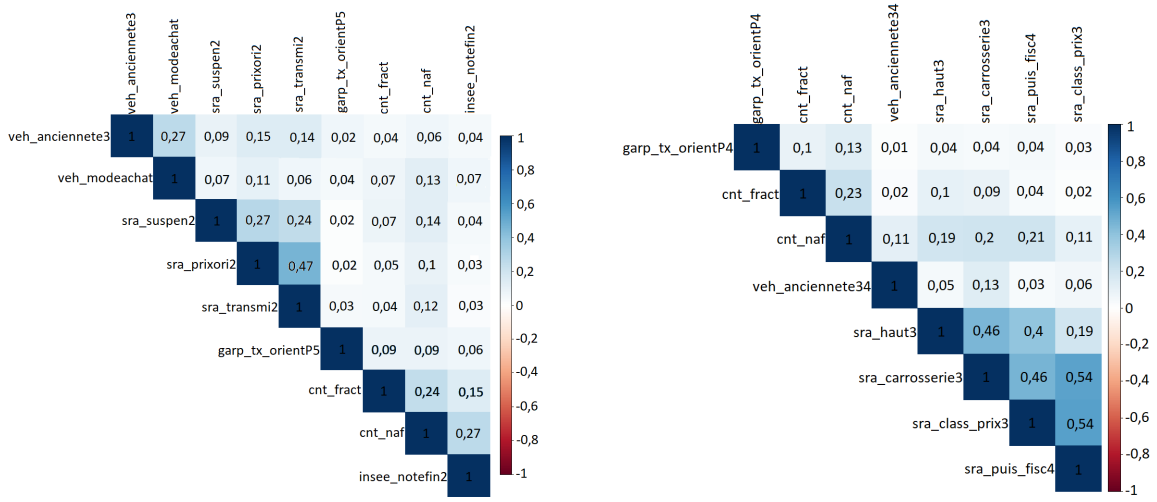


FIGURE 3.28 – Corrélation des variables par V de Cramer

La corrélation au sens du V de Cramer (Figure 3.28) nous alerte sur la corrélation de la transmission du véhicule et le prix à la date de commercialisation du véhicule dans le modèle de charge. De même, dans le modèle de fréquence, quatre variables semblent fortement corrélées entre elles ; la puissance fiscale, la hauteur du véhicule, sa classe de prix et la carrosserie. Cette corrélation est assez facilement explicable. En effet, il s'agit toutes de variables tirées de la base SRA or, on sait que la base de données est constituée de peu de modèles de véhicules proportionnellement au nombre de véhicules. D'ailleurs, cette remarque était à l'origine de la procédure de complétude des codes SRA manquants (sous-section 2.2.3,

page 35). Toutes les variables de la base SRA captent l'information portée par le modèle de véhicule. Il s'agit maintenant de vérifier qu'il n'y a pas de multicolinéarité approximative. Pour ce faire, une mesure supplémentaire sera introduite ; le VIF (*Variance Inflation Factor*).

Le VIF consiste à fournir un indice mesurant de combien la variance d'un coefficient de régression estimée est augmentée en raison de la colinéarité.

Soient $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$ un modèle linéaire de k variables indépendantes, alors l'erreur standard de l'estimation du coefficient β_j, noté $\hat{\beta}_j$, est $\sqrt{(MSE(X^T X)^{-1})_{j+1,j+1}}$ où $X = (X_1, \ldots, X_k)^T$.

On peut définir la variance de $\hat{\beta}_j$,

$$\hat{var}(\hat{\beta}_j) = \frac{s^2}{(n-1)\hat{var}(X_j)} \cdot \frac{1}{1 - R_j^2},$$

où R_j^2 est le coefficient de détermination multiple pour la régression X_j sur les autres X_h.

Le terme $\frac{1}{1-R_j^2}$ est le VIF reflétant tous les autres facteurs influant sur l'incertitude des estimations des coefficients par rapport au premier terme de $\hat{var}(\hat{\beta}_j)$. Le VIF est égal à 1 lorsque le vecteur X_j est orthogonal à chaque colonne de X hormis la $j^{ème}$ colonne. Une version généralisée existe du VIF notée $GVIF = VIF^{-1/2(\times df)}$ disponible dans la fonction *vif* sur **R**.

D'après différents écrits, à partir d'un VIF supérieur à 5, une multicolinéarité de variables est à pressentir.

Dans le cadre de l'étude, une colinéarité a été révélée à partir de la fonction *alias* sur *R* dans le modèle de fréquence et dans le modèle de coût moyen, car pour les deux modèles, respectivement les deux couples de variables (*sra_carrosserie3*, *sra_class_prix3*) et (*sra_transmi*, *sra_suspens*) ont des données manquantes communes. Les données manquantes des variables *sra_carroserie3* et *sra_transmi2* ont été regroupées avec la modalité la plus représentée.

L'information des données manquantes est captée par la seconde variable impliquée dans la colinéarité. L'impact du regroupement n'a donc pas d'effet sur les coefficients. Il est maintenant possible de calculer le VIF des différentes variables du modèle retenu.

```
$vif
                     GVIF Df GVIF^(1/(2*Df))
veh_anciennete3  1.587102 20        1.011615
cnt_naf          2.331419 20        1.021387
sra_prixori2     7.375216 11        1.095076
garp_tx_orientP5 1.144980  9        1.007550
cnt_fract        1.294834  3        1.044004
insee_notefin2   1.447423  3        1.063570
veh_modeachat    1.381761  4        1.041248
sra_transmi2     2.006118  2        1.190115
sra_suspen2      3.584648  2        1.375978
```

```
$vif
                        GVIF Df GVIF^(1/(2*Df))
cnt_naf            2.196556 20        1.019867
veh_anciennete34   1.300214  2        1.067834
sra_haut3         25.351004  7        1.259753
garp_tx_orientP4   1.163662  5        1.015273
sra_class_prix3   10.666495  4        1.344320
cnt_fract          1.236655  3        1.036036
sra_puis_fisc4     3.041640  3        1.203699
sra_carrosserie3  27.813361 10        1.180899
```

FIGURE 3.29 – VIF sur modèle de fréquence (à gauche) et modèle de coût moyen (à droite)

Finalement, les VIF (Figure 3.29) sont toutes strictement inférieures à 2 et donc il n'y a *a priori* pas de problème de multicolinéarité dans les modèles. De plus, le modèle de fréquence et le modèle de coût moyen semblent sélectionner les variables les moins corrélées et avec une multicolinéarité très faible, c'est donc un argument qui vient confirmer le choix de ces deux modèles.

3.4 Modélisation

La modélisation correspond à la prédiction de la charge des sinistres. Dans les étapes précédentes, nous avons :

— défini le modèle linéaire généralisé ou GLM, un modèle paramétrique,

— déterminé la charge à modéliser et les seuils de grave à tester,
— déterminé comment la sélection de variables est réalisée, ainsi que les traitements réalisés sur celles-ci.

Pour la modélisation, plusieurs représentations possibles de la charge seront étudiées.

Dans un premier temps, on peut comparer des modèles dont la charge à modéliser diffère. Par exemple, la charge hors atypique, la charge hors atypiques et graves écrêtées et, enfin la charge hors grave seront testées.

Dans un second temps, il est intéressant de modéliser séparément la responsabilité civile dommage matériel à la responsabilité civile dommage corporel puis de comparer les résultats à un modèle où la responsabilité civile est modélisée dans son intégralité.

Dans un dernier temps, il serait également intéressant de comparer les résultats en utilisant différentes lois pour modéliser la prime pure.

Chacun des modèles sera optimisé par une sélection parcimonieuse des variables et par un lissage. L'objectif est de comparer des modèles tirant le meilleur de chacune des approches. Ainsi, pour un modèle de coût moyen à un autre, il n'y a pas forcément les mêmes variables sélectionnées.

3.4.1 Charge à modéliser ?

Quelle est la charge à retenir pour modéliser au mieux au sens des critères de performance et de validation ?

La charge des sinistres sera traitée de différentes façons :

— charge hors atypique (seuil à 2M€),
— charge hors atypique avec suppression des petits sinistres ($\leqslant$ 120€),
— charge hors grave (seuil à 30k€),
— charge hors grave avec suppression des petits sinistres ($\leqslant$ 120€),
— charge écrêtée des graves (seuil à 30k€).

Le modèle de coût moyen est fortement impacté par la distribution de la charge, contrairement au modèle de fréquence qui sera donc écarté de cette étude. Chaque modélisation d'une charge est optimisée par une sélection parcimonieuse de variables propres à l'outil et un lissage des coefficients par cross-validation (Sous-section 3.3.1, page 63). L'idée est d'optimiser chacun des modèles et, par la suite, retenir la charge impliquant un minimum de biais dans la modélisation. La charge des sinistres graves ont une distribution spécifique, ainsi cette charge ne sera certainement pas correctement modélisée. Il s'agit de mesurer l'impact de l'inclusion et de l'exclusion de la charge des sinistres graves.

Remarque : entre les différents modèles de coûts moyens, les variables et le nombre de variables sélectionnées sont relativement similaires.

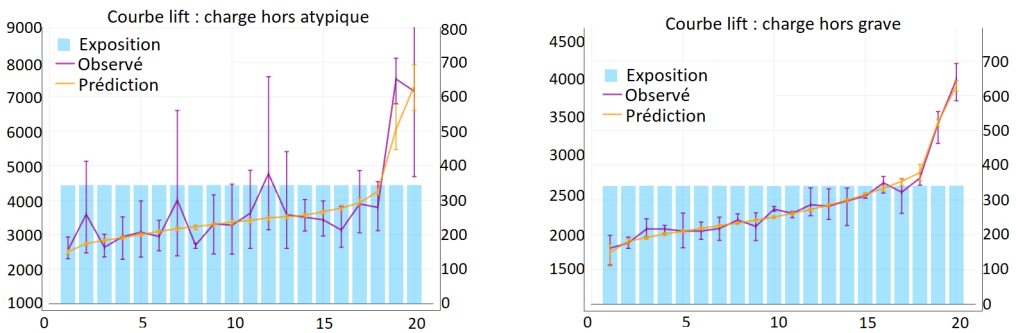


FIGURE 3.30 – Courbe lift, modèle avec charge hors atypique à gauche et hors grave à droite

Graphiquement (Figure 3.30), les charges graves biaisent l'estimation du coût moyen. Le Gini varie fortement selon les folds, indiquant que l'apprentissage est au-delà d'un critère de performance satisfaisant.

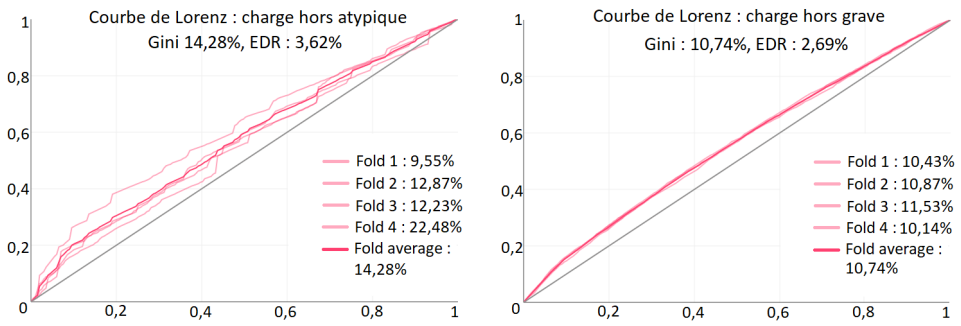


FIGURE 3.31 – Courbe de Lorenz, modèle avec charge hors atypique à gauche et hors grave à droite

De plus, la lift curve de la prédiction (en orange) a une tendance grandement différente de la lift curve des observations [32]. Ce sur-apprentissage est lié aux sinistres graves. En observant les courbes de la modélisation de la charge hors grave, les fortes variations des critères de performance disparaissent.

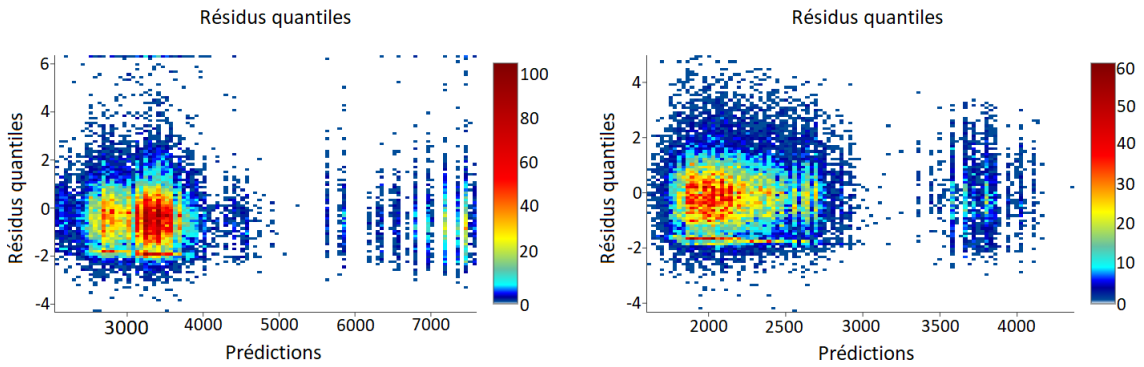


FIGURE 3.32 – Analyse des résidus quantiles modèles avec charge hors atypique à gauche et hors grave à droite

On a supposé *a priori* que les modèles étaient valides. Or, d'après la distribution des résidus quantiles, l'hypothèse sur le modèle de la charge hors atypique est remise en question à cause d'une forte asymétrie des résidus. Cette asymétrie est partiellement supprimée dans le second modèle. Il serait intéressant d'étudier la provenance de cette asymétrie encore présente dans la modélisation de la charge hors grave. De ce fait, la détermination d'un seuil hors grave prend sens, car les sinistres graves suivent une distribution spécifique des autres sinistres et semblent expliquer l'asymétrie.

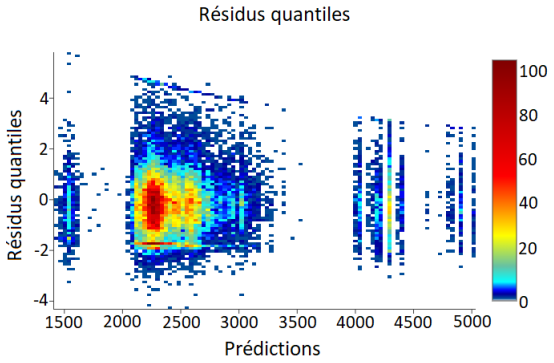


FIGURE 3.33 – Résidus quantiles du modèle de coût moyen avec charge graves écrêtée

32. Pour rappel : la lift curve de la prédiction ou de l'observation correspond à la moyenne respectivement de la prédiction ou de l'observation par classe de 5% de l'exposition et ordonnée par la prédiction.

La modélisation de la charge hors graves ou de la charge écrêtée donne des résultats de performance similaires. Cependant, la charge écrêtée a pour conséquence "d'écrêter" les résidus, c'est-à-dire de créer au niveau des résidus, une tendance non centrée qui n'est pas souhaitée. En effet, en écrêtant, une Dirac est créée égale au seuil des graves qui a un impact négatif sur la modélisation à cause du nombre relativement conséquent de graves.

Peut-on optimiser le seuil des graves afin de traiter seulement les sinistres ayant une distribution commune et représentant une grande partie de la sinistralité ?

Ainsi, deux nouveaux seuils sont introduits en supplément de celui défini par AXA :

— seuil calculé par la Théorie des Valeurs Extrêmes d'un montant de 18 000€,
— seuil introduit par le dAMSE[33] d'un montant de 16 000€. Ce seuil est défini afin d'étudier la sensibilité du seuil et son impact sur la représentation graphique des résidus quantiles.

Seuil	30k	18k	16k
Nombre sinistres au-delà du seuil	294	574	691
en %	0.9%	1.7%	2.0%
Charge au-delà du seuil	44 498 604	50 747 574	52 726 044
en %	36%	41%	43%

FIGURE 3.34 – Sinistres au-delà de chaque seuil

Le seuil qui sera retenu, est celui dont la représentation graphique des résidus quantiles se rapproche au plus d'une loi normale centrée et réduite en supprimant un nombre minimal de sinistres graves (Figure 3.34).

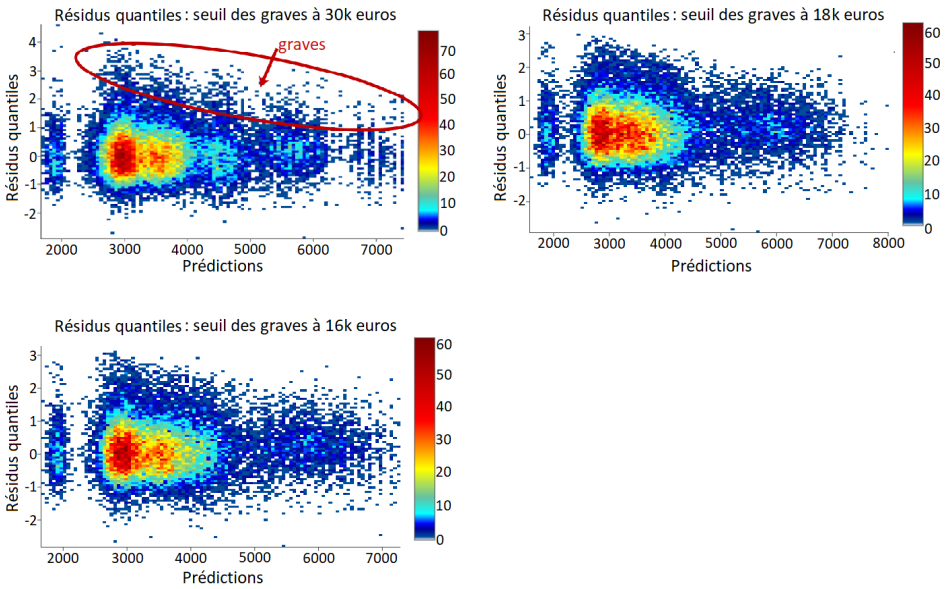


FIGURE 3.35 – Analyse des résidus quantiles du modèle coût moyen sur différents seuils de graves

Figure 3.35, le seuil défini par la TVE améliore l'ajustement du modèle. Une grande partie des résidus non centrée est supprimée (les échelles entre les graphiques sont différentes) et un léger gain de performance est constaté. Les charges graves ont été mutualisées pour comparer des modèles ajustés sur des

33. Pour rappel, le dAMSE est la minimisation de l'erreur quadratique de l'estimateur de Hill

bases équivalentes en matière de données, et charge totale. Un seuil fixé à 16 000€ n'engendre pas d'amélioration apparente du modèle. Les performances entre les trois modèles sont relativement similaires, c'est pourquoi ils n'ont pas été détaillés.

Au regard des résultats, un seuil des sinistres graves à 18 000€ est défini.

Enfin, de petits montants de charge peuvent biaiser le modèle. En effet, (Sous-section 3.2.1, page 49)cette sinistralité est liée à des indemnisations causé par les frais d'expertise. Un sinistre peut être assujetti à une expertise et peut être contesté en cas d'exclusion des clauses du contrat. En général, les petits montants de sinistre ne reflètent donc pas la réalité de la sinistralité du risque étudiée. En comparant un modèle où ces petits montants sont supprimés et un modèle comprenant ces petits montants, aucune différence significative n'est constatée. Ainsi, la charge des petits sinistres sera conservée.

La fréquence et le coût moyen sont entrainés sur un nombre différent de données selon la charge étudiée. Il a paru donc judicieux de comparer les ajustements au sens des critères de performances sur les données test après agrégation des modèles de fréquence et de coût moyen.

	hors grave >= 120	hors grave
Gini	33.2%	33.1%
Gini normalized	33.6%	33.5%
RMSE	1187	1187
MAE	171	173

FIGURE 3.36 – Performance des modèles

3.4.2 Mutualisation des graves ?

Comment la charge grave est-elle mutualisée ?

La modélisation est construite hors sinistres grave (charge supérieure ou égale à 18000€), car elle présente de meilleurs résultats sans ces charges qui biaisent la distribution du modèle paramétrique. Cependant, la charge de ces sinistres doit tout de même être prise en compte et redistribuée. Ces sinistres sont rares et par conséquent ils sont difficiles à modéliser de façon paramétrique.

Chez AXA Entreprises, la charge des atypiques est prise en compte dans le calcul du PLR [34] segmenté par, notamment, le segment NAF. La Nomenclature d'Activité Française est une nomenclature des activités économiques productives, principalement élaborée afin de faciliter l'organisation de l'information économique et sociale. Ainsi, lors du passage de la prime pure à la prime commerciale grâce au PLR, les sinistres atypiques seront automatiquement captés.

Pour la sinistralité, plusieurs choix sont possibles afin d'éviter que les sinistres graves biaisent la modélisation :

— charge hors grave : il est intéressant dans un premier temps d'exclure les graves. Un taux moyen observé sur les dernières années serait alors appliqué pour prendre en compte cette charge supplémentaire,
— charge écrêtée : un écrêtement consiste à plafonner le montant des sinistres selon le seuil des graves. Cette méthode permet de ne pas exclure les extrêmes en les comptabilisant dans la modélisation, mais en minimisant leur impact dans la régression, ce qui revient à modéliser seulement la charge attritionnelle. Cependant, cette méthode est en général déconseillée, car elle présente le défaut de créer une Dirac autour du seuil des graves et donc de tirer à la hausse les charges prédites. C'est d'ailleurs ce qui a été observé à la sous-section précédente.

La charge des graves ou la sur-crête doit être prise en compte dans le tarif, car elle fait partie intégrante de la charge et donc de l'estimation de la prime pure. Il existe plusieurs moyens de redistribuer ces montants de sinistres : sous différentes mailles (maille contrat, maille agent, courtage ...) ou différents segments (NAF, catégorie de véhicules ...), ou encore en segmentant cette charge par un modèle paramétrique type GLM, si la distribution appartient à une famille exponentielle, ou par des modèles

34. Pour rappel, le PLR est un indice économique de passage entre la prime pure et la prime commerciale, prenant en considération l'ensemble des coûts liés à l'activité d'assurance.

non-paramétriques, tels que des arbres de régression. Cependant, la quantité faible de sinistres graves due à un jeu de données contenant un faible historique, ne nous permettrait pas de réaliser une segmentation suffisamment robuste. Ainsi, il semble préférable d'émettre l'hypothèse que les graves suivent approximativement la même segmentation que les attritionnels. La charge des graves est ajoutée au prorata de la prime des sinistrés calculés hors grave.

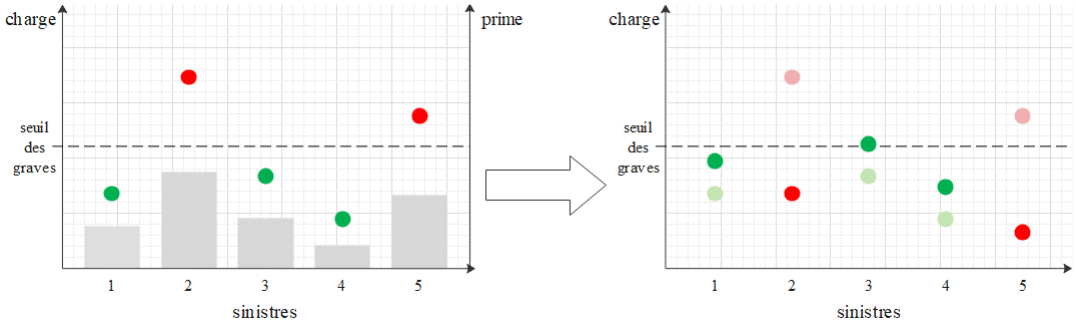


FIGURE 3.37 – Mutualisation au prorata de la prime

À chaque sinistre est associé une charge, point vert ou rouge, et une prime, barre en gris (Figure 3.37). L'ensemble des charges associé à des sinistres graves, c'est-à-dire dont le montant dépasse le seuil des graves (points rouges) est mutualisé sur l'ensemble des sinistres. Ainsi, la charge d'un sinistre sera estimée selon la formule suivante :

$$charge\ mutualisée_i = charge_i + charge\ totale\ des\ graves \times \frac{prime_i}{prime\ totale\ des\ sinistres},$$

où $i \in S$ et S est l'ensemble des sinistres.

Remarque : pour un sinistre grave, $charge_i = 0$, car il s'agit de la charge hors grave.

3.4.3 RC ou RC CORP/RC MAT ?

Est-il préférable de modéliser séparément d'un côté la responsabilité civile liée à des dommages corporels (RC CORP) et de l'autre, la responsabilité civile liée à des dommages matériels (RC MAT) ?

Cette étude a été faite dans un premier temps hors atypique et dans un second temps hors graves avec le seuil défini par AXA (seuil à 30 000€). Seule l'étude hors grave sera décrite ci-après, car les résidus quantiles des modèles hors atypique présentaient une asymétrie pour la RC CORP et la RC MAT (ainsi que la RC). Trois arguments viennent en défaveur du traitement à part de la RC CORP et de la RC MAT :

— le modèle de coût moyen de la RC CORP a une performance insatisfaisante,
— la RC CORP est composée en majorité de graves. Ainsi, hors graves, les performances ne sont pas améliorées. De plus, le modèle de fréquence est impacté par une forte variation selon les folds,
— un modèle fréquence×coût moyen pour les deux garanties a tendance à augmenter le nombre de variables à conserver pour des raisons de performances. Le modèle agrégé est donc complexifié pour des performances relativement similaires au modèle RC.

Pour conclure, le modèle final sera ajusté sur la charge RC hors grave, seuil à 18 000€. La charge des graves est mutualisée au prorata de la prime des sinistrés avant modélisation.

3.4.4 Lois ?

Quelle loi est la plus adaptée au jeu de données ?

La charge à modéliser a été définie, ainsi que le risque. La charge est hors grave au seuil de 18 000€ et le risque correspond à la RC. En revanche, dans un modèle paramétrique tel que le GLM, il est supposé

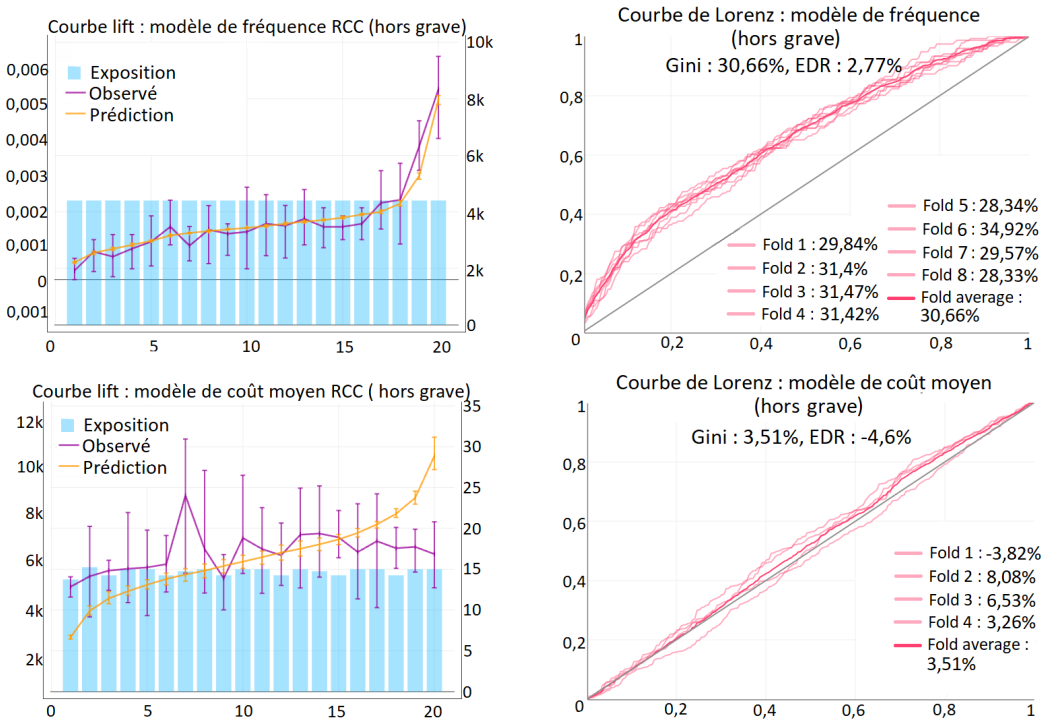


FIGURE 3.38 – Analyse des performances du modèle RCC

que les données suivent une loi. Dans un premier temps, pour vérifier l'adéquation d'une loi, il suffit d'observer les résidus quantiles après modélisation. Si les résidus suivent une loi normale centrée et réduite, on peut en déduire que la loi utilisée est en adéquation avec la variable à prédire. Le modèle de départ est valide.

Pour le modèle de fréquence, la loi de Poisson semble parfaitement suivre l'allure des données :

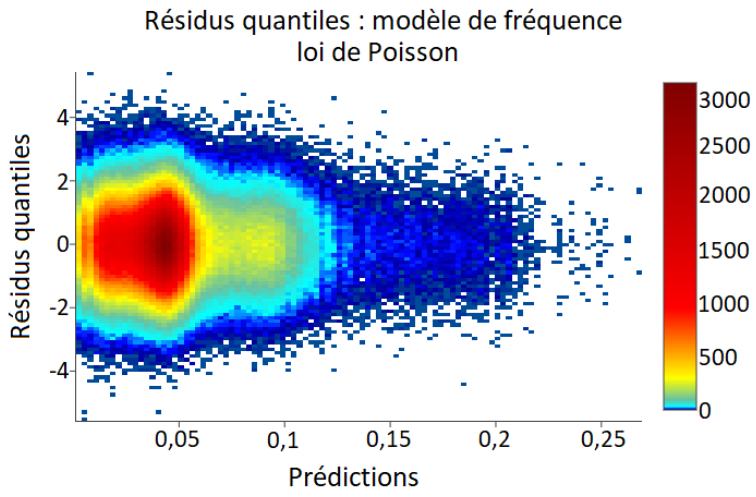


FIGURE 3.39 – Résidus quantiles du modèle de fréquence RC

En conclusion de ces résultats, le modèle de fréquence est modélisé avec la loi de Poisson. Couramment utilisée pour le modèle de fréquence, la loi binomiale négative est une alternative à loi poisson qui ne se limite plus à une moyenne empirique égale à la variance empirique. Elle permet de décrire des observations sur-dispersées grâce à un paramètre supplémentaire dissociant la variance de la moyenne. De plus, la loi binomiale négative tend faiblement vers une loi de Poisson. Ainsi, si les données sont en adéquation avec une loi de poisson, il est en général de même pour la loi binomiale négative et il devient donc difficile de déduire laquelle des deux lois conviennent le mieux.

En revanche, pour le modèle de coût moyen, une légère asymétrie est perceptible qui s'explique principalement par la présence de sinistres graves. Il serait intéressant de regarder si l'asymétrie résiduelle peut être corrigée par l'utilisation d'une autre loi que la loi gamma jusqu'alors conservée. Il se trouve que la loi inverse gaussienne ne sera pas retenue. La représentation des résidus quantiles présente un centrage de 2 au lieu de 0 (Figure 3.40).

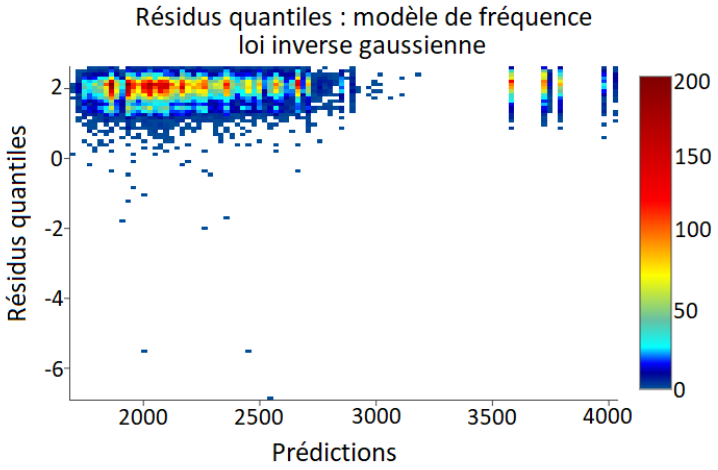


FIGURE 3.40 – Résidus quantiles modèle de coût moyen, loi inverse gaussienne

Enfin, la méthode utilisée jusqu'à présent est le modèle collectif supposant l'indépendance de la fréquence du coût moyen. Il est cependant possible de déterminer la prime pure directement à partir de la loi Tweedie qui est une loi de poisson composée avec une loi gamma.

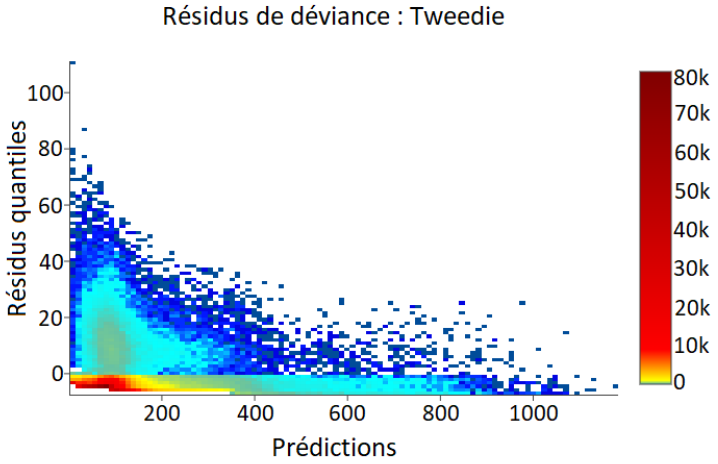


FIGURE 3.41 – Résidus de déviance modèle de prime pure, loi Tweedie

Les résidus de déviance non standardisés pour une loi Tweedie suivent à peu près une loi normale dont le centrage et la réduction ne sont pas attendus. Graphiquement, les résidus ne suivent pas une loi gaussienne, ainsi le modèle n'est *a priori* pas valide. D'autre part, la performance du modèle est similaire au modèle collectif agrégé au sens des métriques définies précédemment (comme le Gini) sur les données test.

Dans la plupart des cas, le modèle collectif est plus intéressant que le modèle de prime pure pour sa transparence et son interprétabilité. En effet, deux modèles, fréquence et sévérité, permettent de

différencier les effets des variables explicatives différentes selon le coût moyen et la fréquence. Ainsi, des informations plus précises sur les comportements sont accessibles.

3.5 Zonier

Plus le nombre de variables et/ou plus le nombre de modalités dans la modélisation est important, plus le modèle devient saturé, c'est-à-dire qu'il y a un risque de surapprentissage. Or, la variable géographique, selon la maille retenue, peut être composée de milliers de modalités. Elle implique donc la création de milliers de paramètres et une difficulté d'estimation liée à la disparité des données. En effet, certaines zones peuvent être caractérisées par une exposition faible, voire nulle. Pourtant, la géographie est une segmentation intéressante. Il convient donc de réfléchir à une approche spécifique afin de prendre en compte l'information géographique.

Un zonier a pour objectif de construire un facteur exprimant l'effet géographique. En fait, lors de la modélisation, les variables portant sur la géographie du risque ont été expressément omises. La prise en compte de la dimension géographique est réalisée sur les résidus de la modélisation. Un zonier est ainsi construit par projection sur une dimension géographique : une carte. On émet l'hypothèse que l'information géographique explique entièrement la variation résiduelle. Lors de la création d'un zonier, il est attendu une amélioration sur les critères d'évaluation (Gini) et les critères de validation (variation des résidus) du modèle.

Remarque : les variables incluses dans un modèle ne sont seulement qu'un sous-ensemble de l'univers des variables explicatives. Ainsi, en plus du bruit (partie aléatoire du risque ne pouvant être expliquée), une partie de la variation résiduelle non expliquée par les variables retenues subsiste dans l'effet résiduel.

En pratique, il est préférable de considérer une unité géographique la plus fine possible, puis à l'aide d'étapes parcimonieuses, de calibrer un zonier mieux défini au sens des critères choisis. Trois mailles seront considérées :

— la maille la plus fine est le code INSEE lié à un couple unique nom de commune et code postal,
— le code postal,
— le département pour pouvoir comparer les données propres au risque étudié et les données nationales agrégées à la maille département notamment sur l'état des routes ou la géographie des accidents de la route nationale.

Pour éviter une granularité trop importante d'une zone à une autre, c'est-à-dire des coefficients relativement différents, l'effet doit être lissé avec celles des voisins géographiques. En fait, le signal géographique est considéré continu. Par conséquent, il est attendu à ce que les coefficients évoluent en douceur entre les points voisins. La création d'une carte par projection prend son sens pour visualiser le bon lissage des coefficients.

Le lissage peut permettre une amélioration de la prédiction du modèle au sens du critère d'évaluation. En effet, cela permet de remettre en cause la fiabilité des observations liées à une faible exposition et de prendre en compte la vitesse à laquelle des emplacements proches ou éloignés peuvent s'influencer mutuellement. Un lissage est un compromis entre ces deux phénomènes. Plusieurs niveaux de sensibilité au signal observé dans la base de données seront réalisés et visualisés sur une grille évaluant le compromis entre les performances (Gini) et la complexité (indice de Moran). Un niveau de lissage élevé correspond à des coefficients dont la variance est faible sur de petites distances. Réciproquement, un faible lissage crée des coefficients avec une plus grande oscillation sur une petite échelle géographique.

Pour construire un lissage des plus proches voisins, on définit une distance, en l'occurrence ici la distance de Moran[35]. La distance de Moran correspond au rayon d'une région circulaire qui contient des points ayant un score de corrélation de 50% entre les distances relatives et les coefficients. Pour un observateur donné, placé dans une zone fixée, la distance de Moran donne une distance moyenne à parcourir pour observer un changement important du coefficient de risque associé à la variable géographique.

Ainsi, plusieurs modèles seront construits avec différentes distances de Moran. L'étude sera suivie d'une sélection qui retiendra un zonier maximisant les critères de performance tout en apportant un

35. La théorie de cet indice ne sera pas présentée ici, ne faisant pas l'objet de ce mémoire, pour plus de détails se référer à l'étude de OLIVEAU, 2010

lissage suffisamment important, afin que d'une année de survenance à l'autre, le zonier soit relativement similaire, *time consistency*.

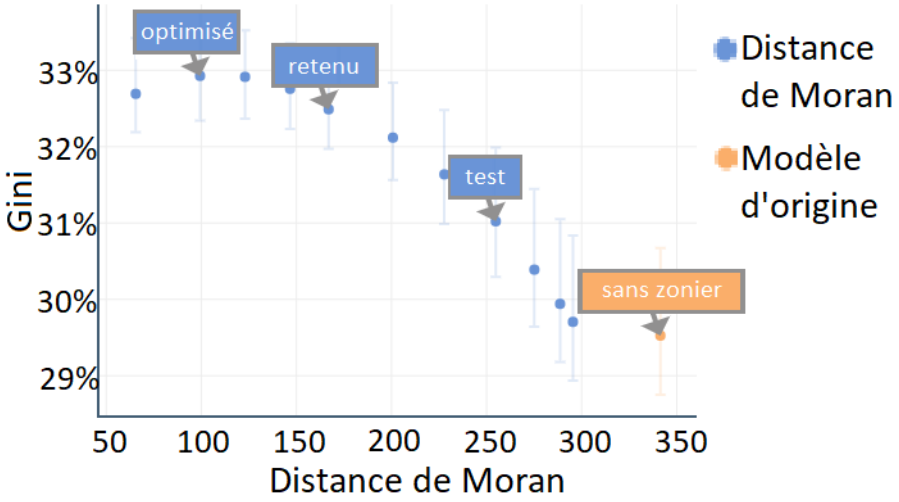


FIGURE 3.42 – Grille de recherche parcimonieuse du zonier modèle de fréquence

Sur ce graphique (Figure 3.42) est représenté un exemple type de sélection judicieuse du zonier. En fait, le zonier optimisé donne de meilleurs résultats au sens du Gini, mais il n'est pas retenu, car le lissage n'est pas assez approfondi. C'est visuellement et à l'aide de la dimension *time consistency* que le zonier est choisit. La dimension *time consistency* n'est pas développée afin de ne pas alourdir ce mémoire.

Afin de regrouper des zones entre elles et ainsi d'obtenir un zonier à une maille moins fine, l'étape parcimonieuse est réalisée à partir d'un paramètre de quantisation correspondant au nombre de zones, introduit dans le zonier. L'approche générale, inclue dans l'outil pour le calcul des coefficients géographiques, suit le principe selon lequel chaque zone (code INSEE) se verra attribuer son propre coefficient. Cependant, il est possible d'effectuer un regroupement automatique des coefficients, adapté à une approche zonale par la quantisation. Par conséquent, il n'y aura au final qu'un nombre préfixé de zones. La variable géographique sera recatégorisée. Le regroupement des zones a l'avantage de créer de nouvelles zones avec une plus grande exposition et donc une représentation plus robuste sur une dimension géographique. En revanche, par exemple, une quantisation à deux zones, aura pour conséquence de dissimuler des informations sur certaines régions. Ainsi, il faut trouver un compromis entre lissage, performance et quantisation. En produisant plusieurs étapes de lissage sur plusieurs paramètres de quantisation, 20 zones sont retenues pour le modèle de fréquence et le modèle de coût moyen.

Remarque : couramment, un zonier est produit sur le seul modèle de fréquence. Cependant, un zonier sur le modèle de charge semble pertinent, car un gain de performance relativement élevé est constaté. On soupçonne que ce gain de performance est causé par la mutualisation des sinistres graves au prorata des primes sinistrés. En effet, avant mutualisation, ce gain de performance était relativement faible.

Il serait maintenant intéressant d'expliquer les zones à risques et les zones à faible risque, c'est-à-dire respectivement les zones de coefficient élevé (> 1) et les zones de coefficient faible (< 1).

Deux bases de données externes ont été appelées afin d'expliquer l'effet géographique :

— fichier BAAC 2019 : cette base constituée par l'ONISR [36] recense les accidents corporels survenus en 2019 en France sur une voie ouverte à la circulation publique, impliquant au moins un véhicule

36. Observatoire National Interministériel de la Sécurité Routière

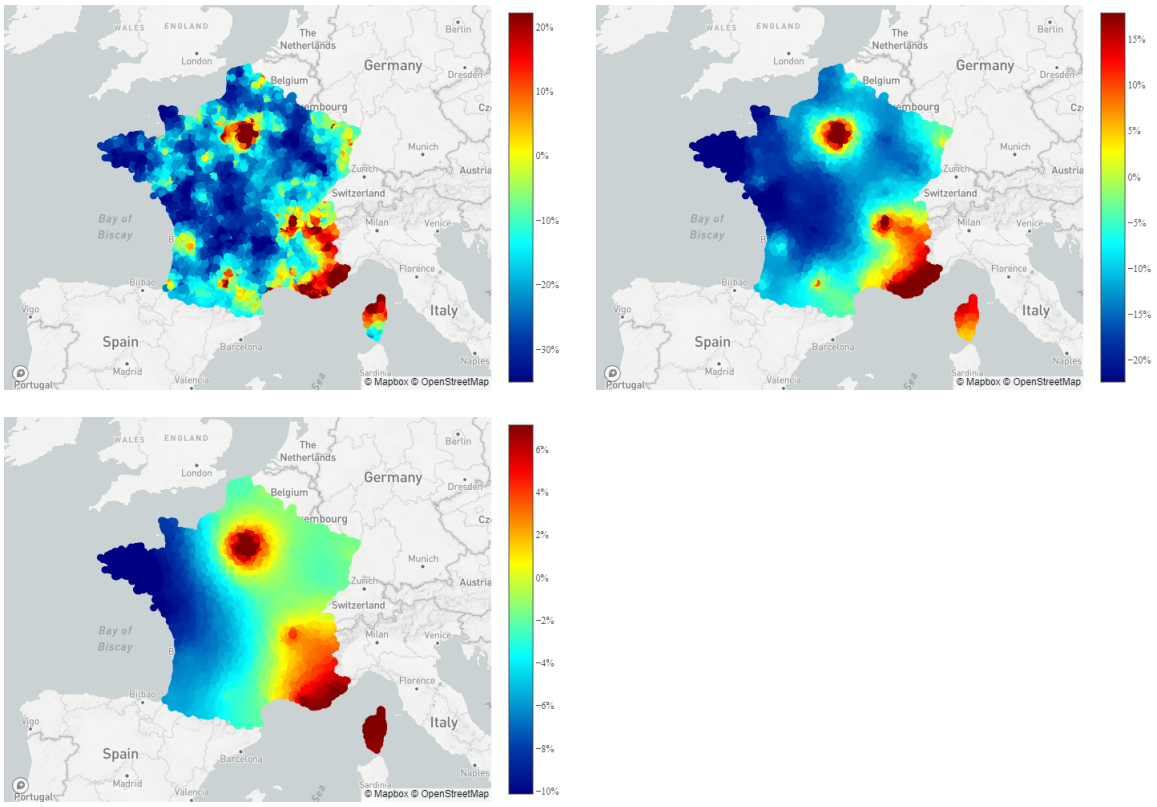


FIGURE 3.43 – Zonier sur le modèle de fréquence *(optimisé en haut à gauche, retenu en haut à droite et test en bas à gauche)*

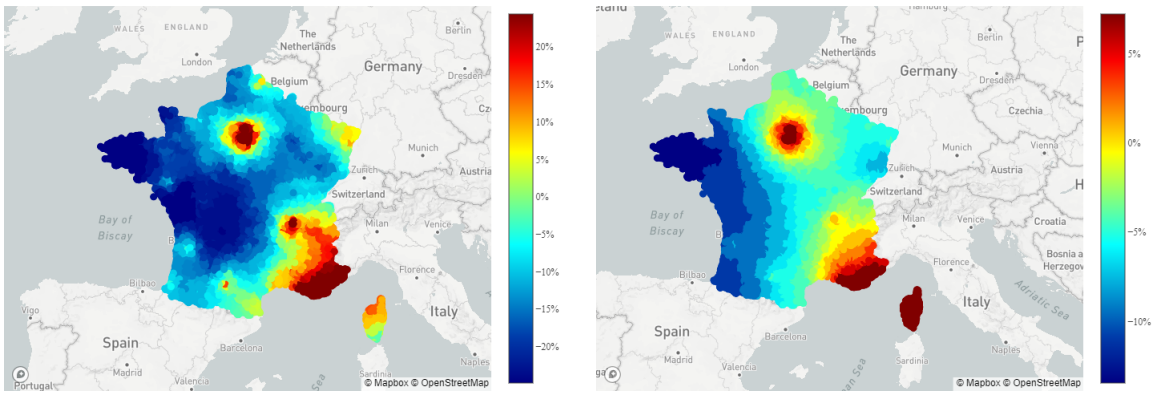


FIGURE 3.44 – Zonier sur le modèle de fréquence à gauche et zonier sur le modèle de charge à droite

et au moins une victime ayant nécessité des soins. Des informations décrivant l'accident sont saisies par les forces de l'ordre. Les données, dont le nombre de sinistres et la météo lors de l'accident, sont agrégés à la maille département,

— état des routes : cette base a été tirée de DATA.GOUV, 2020 et donne une évaluation de l'état des chaussées du réseau routier national non concédées à partir de 2019. Les données de notation sur l'état de chaque route ont été agrégées à la maille département selon la moyenne de la note pondérée par la distance des routes.

Après agrégation de la fréquence de sinistre à la maille département pondérée par l'exposition, les corrélations sont comparées entre les différentes données sur la variable réponse.

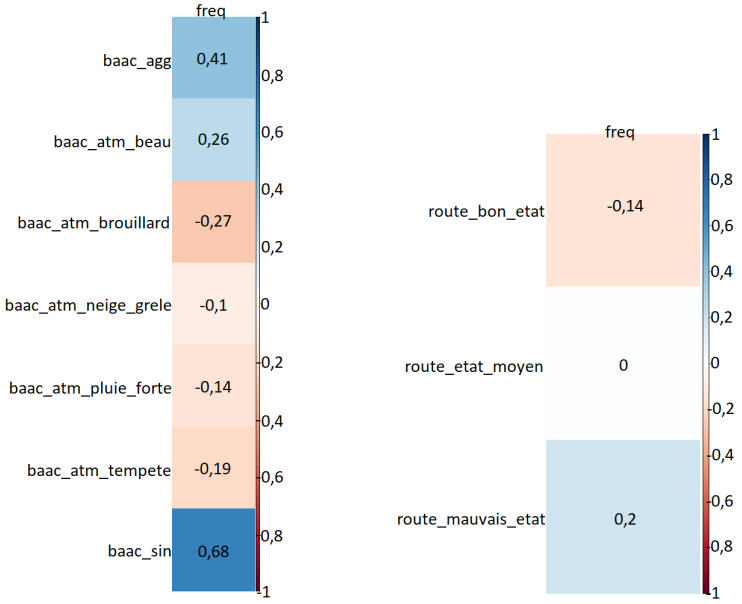


FIGURE 3.45 – Corrélation entre la fréquence RC et les informations géographiques

La fréquence de sinistre semble être plus élevée dans les départements où la notation moyenne de l'état des routes est mauvaise par rapport à la moyenne observée et où la météo est plus propice à des journées ensoleillées (sud-est de la France et Île-de-France). Les sinistres BAAC sont fortement corrélés aux sinistres RC. Pour s'assurer qu'il n'y ait pas de mauvaises interprétations, on peut tracer les cartes de ces différentes variables :

Cette étude est à nuancer. Les corrélations entre les variables sont faibles et les cartes s'éloignent de l'information extraite des zoniers. En effet, la carte sur le mauvais état des routes recense des routes de mauvais états autour du pays basques pourtant le zonier ne présente pas de sur-sinistralité dans cette région et inversement pour l'Île-de-France. La carte à propos du beau temps capte mal la sur-sinistralité du coté nord-est de la France.

Remarque : la notion de météo est définie selon le sinistre corporel associé dans le fichier BAAC, cela ne correspond pas a priori ici aux départements où le nombre d'heures ensoleillées est le plus élevé. À propos du nombre d'heures d'ensoleillement par département, on sait que le nord de la France est moins ensoleillé que le sud de la France, à l'exception du sud-ouest.

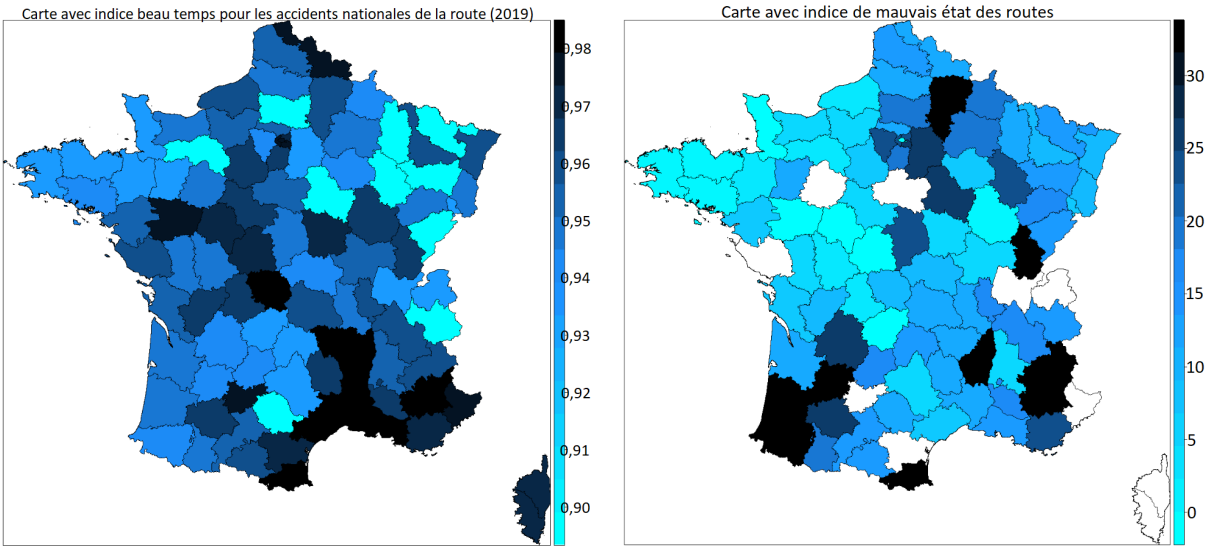


FIGURE 3.46 – Cartes de quelques informations géographiques

3.6 Interactions

Dans le cadre de la modélisation avec des méthodes de type GLM, les variables explicatives catégorielles x^{l_1} et x^{l_2}, composées respectivement de K_1 et K_2 modalités, interviennent dans la composante systématique du modèle via les coefficients respectifs $\beta_{(l_1,1)},\ldots,\beta_{(l_1,K_1-1)}$ et $\beta_{(l_2,1)},\ldots,\beta_{(l_2,K_2-1)}$.

$$\eta = \underbrace{\sum_{k_1=1}^{K_1-1} \beta_{(l_1,k_1)} \mathbb{1}(x^{(l_1)} = a_{k_1}^{(l_1)})}_{effet\ de\ x^{l_1}} + \underbrace{\sum_{k_2=1}^{K_2-1} \beta_{(l_2,k_2)} \mathbb{1}(x^{(l_2)} = a_{k_2}^{(l_2)})}_{effet\ de\ x^{l_2}} + \underbrace{\ldots}_{effet\ des\ autres\ var.},$$

où $\eta = \sum_{j=0}^{d} \beta_j x(j) = g(\mu)$ avec g la fonction de lien et $\mu = \mathbb{E}(Y|X = x)$.

L'effet des variables explicatives sur η est donc purement additif. L'utilisation de termes d'interaction entre $x^{(l_1)}$ et $x_{(l_2)}$ dans l'ajustement permet de modéliser un effet conjoint de ces deux variables sur η. Deux manières classiques d'introduire ces termes sont classiquement utilisées :

— interaction dite complète où une variable catégorielle $x^{(l_1,l_2)} = (x^{(l_1)}, x^{(l_2)})$ décrit les valeurs conjointes de $x^{(l_1)}$ et $x^{(l_2)}$. Cette méthode a été appliquée en utilisant certaines variables côté PP (professionnel-particulier) comme le ratio poids puissance. Ces variables n'ont pas été sélectionnées par le modèle,
— interaction dite marginale où les coefficients individuels associés à $x^{(l_1,l_2)}$ sont conservés en ajoutant des coefficients supplémentaires liés aux valeurs conjointes de $x^{(l_1,l_2)}$.

L'hypothèse de l'indépendance des variables explicatives entre elles est validée par l'analyse des corrélations. Cependant, il est possible de choisir de ne pas ignorer la relation entre deux variables dans le modèle par des termes d'interaction, ainsi l'indépendance des variables n'est plus validée. Des termes d'interactions marginales sont ajoutés en fin de modélisation afin d'enrichir la qualité du modèle au sens du critère de performance, le Gini. La sélection des termes d'interaction se déroulent de la même façon que pour les variables précédemment définies (Sous-section 3.3.1, page 63). Cependant, une première étape consiste à réaliser une présélection des termes d'interactions :

— ces termes doivent avoir un sens opérationnel,
— un écart entre observé et prédit doit être relativement élevé sur certains termes d'interaction afin de capter une nouvelle information,
— une interaction statistiquement intéressante à étudier a un score élevé défini dans l'outil, qui est un indicateur univarié. Ce score donne une bonne indication si la variable est potentiellement prédictive ou non. Le niveau d'exposition est pris en compte.

Dans un premier temps, on peut tracer l'observé selon la prédiction pour chaque couple de modalités croisées (Figure 3.47). Dans un second temps, il est possible de visualiser les écarts (couleur du point) de

ces modalités avec un poids sur l'exposition (taille du point). Dans l'exemple suivant, c'est le croisement entre la variable *insee_notefin* et *sra_prixori* qui est étudié.

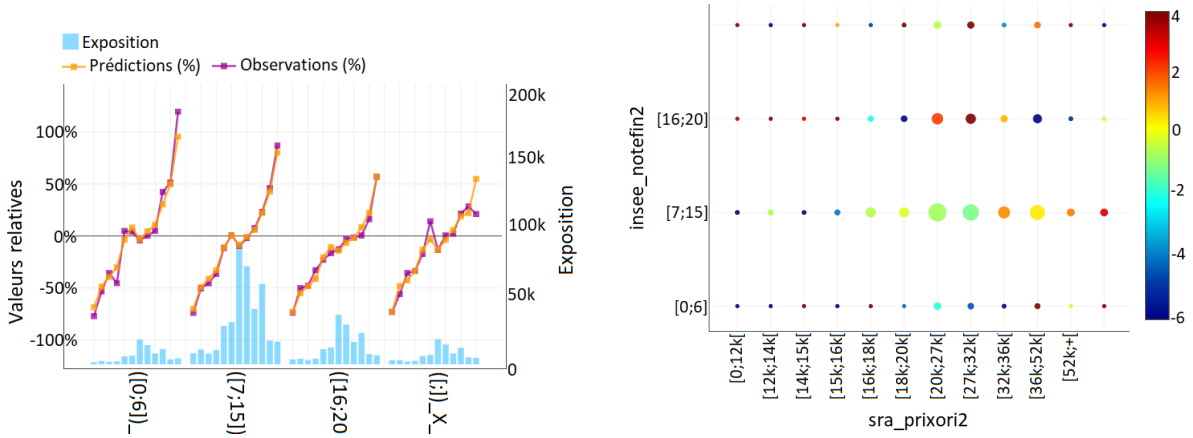


FIGURE 3.47 – Analyse de l'interaction marginale

Pour le modèle de fréquence ou le modèle de coût moyen, aucun terme d'interaction n'est conservé.

Chapitre 4

Véhiculier

Le regroupement de modalités et le zonier sont deux méthodes permettant de conserver l'information importante des variables. Cependant, le nombre de variables a été limité par une étape de sélection. Ainsi, ces deux étapes successives ont peut-être supprimées des informations importantes. Les variables véhicules apportent peu d'informations sur la réponse d'après la significativité des variables. Afin de s'en assurer et de capter un maximum les effets de ces variables, un véhiculer sera construit. De plus, peu de renseignements sur leurs effets sont à disposition. Une analyse devient donc nécessaire.

Trois méthodes de véhiculier seront étudiées :
- un véhiculier construit à partir d'un arbre CART sur l'effet véhicule,
- un véhiculier élaboré également par un arbre de décision, mais précédé par une crédibilisation de l'effet véhicule en fonction des années d'exercice,
- un véhiculier fondé par un arbre de régression dont la variable à expliquer est l'effet véhicule lissé par une cartographie des caractéristiques techniques des voitures.

4.1 Classification des véhicules

Le modèle sans contrainte apporte une information supplémentaire sur les variables véhicules par rapport au tarif actuel. Il serait donc intéressant d'étudier l'impact de l'effet véhicule sur les performances des ajustements. Un traitement du signal est proposé à partir de plusieurs véhiculier permettant de classifier l'effet véhicule à partir des caractéristiques techniques.

4.1.1 Segmentation

Dans un marché qui devient de plus en plus concurrentiel, une segmentation précise du risque permet de refléter de façon plus juste la sinistralité. Cependant, une segmentation trop fine peut remettre en cause le principe de mutualité des risques qui est la notion clé en assurance. En assurance automobile, le véhiculier est utilisé à des fins de regroupement de véhicules en risques homogènes.

L'intérêt de la segmentation est d'affiner l'appréhension du risque et d'adapter la prime au profil de risque des assurés évitant ainsi un phénomène d'antisélection résultant de l'asymétrie d'information entre assuré et assureur. En effet, une segmentation demande une plus grande quantité d'informations sur le risque à couvrir et permet une distinction de certains groupes homogènes qui sont plus exposés au risque à ceux qui ont une tendance à moins l'être.

Mathématiquement, dans le cadre d'une segmentation nulle le transfert de risqué s'opère de la manière suivante :

	Assuré	Assureur
Coût	$\mathbb{E}[S]$	$S - \mathbb{E}[S]$
Coût moyen	$\mathbb{E}[S]$	0
Volatilité	0	$\mathbb{V}[S]$

L'assureur porte toute la variabilité de la sinistralité en échange d'une prime. Dans le cadre où Ω représente l'ensemble des informations des assurés influant le risque que l'assureur est amené à utiliser, le transfert de risque s'opère de la façon suivante :

	Assuré	Assureur
Coût	$\mathbb{E}[S\|\Omega]$	$S - \mathbb{E}[S\|\Omega]$
Coût moyen	$\mathbb{E}[S\|\Omega]$	0
Volatilité	$\mathbb{V}[\mathbb{E}[S\|\Omega]]$	$\mathbb{V}[S - \mathbb{E}[S\|\Omega]]$

La prime payée par un assuré au hasard est une variable aléatoire car l'espérance pour un individu i de cotiser une somme $\mathbb{E}[S|\Omega]$ dépend des caractéristiques des informations ω_i données à l'assureur. L'assureur restitue alors à l'assuré la partie de volatilité qui correspond à Ω. D'ailleurs, $\mathbb{V}[S] = \mathbb{V}[\mathbb{E}[S|\Omega]] + \mathbb{E}[\mathbb{V}[S|\Omega]] = "Solidarité" + "Risque" ou "Biais" + "Variance"$. Dans le cadre où Ω est connu, l'assureur conserve et opère une mutualisation du risque pur. En revanche, la mutualisation des risques est dissociée de la notion de solidarité. Un groupe homogène qui a une exposition plus faible par rapport à un autre groupe ne se verra pas attribuer une partie du risque de ce deuxième groupe. Néanmoins, dans la pratique, Ω est inconnu. Seul $X \subset \Omega$ un sous-ensemble est connu. Il s'agit d'une segmentation avec information imparfaite :

	Assuré	Assureur
Coût	$\mathbb{E}[S\|X]$	$S - \mathbb{E}[S\|X]$
Coût moyen	$\mathbb{E}[S\|X]$	0
Volatilité	$\mathbb{V}[\mathbb{E}[S\|X]]$	$\mathbb{V}[S - \mathbb{E}[S\|X]]$

La tarification de la prime pure est une approche moyenne et n'assure donc pas la probabilité. Une segmentation du risque, c'est aussi la connaissance du risque. Cela permet une sélection positive protégeant la profitabilité à court terme et générant de la croissance à moyen et long terme.

Il existe néanmoins des règles en matière de segmentation, par exemple il est interdit de tarifer selon le sexe de l'assuré suite à l'arrêt de la Cour de Justice de l'Union Européen (CJUE) en mars 2011. Ces règles à but éthiques sont fixées pour s'opposer aux phénomènes de discrimination.

4.1.2 Introduction au véhiculier

Dans un véhiculier, c'est l'effet véhicule qui est expliqué et segmenté. L'effet véhicule se définit comme la part de variable d'intérêt (fréquence ou coût moyen) liée aux véhicules, c'est-à-dire l'effet des variables véhicules pour prédire la variable réponse. À ce jour, aucune étude actuarielle n'a été réalisée sur les facteurs de risque véhicule à partir d'un véhiculier en auto-entreprise chez AXA. En revanche, d'après plusieurs écrits (LAVENU, 2016) et à partir de la segmentation intégrée dans les GLM, certaines caractéristiques véhicules ont *a priori* un fort impact sur la segmentation. Par exemple, il y a la marque du véhicule, le nombre de portes, le nombre de places, le type de transmission ou encore le carburant.

D'après LAVENU, 2016, il existe trois approches avancées de classification automobile :

— un GLM permet d'introduire une variable composée d'un nombre conséquent de modalités dont l'effet est lissé par crédibilité,
— un arbre classifie l'effet véhicule. La théorie de la crédibilité peut être ajoutée à des fins de robustesse,
— une cartographie à partir des corrélations des informations techniques véhicules définit un lissage de l'effet véhicule qui sera par la suite catégorisé par une arbre.

Le fichier SIV et la base SRA sont les deux clés à la constitution d'un véhiculier. En effet, ils permettent de compléter les données manquantes et de fiabiliser les données existantes. Le code SRA, noté *sra_ code*, définit un type de véhicule et correspond à une ligne de la base SRA, c'est donc la maille la plus fine pour définir les caractéristiques techniques d'un véhicule.

Traiter la variable *sra_ code* comme une variable catégorielle dans un GLM n'est pas envisageable, car elle présente de nombreuses modalités, noté $x^{(l)}$. Il est possible de traiter l'effet de $x^{(l)}$ comme une variable aléatoire, en utilisant une approche de type crédibilité pour l'estimation. L'approche proposée par *OHLSSON E.* en 2007 conduit à écrire pour un GLM log-Poisson, lorsque $x^{(l)} = a$, les variables $x^{(k)}$ pour $k \neq l$ étant encodées par les $x^{(j)}$, $1 \leqslant j \leqslant d^{(-l)}$,

$$\mathbb{E}(Y|X = x, U_a) = U_a.exp(\sum_{j=0}^{d_{-l}} \beta_j x^{(j)}),$$

où l'effet aléatoire U_a est estimé par $\hat{U}_a = z_a\bar{\bar{y}} + (1 - z_a)$, les z_a sont des poids de crédibilité et $\bar{\bar{y}}_a$ est une moyenne pondérée des valeurs normalisées des réponses y_i lorsque $x_i^{(l)} = a$.

Une variante de ce modèle introduit d'autres caractéristiques véhicule dans le modèle. Cette méthode n'apporte en pratique que très peu d'intérêt, car elle ne permet pas de construire un véhiculier interprétable et contrôlable. De plus, elle nécessite une quantité de données considérable.

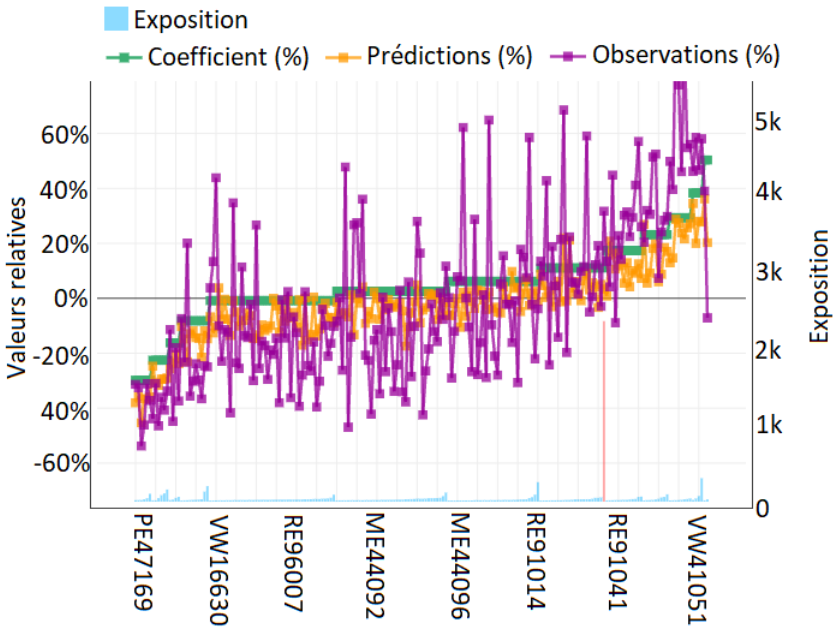


FIGURE 4.1 – Véhiculier traité par crédibilité dans un GLM (modèle de coût moyen)

Les codes SRA sont classés selon leur coefficient, mais il est difficile de justifier le rapprochement d'un code à un autre. De plus, le modèle avec véhiculier donne des performances plus faibles que le modèle de départ, c'est-à-dire sans véhiculier et sans variable véhicule. Ainsi, cette méthode ne sera pas retenue par la suite.

4.1.3 Regroupement des véhicules par l'algorithme CART

En 2012, *SIPULSKYTE R.* propose de combiner le pouvoir de prédiction du GLM avec la clusterisation par arbre et à l'aide de la théorie de la crédibilité pour la robustesse. Une variante de la méthode sera décrite et utilisée ci-après.

L'effet véhicule peut être capté de deux manières différentes. L'effet véhicule correspond dans un premier cas à la prédiction du modèle de fréquence ou de coût moyen lorsque les variables véhicules sont comprises dans le modèle. La prédiction est ensuite divisée par le produit des coefficients non-véhicules. Les effets non-véhicules sont ainsi supprimés. Dans le second cas, l'effet véhicule sont les résidus du modèle dont les variables véhicules ont été intentionnellement omises. Ces deux extractions de l'effet véhicule sont envisageables, cependant le premier cas ne sera pas retenu, car il présente des résultats moins robustes que le second.

Remarque : le travail sur un modèle où les variables explicatives véhicules sont exclues présentes plusieurs limites dont une potentielle instabilité dans la mesure où certaines variables explicatives pourraient être négligées. De plus, les interactions et corrélations entre les variables véhicules et non véhicules peuvent rendre difficile l'interprétation a posteriori du véhiculier. Cela dit, les variables non-véhicules sont très peu corrélées aux variables véhicules et la sélection de variables est effectuée sur un nombre important de variables.

Le principe de la méthode résiduelle est de considérer que les résidus du modèle s'expliquent entièrement par l'effet véhicule. Cette hypothèse est forte. En effet, une part de l'effet non-véhicule et le bruit composent l'effet résiduel. Le résidu se définit ici comme le rapport des valeurs observées sur les valeurs estimées de la variable à expliquer sous l'influence d'un poids, l'exposition.

Les résidus, les observés, les prédictions et l'exposition sont agrégés à la maille code SRA. Ainsi, pour chaque code SRA, on attribue un résidu moyen que l'on peut mettre sous la formule suivante,

$$\hat{r}_k = \frac{\sum_{i \in k} obs_i / expo_i}{\sum_{i \in k} \hat{obs}_i},$$

où $k \in E$ est l'ensemble des véhicules ayant un même code SRA, E est l'ensemble des codes SRA, obs_i est l'observé (fréquence ou charge) du véhicule i ayant pour code SRA k, $expo_i$ est l'exposition du véhicule i et $\hat{obs}_i$ est la prédiction sortie du GLM.

L'algorithme CART permet de segmenter la population en fonction de l'ensemble des variables explicatives en divisant, de manière itérative, les données en deux groupes. Les arbres de régression ont une représentation graphique facile à interpréter grâce à une hiérarchisation binaire. Plus l'arbre se divise, moins l'information résiduelle est significative. La valeur prédite de Y, associée à un groupe, correspond à la moyenne des valeurs observées de Y pour les individus appartenant à cette feuille. En effet, $\hat{f}(x) = \sum_{j=1}^{J} \bar{Y}_j \times \mathbb{1}_{x \in F_j}$, où $\hat{f}$ est l'estimateur de $f(X) = \mathbb{E}[Y|X]$, $\mathbb{1}_{x \in F_j}$ est la fonction indicatrice associée à la feuille F_j et $\bar{Y}_j$ désigne la moyenne empirique dans le groupe j.

La croissance de l'arbre repose sur la notion d'hétérogénéité. Soient N le nœud initial, N_G le nœud gauche engendré par N et N_D le nœud droit engendré par N. Parmi toutes les divisions admissibles du nœud N, l'algorithme retient celle qui maximise,

$$\hat{\Delta} = \sum i \in N(y_i - \bar{y}_N)^2 - (\sum_{I \in N_G} (y_i - \bar{y}_{N_G})^2 + \sum_{I \in N_D} (y_i - \bar{y}_{N_D})^2),$$

où $\bar{y_N}$ est la moyenne empirique de Y pour les observations appartenant au nœud N.

Cette partition classe théoriquement au mieux les données et minimise l'erreur de prédiction. L'hétérogénéité prend son sens, car plus $\hat{\Delta}$ est proche de 0, plus les individus prennent la même valeur. *A contrario*, plus $\hat{\Delta}$ s'éloigne de 0, plus les valeurs de Y sont équiprobables ou très dispersées. Pour une variable réponse continue, cette fonction est la variance intra-nœud. Le processus se termine lorsque tous les nœuds créés sont homogènes, nœuds vides ou arrêté selon un critère.

Afin d'optimiser l'algorithme CART dans la prédiction, le paramètre *cp*, un critère d'arrêt, sera utilisé par la suite afin d'élaguer de manière intelligente l'arbre. L'idée du paramètre est de donner un score à chaque nœud. Dès qu'un nœud a un score plus faible que le nœud précédent alors le dernier nœud est supprimé.

L'arbre CART implique l'application de plusieurs opérations préliminaires :

— les variables avec un nombre non-négligeable de modalités, notamment le modèle ou le nom commercial du véhicule, sont écartées, car l'algorithme CART a tendance à favoriser ces variables, pour plus de souplesse,
— les variables catégorielles ordinales et quantitatives sont numérisées,
— les variables avec un nombre non-négligeable d'informations manquantes sont écartées malgré une gestion par variable substitut [1] intelligente de l'algorithme CART,
— l'apprentissage est réalisé sur la base d'entraînement comme pour un GLM pour tester la robustesse sur les données restantes. Pour rappel, la base d'apprentissage représente 80% du jeu de données,
— chaque feuille représente au minimum 2% de l'exposition.

L'effet véhicule est extrait des résidus à la maille SRA des modèles de coût moyen et de fréquence retenus où les variables véhicules sont omises, hormis l'ancienneté. Ensuite, les arbres sont construits à partir de la fonction *rpart* sur **R** avec la fonction de complexité *cp* comme critère d'arrêt. Cependant, le pouvoir explicatif des arbres est faible, l'erreur relative est proche de 1 quelle que soit la valeur *cp*

1. Une variable discriminante sélectionnée par l'arbre dont une donnée serait manquante pour un individu, est remplacée localement par une variable dite substitut. Cette variable substitut est sélectionnée parmi l'ensemble des variables mises en entrée. C'est la variable qui produit localement les feuilles les plus similaires aux feuilles produites par la variable de départ qui sera choisie. Si pour la variable substitut, la donnée est également manquante, c'est la deuxième variable qui contribuera à la prédiction, et ainsi de suite jusqu'à une limite définie par un critère de qualité.

choisie. Par ailleurs, l'écart-type de l'erreur tracé sur le graphique ci-dessous a une amplitude importante suggérant des écarts relativement conséquents. Cela semble indiquer une importante variabilité de la variable d'intérêt.

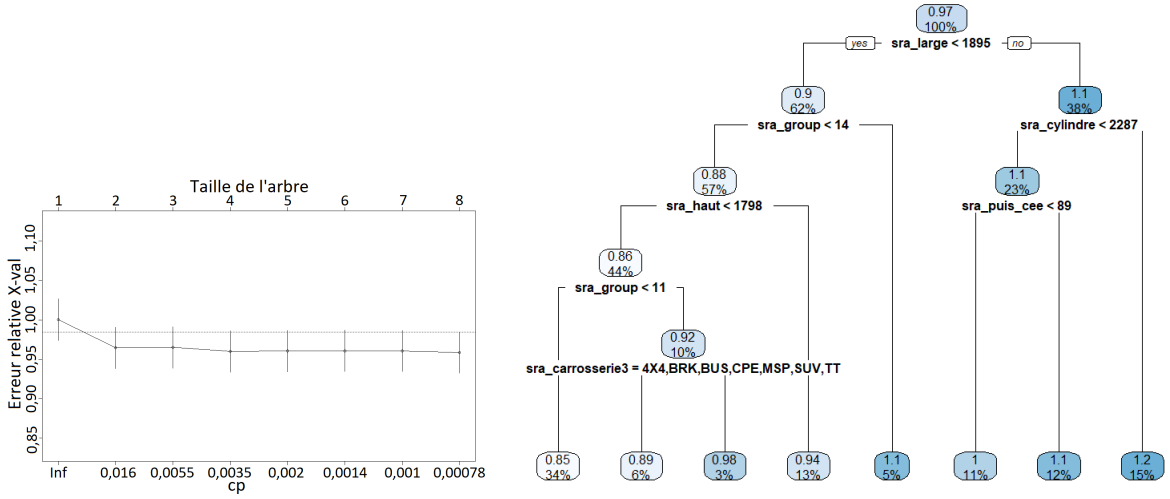


FIGURE 4.2 – Arbre CART sur les résidus du modèle de coût moyen ($cp = 0.0007$)

L'arbre de fréquence renvoie une erreur relative croissante en fonction de la complexité. L'arbre optimal serait donc réduit à une racine. Cependant, un arbre avec 8 racines sera tout de même retenu afin de comparer cette méthode à celle qui suivra. En fait, la forte variabilité de la variable d'intérêt et la sensibilité de la validation croisée, pour un nombre de données faibles, pourrait expliquer la difficulté du modèle à relever un effet véhicule pertinent.

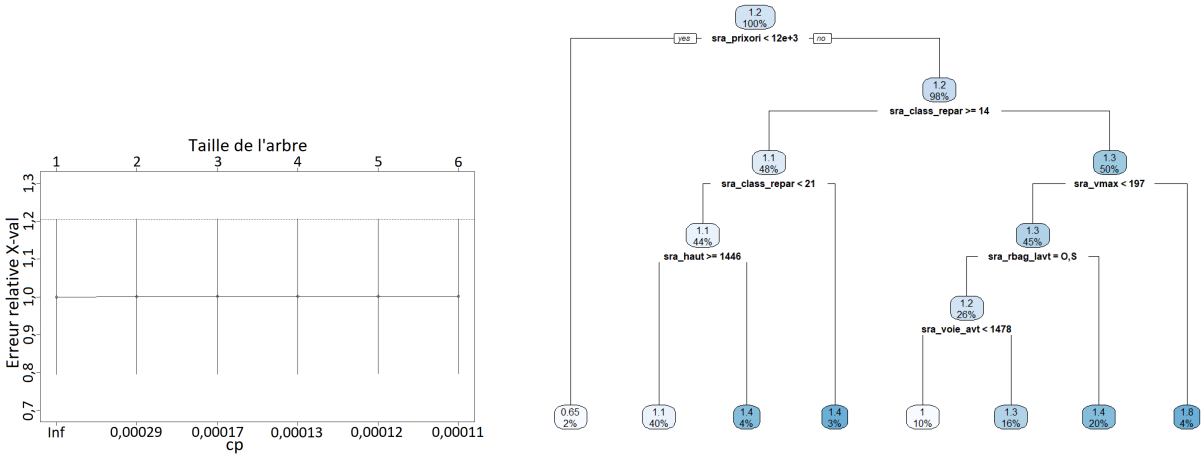


FIGURE 4.3 – Arbre CART sur les résidus du modèle de fréquence ($cp = 0.0005$)

Chaque feuille est constituée de véhicules associés à de nombreuses images ce qui conforte la crédibilité de chaque classe créée.

Afin de vérifier la pertinence de ces résultats, nous proposons d'utiliser la théorie de la crédibilité. Le modèle de Bühlmann-Straub sera plus amplement décrit par la suite. La fonction *cm* sous **R** est appliquée afin de conserver les informations jugées crédibles. Pour chaque modèle, l'information est crédibilisée par année de survenance, et l'effet moyen est partitionnée par le groupe SRA. Un poids est accordé selon l'exposition. Seuls les véhicules ayant un facteur de crédibilité supérieur ou égal à un seuil seront utilisés en entrée de l'arbre CART.

Remarque : les hypothèses prérequises à l'application du modèle de Bühlmann-Straub sont supposées vérifiées.

Le choix du seuil de crédibilité est défini graphiquement à partir de trois courbes :

— le pourcentage de charges ou sinistres,
— le pourcentage de véhicules à la maille SRA,
— le pourcentage d'exposition minimum par véhicule à la maille SRA.

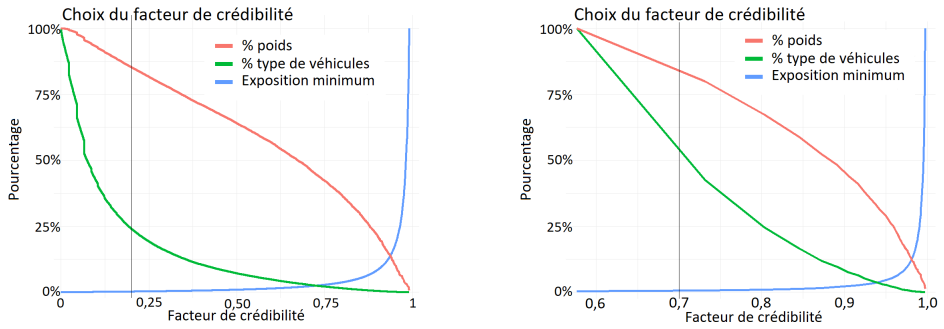


FIGURE 4.4 – Seuil de crédibilité, à gauche celui du modèle de fréquence et à droite celui du modèle de coût moyen

En arbitrant sur ces trois courbes où le but est de supprimer un minimum de charges, de sinistres ou d'exposition, mais un maximum de véhicules à la maille SRA, plusieurs seuils de crédibilité sont fixés. L'exclusion des véhicules les moins crédibles ne modifie pas significativement les règles de l'arbre et des variables similaires sont sélectionnées pour le modèle de coût moyen, contrairement au modèle de fréquence. Cela s'explique par le fait que les arbres construits sans crédibilité sont constitués de nœuds et de feuilles avec un poids élevé. En revanche, pour le modèle de fréquence, l'erreur relative est toujours supérieure à 1, d'où une variation des règles de l'arbre. En augmentant le seuil de crédibilité de 5% à 10% aucun changement n'est constaté. Ainsi, un seuil de 20% et de 65% sera fixé respectivement pour le modèle de fréquence et pour le modèle de coût moyen.

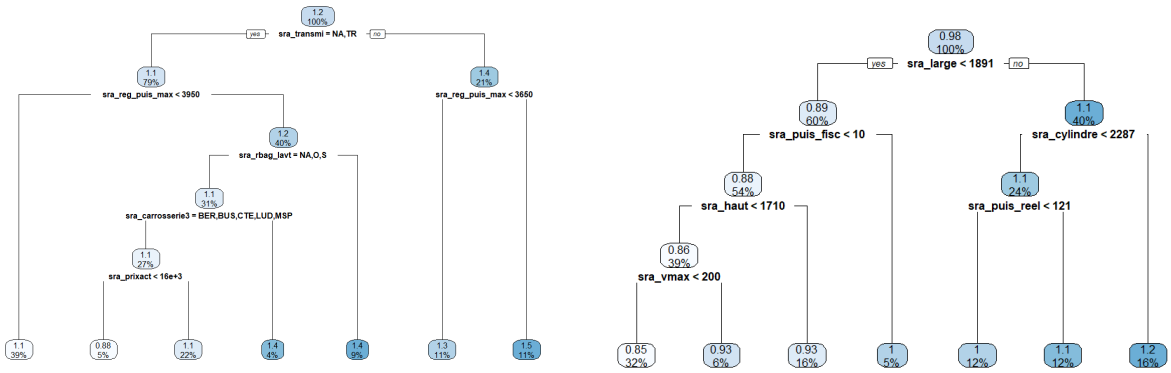


FIGURE 4.5 – Arbres CART du modèle de fréquence à gauche ($cp = 0.0004$) et du modèle du coût moyen à droite ($cp = 0.002$)

Il est difficile à ce stade de juger de la pertinence de ces arbres. Par la suite, ces véhiculiers retenus, ou arbres de régression, seront décrites par une variable. Chaque feuille de l'arbre est retranscrite sous forme d'une modalité d'une variable décrivant la globalité de l'arborescence. Ensuite, pour tester l'efficacité des véhiculiers, une mesure des performances des GLM composés des véhiculiers sera réalisée et comparée au modèle GLM retenu précédemment.

4.2 Analyse Factorielle des Données Mixtes

L'Analyse Factorielle des Données Mixtes, notée AFDM, est une méthode destinée à l'analyse des données contenant à la fois des variables quantitatives et qualitatives. Cet algorithme propose une pro-

jection des individus sur des axes, appelés axes factoriels, décrites par les variables. L'AFDM combine les avantages de l'Analyse en Composantes Principales (ACP) et l'Analyse des Correspondances Multiples (ACM). L'étude sera réalisée à partir des paquets *FactoMineR*, pour l'analyse, et *factoextra*, pour la visualisation des données, sur **R**.

Couramment, l'AFDM ou toutes autres méthodes d'analyse des variables sont produites avant modélisation afin de comprendre les liens ou corrélations entre les différentes variables. Dans le cas de l'étude, le nombre de variables était trop important pour réellement réaliser une analyse qui aie du sens. L'objectif ici, est d'apporter une explication aux liens entre les variables véhicules (Sous-section 3.3.5, page 75) et voire de discerner des effets différents et suffisamment prononcés entre celles-ci. L'utilisation de la base SRA comme appui à la modélisation est une nouveauté côté Auto-entreprises. Les variables véhicule qu'elle propose sont donc encore inconnues. L'AFDM aura par la suite, une seconde utilité, grâce à sa propriété de projection et donc de mappage multi-dimensionnels des corrélations des variables.

4.2.1 Algorithme AFDM

Soient $n = 40\ 000$ le nombre d'individus, $C = 19$ le nombre de variables quantitatives, $D = 25$ le nombre de variables qualitatives et K le nombre total de variables tel que $K = C + D = 44$.

Les variables quantitatives sont traitées comme dans une ACP par centrage et réduction,

$$z_{ik} = \frac{x_{ik} - \mu_k}{\sigma_k},$$

où $i = 1, \ldots, n$ est l'individu, $k = 1, \ldots, C$ est la variable, x_{ik} la valeur prise par l'individu i pour la $k^{ème}$ variable, μ_k la moyenne de la variable, σ_k l'écart-type de la variable et z_{ik} la valeur standardisée.

Parallèlement, les variables qualitatives sont transformées de la même manière que dans un GLM, par un codage disjonctif complet en m_k indicatrices (valeurs dans $\{0, 1\}$). Ainsi, il y a autant de variables formées par une variable catégorielle que de nombre de modalités qui la compose. Enfin, ces variables sont standardisées comme suit,

$$z'_{ik} = \frac{x'_{ik}}{\sqrt{p_k}},$$

où $i = 1, \ldots, n$, $k = C + 1, \ldots, P$, $x'_{ik} = \begin{cases} 1 & \text{si l'individu i prend la modalité k} \\ 0 & \text{sinon} \end{cases}$, p_k la proportion de la modalité dans la variable de départ, z'_{ik} la valeur standardisée et P le nombre total de variables dans la base transformée.

La base de données utilisée en entrée de l'algorithme est donc composée de $P = C + M = 279$ variables, avec M le nombre total d'indicatrices. C'est l'ACP qui est utilisée. En effet, le tableau de données transformées est à présent composé uniquement de variables quantitatives normalisées.

Remarque : la normalisation permet d'équilibrer l'influence des variables dans le calcul des ressemblances et de fournir des priorités intéressantes notamment sur l'inertie. On s'affranchit des unités de mesure.

L'idée générale de l'analyse factorielle est de trouver un facteur F_1 qui explique au mieux la variance résiduelle par les variables. Un second facteur F_2 explique au mieux l'information résiduelle après F_1, et ainsi de suite, jusqu'au $P^{ème}$ facteur.

Concrètement, on cherche à maximiser l'inertie projetée du nuage N_k comportant les P variables. Le vecteur unitaire u qui maximise la variance empirique, ou inertie expliquée, de la projection du nuage sur u est le vecteur propre de c, la matrice de covariance, associé à la valeur propre λ_1,

$$\lambda_1 = \sum_{k=1}^{C} r^2(F_1, X_k) + \sum_{k=C+1}^{P} \eta^2(F_1, X_k),$$

où $0 \leqslant r^2 \leqslant 1$ est le carré du coefficient de corrélation calculé entre le facteur F_1 et les C variables quantitatives et $0 \leqslant \eta^2 \leqslant 1$ est le carré du rapport de corrélation [2] calculé entre le facteur F_1 et les M

2. Le rapport de corrélation se définit comme,

$$\eta^2(y, x) = \frac{\sum_j \sum_{i \in J_j} J(\bar{y}_{.j} - \bar{y})^2}{\sum_j \sum_{i \in J_j} (y_{ij} - \bar{y})^2},$$

où J est le nombre de modalités de la variable x, J_j l'ensemble des individus appartenant à la modalité j, y_{ij} la

variables restantes, utilisé dans l'ACM.

L'inertie de chaque variable vaut 1, de par la normalisation. Les variables qualitatives engendrent un sous-espace de dimension $m_k - 1$, mais par projection sur une direction quelconque l'inertie vaut tout de même 1. De ce fait, la somme des valeurs propres λ_i est égale à K.

4.2.2 Critères de choix

Chaque axe factoriel capte une information contenue dans les variables de départ. En revanche, cette dernière n'est pas toujours judicieuse. Ainsi, nous cherchons à retenir seulement les axes contenant une information pertinente. La pertinence des axes est définie par les trois critères suivants :

— **critère de Kaiser** : ce critère propose de retenir les axes dont l'inertie est supérieure à l'inertie moyenne qui est égale 1. En effet, pour rappel, les variables ont été standardisées.
— **critère de Karlis-Saporta-Spinaki (KSS)** : dans le cadre d'une analyse standardisée, les axes pour lesquels les valeurs propres sont supérieures ou égales au seuil KSS sont considérés pertinentes. Ce seuil se caractérise par,

$$seuil_{KSS} = 1 + 1.65\sqrt{\frac{P-1}{N-1}},$$

où P est le nombre total de valeurs propres théoriquement non nulles et N le nombre d'observations.
— **règle du coude** : ce critère graphique est couramment utilisé, car s'adapte plus facilement au jeu de données. À la détection d'un coude ou rupture de pente, sur l'histogramme des valeurs propres, le nombre de dimensions potentiellement intéressant correspond aux axes d'inertie situés avant le coude (Figure 4.6).

Les critères de Kaiser et de KSS retiennent respectivement 105 et 61 axes factoriels contrairement au critère du coude qui déclare seulement 3 axes comme significativement intéressants.

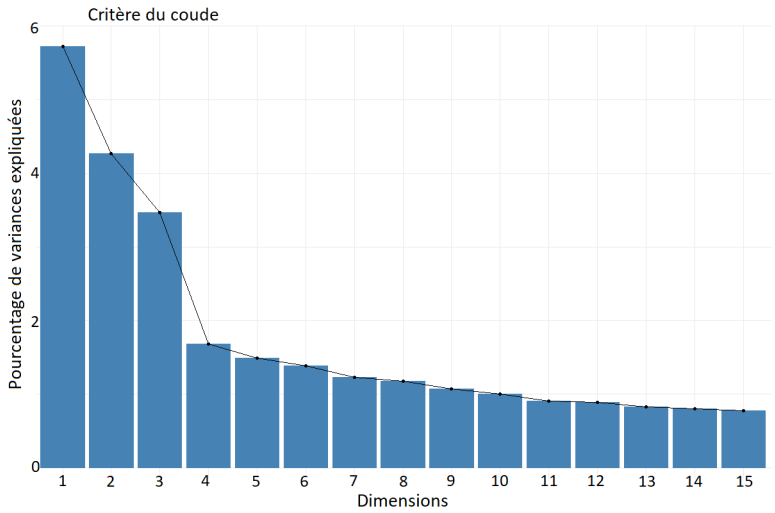


FIGURE 4.6 – Valeurs propres des axes factoriels

Finalement, le critère du coude est retenu. Les 3 premiers axes semblent mieux décrire l'information, car la perte d'inertie est faible à partir du 4ème axe. L'analyse des données est limitée à cause du nombre important de variables. Pour la suite (section 4.3, page 102), retenir 3 axes est pertinent afin de pouvoir représenter graphiquement les résultats et afin de caractériser un voisinage, avec un minimum de sens. Les axes conservées expliquent $13,5\%$ de l'inertie. Ce taux est en revanche biaisé par le nombre important de variables. En effet, *Benzécri* invite à ne s'intéresser qu'à l'information utile en recalculant les valeurs propres sous la forme,

$$\lambda' = [(\frac{P}{P-1}) \times (\lambda - 1)]_+$$

proportion de l'individu i dans la modalité j, $\bar{y}$ est la proportion moyenne de tous les individus et $\bar{y}_{.j}$ est la proportion moyenne des individus de la modalité j.

D'après *Benzécri*, les 3 premières dimensions expliquent 41% de l'inertie corrigée.

4.2.3 Analyse

L'AFDM utilise deux mesures pour rendre compte de l'importance d'une variable sur un axe.

Le cos^2 correspond à la qualité de représentation d'une variable X_j. En effet, en trigonométrie, le cosinus de l'angle correspond au rapport entre le côté adjacent et l'hypoténuse d'un triangle rectangle. Le cosinus est ainsi défini comme la longueur du cercle unité délimité par le point d'intersection entre le côté adjacent et le côté opposé à l'hypoténuse du triangle. La mise au carré permet d'additionner les qualités de représentation des deux axes. La somme des qualités d'une variable sur l'ensemble des axes factoriels est égale à 1. La qualité de représentation d'une variable X_j sur un plan est liée à sa proximité avec le bord du cercle.

La contribution traduit l'importance prise par la variable X_j dans la construction de l'axe (k). Il s'agit de sa part dans l'inertie globale de l'axe,

$$I_k = \sum_{j=1}^{n} m_j \times c_{jk}^2,$$

où c_{jk}^2 est l'éloignement au carré sur l'axe (k) par rapport au centre du repère et $m_j = 1$ la masse.

La contribution est en général exprimée en pourcentage de l'inertie I_k de l'axe (k) sous la forme,

$$CTR_k(X_j) = \frac{m_j \times c_{jk}^2}{I_k}$$

Cette mesure ainsi que la précédente sont reproductibles sur les individus. Une trop forte contribution d'un individu a pour conséquence de modifier l'axe et pourrait nous pousser à supprimer de l'individu.

Remarque : dans l'étude, l'influence d'un individu est fortement diluée par la contribution des autres. En effet, chaque individu contribue sous le poids $m_i = \frac{1}{n}$ où n est le nombre d'individus.

Les variables déterminantes dans l'AFDM sont usuellement sélectionnées selon deux règles :

— la contribution doit être supérieure à $1/K$ où K est le nombre total de variables étudiées
— le cosinus carré doit être supérieur à $1/P$ où P est l'inertie totale.

Le tableau ci-dessous (Figure 4.7), résume l'ensemble des informations citées précédemment, et donne une idée du nombre de données manquantes par variable. L'AFDM ne fonctionne pas en présence de données manquantes. Il convient de retirer les variables SRA composées de plus de 30% de données manquantes. La suppression de l'ensemble des informations non renseignées sur les variables restantes engendre une perte de 26% des véhicules à la maille SRA.

	Dim. 1			Dim. 2			Dim. 3			Données
Variables	coord	cos2	contrib	coord	cos2	contrib	coord	cos2	contrib	manquantes
sra_abs	0.07	2%	2%	0.00	0%	0%	0.20	5%	5%	5%
sra_antivol	0.06	1%	2%	0.00	0%	1%	0.21	2%	5%	5%
sra_ass_frein_urg	0.08	3%	2%	0.01	0%	1%	0.34	11%	6%	12%
sra_boit_vit	0.31	2%	3%	0.06	0%	2%	0.04	0%	2%	0%
sra_carrosserie3	0.05	1%	1%	0.45	5%	6%	0.04	0%	2%	0%
sra_class_prix	0.72	3%	5%	0.12	0%	3%	0.05	0%	2%	0.024%
sra_class_repar	0.22	1%	3%	0.02	0%	1%	0.11	0%	3%	0%
sra_cout_actu_optique	0.14	14%	2%	0.04	4%	2%	0.00	0%	0%	12%
sra_cout_actu_pbrise	0.16	16%	2%	0.00	0%	0%	0.00	0%	0%	12%
sra_cout_actu_piece	0.31	31%	3%	0.01	1%	1%	0.00	0%	0%	12%
sra_ctrl_dyn_stab	0.12	4%	2%	0.03	1%	1%	0.33	11%	6%	12%
sra_cylindre	0.19	19%	3%	0.00	0%	0%	0.09	9%	3%	1%
sra_dir_ass	0.04	1%	1%	0.00	0%	0%	0.08	3%	3%	1%
sra_dispo_cyl	0.18	4%	3%	0.02	0%	1%	0.42	8%	7%	12%
sra_empat	0.05	5%	1%	0.25	25%	4%	0.00	0%	1%	12%
sra_energie	0.01	0%	1%	0.05	1%	2%	0.00	0%	0%	0%
sra_frein	0.18	6%	3%	0.00	0%	0%	0.01	0%	1%	0%
sra_group	0.53	2%	5%	0.14	1%	3%	0.03	0%	2%	0.024%
sra_haut	0.00	0%	0%	0.38	38%	5%	0.00	0%	0%	18%
sra_large	0.17	17%	3%	0.14	14%	3%	0.00	0%	1%	12%
sra_long	0.16	16%	3%	0.13	13%	3%	0.00	0%	1%	12%
sra_marque2	0.08	0%	2%	0.11	0%	3%	0.03	0%	2%	0%
sra_nb_cyl	0.08	1%	2%	0.05	1%	2%	0.04	0%	2%	1%
sra_nbplace	0.01	0%	1%	0.45	5%	6%	0.01	0%	1%	0%
sra_poids_vide	0.33	33%	4%	0.03	3%	2%	0.01	1%	1%	1%
sra_pos_moteur	0.00	0%	0%	0.00	0%	0%	0.00	0%	0%	0%
sra_prixact	0.47	47%	4%	0.01	1%	1%	0.01	1%	1%	0%
sra_prixori	0.48	48%	4%	0.01	1%	1%	0.01	1%	1%	0%
sra_ptac	0.09	9%	2%	0.22	22%	4%	0.02	2%	1%	1%
sra_puis_cee	0.43	43%	4%	0.03	3%	1%	0.00	0%	1%	0%
sra_puis_fisc	0.17	17%	3%	0.02	2%	1%	0.05	5%	2%	0%
sra_puis_reel	0.43	43%	4%	0.03	3%	1%	0.00	0%	1%	0%
sra_rbag_cond	0.07	2%	2%	0.00	0%	0%	0.34	11%	6%	12%
sra_rbag_larr	0.08	3%	2%	0.15	5%	3%	0.36	12%	6%	13%
sra_rbag_lavt	0.07	2%	2%	0.33	11%	5%	0.40	13%	7%	12%
sra_rbag_pavt	0.07	2%	2%	0.39	13%	5%	0.40	13%	7%	12%
sra_reg_puis_max	0.00	0%	0%	0.13	13%	3%	0.00	0%	0%	12%
sra_segment	0.23	3%	3%	0.54	8%	6%	0.18	3%	4%	3%
sra_suspen	0.00	0%	0%	0.16	4%	3%	0.04	1%	2%	0%
sra_transmi	0.05	1%	1%	0.01	0%	1%	0.03	1%	2%	0%
sra_typ_alim	0.10	1%	2%	0.02	0%	1%	0.06	0%	2%	0%
sra_vmax	0.09	9%	2%	0.24	24%	4%	0.00	0%	0%	0%
sra_voie_arr	0.17	17%	3%	0.12	12%	3%	0.00	0%	1%	14%
sra_voie_avt	0.15	15%	2%	0.14	14%	3%	0.00	0%	1%	14%

FIGURE 4.7 – Tableau des coordonnées, contributions et cosinus carré des variables

Un total d'environ 40 000 codes SRA est mis en entrée de l'AFDM. Un traitement des données manquantes a été réalisé et a conduit à des résultats probants.

Audigier V., *Husson F.* et *Josse J.* en 2013 proposent deux algorithmes dont l'algorithme itératif AFDM régularisé à partir de la fonction *imputFAMD*. Cet algorithme s'initialise en imputant les valeurs manquantes avec des valeurs telles que la moyenne de la variable pour les variables continues, et la proportion de la catégorie pour chaque variable indicatrice normée en utilisant les entrées non-manquantes. La deuxième étape de l'algorithme itératif AFMD régularisé consiste à exécuter *FAMD* sur l'ensemble des données complétées. Ensuite, il impute les valeurs manquantes avec les formules de reconstruction régularisées d'ordre *ncp* [3] (la matrice ajustée avec les composantes *ncp* pour les scores et les chargements régularisés). Ces étapes d'estimation des paramètres via AFDM et d'imputation des valeurs manquantes à l'aide de la matrice ajustée et régularisée sont itérées jusqu'à convergence.

Malheureusement, la méthode n'a pas été concluante. En effet, au niveau du critère de liens aberrants (sous-section 4.3.1, page 103), les véhicules à la maille SRA composés de données manquantes et complétés par l'algorithme, sont proches de certains véhicules dont la classe de prix ou le groupe SRA est très éloigné.

Une seconde approche transforme les données manquantes des variables quantitatives en la moyenne de ces variables par classe de prix. La classe de prix est une variable qui contribue fortement aux axes lorsque l'on supprime les 26% des données manquantes. Le groupe SRA est également une variable intéressante. Cependant, à propos du critère de liens aberrants décrit par la suite, regrouper par classe de prix donne des résultats légèrement moins variants. Cette approche sera retenue pour traiter les données manquantes, car elle fournit des résultats convenables au niveau des corrélations des variables et de la proximité des points.

Les plans factoriels des différents axes peuvent aider à l'interprétation des variables. Les coordonnées sur un axe sont définies par le coefficient de corrélation r pour les variables quantitatives et le rapport de corrélation η pour les variables qualitatives. De plus, les cercles de corrélation concentrent l'analyse sur les seules variables quantitatives afin d'observer leur qualité de représentation et leurs contributions aux axes.

3. Entier correspondant au nombre de composants utilisés pour prédire les entrées manquantes

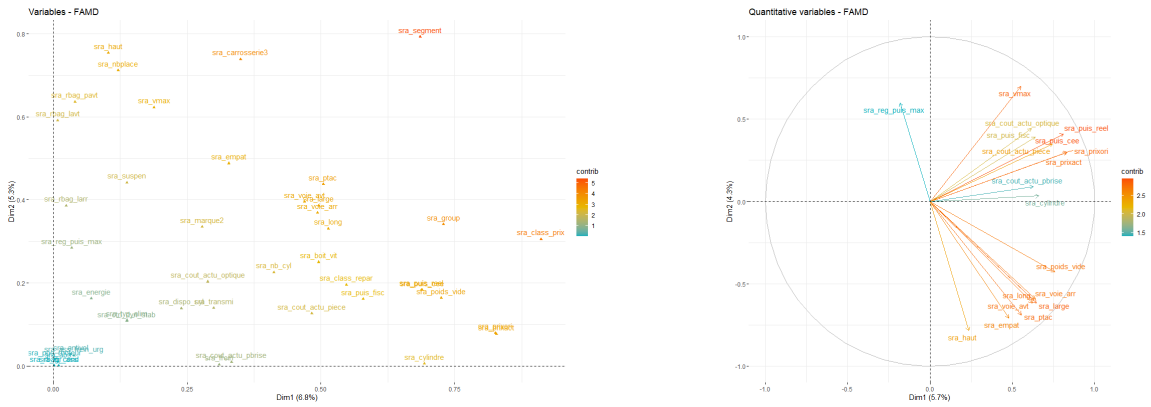


FIGURE 4.8 – Représentation graphique des variables sur les axes factoriels

Sur le cercle de corrélation, des groupes de variables partagent un sens et une direction commune. En effet, leur corrélation est forte, pour citer les variables de dimension du véhicule et de poids ou les variables de prix et de puissance du véhicule. La qualité de la représentation, ou cos^2, confirme la relation entre ces variables. Afin d'apporter une explication aux axes :

— l'axe 1 est un axe fortement porté par le prix du véhicule,
— l'axe 2 est une dimension influencée par la carrosserie du véhicule et sa dimension (seule la dimension semble posséder une bonne qualité de représentation),
— l'axe 3 représente la sécurité du véhicule.

Maintenant, les individus peuvent être représentés sur un graphique à trois dimensions correspondant au nombre d'axes retenus. L'espace ainsi créé correspond à la carte des individus au même titre qu'un zonier avec un mappage sur la longitude et la latitude.

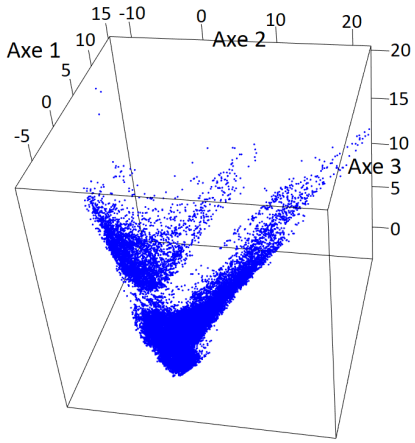


FIGURE 4.9 – Graphique 3D de l'AFDM

4.3 Lissage des résidus par la carte des voisins

Les individus ont été précédemment cartographiés sur un axe à trois dimensions. Les voisins de chaque profil SRA seront alors définis dans l'espace. D'ailleurs, un lien peut être fait avec l'information géographique qui a été classifiée à l'aide d'une projection sur un plan. Par conséquent, une carte des voisins sera construite, communément appelée table d'adjacence.

L'objectif visé serait de crédibiliser puis de lisser l'effet résiduel véhicule de chaque individu par les informations de leur plus proche voisin. Ce modèle peut être assimilé à la distance de Moran dans un zonier.

4.3.1 Définition de la carte des voisins

Pour définir un voisin, Delaunay propose une triangulation de l'espace euclidien.

Soit $P = p_1, \ldots, p_n$ un ensemble de points de l'espace euclidien à deux dimensions. La triangulation consiste en la subdivision de cet espace en triangles en reliant trois points. Il s'agit d'un réseau inter-connecté dans lequel aucun lien entre deux points ne peut croiser un autre. Pour un même espace, plusieurs méthodes de triangulation sont possibles.

La triangulation de Delaunay propose une unique triangulation dont l'idée générale n'est de créer que les triangles, ou liens entre trois points, dont le cercle circonscrit à ce triangle ne contient aucun autre point. Cela permet de valider que les points du triangle sont des voisins au sens des plus proches points dans l'espace, le vide les séparant. Cette triangulation est le diagramme dual des polygones de Voronoï.

Soit n le nombre d'observations dans un espace normé, on construit un pavage de l'espace où chaque polygone élémentaire, nommé polygone de Voronoï, ne renferme qu'une observation et que tous points intérieurs à ce polygone sont plus près de son centre que des autres. La construction de ces polygones se traduit par le tracé des médiatrices de chaque segment.

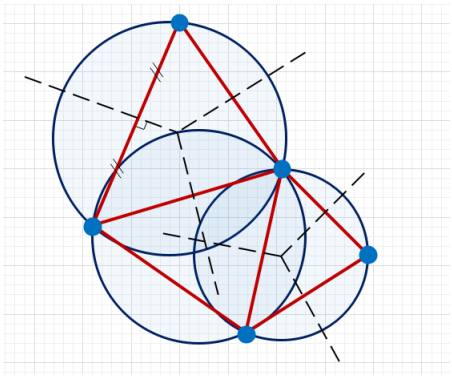


FIGURE 4.10 – Triangulation de Delaunay

Remarque : le centre du cercle circonscrit d'un triangle est le point de concordance des médiatrices d'un triangle.

Cette méthode présente également d'autres avantages. En effet, on essaie de maximiser les angles pour chaque triangle et donc de se rapprocher le plus possible de triangles équilatéraux supprimant les longues arrêtes. Son implémentation sur **R** est rapide avec la fonction delaunay. La triangulation de Delaunay forme des tétraèdres et peut se généraliser à toute dimension supérieure, mais ne peut pas être représentée.

La triangulation de Delaunay crée un réseau fermé ne laissant aucun point isolé. Chaque lien correspond à un voisinage. En revanche, graphiquement, certains points sont éloignés et pourtant sont considérés comme voisins. La notion de triangulation ne contient pas la notion de distance, mais de vide. Ainsi, afin de construire une carte des voisins cohérente, l'idée serait de supprimer certains liens selon les deux critères suivants :

— critère de distance : l'idée est de couper les liens entre deux véhicules à la maille SRA où la distance euclidienne est supérieure à une valeur seuil. La distance euclidienne entre deux points A et B dans l'espace à 3 dimensions se définit comme,

$$d(A, B) = \sqrt{(X_A - X_B)^2 + (Y_A - Y_B)^2 + (Z_A - Z_B)^2}.$$

Le seuil a été fixé à $0,87$ et 5% de liens ont été supprimés. Ce seuil a été choisi graphiquement en traçant 100 quantiles sur les 10% plus grandes distances euclidiennes,

— critère de liens aberrants : ce critère s'ajoute au précédent et est lié à un sens métier. En effet, la classe de prix et le groupe SRA sont des variables qui contribuent beaucoup aux différentes dimensions. Le groupe SRA résume une grande partie de l'information captée par les autres variables SRA. Ainsi, cela n'a pas de sens de considérer deux véhicules à la maille SRA comme voisins alors qu'ils ont plus de 4 classes de différences. Ce chiffre 4 a été fixé à partir d'un tableau croisé des modalités où pour chaque classe, le nombre de voisins est recensé. Au-delà de 4 classes

de différences, le nombre de liens est faible, preuve que la triangulation, l'AFDM et le critère de distance représentent bien le voisinage.

	D	E	F	G	H	I	J
D	688						
E	1356	1284					
F	643	2021	2053				
G	302	1068	3083	2632			
H	288	581	1635	4006	4438		
I	137	341	831	1871	6158	5904	
J	62	106	285	731	2187	6595	7090
K	15	38	78	185	629	2195	8346
L	5	3	10	37	131	617	2581
M	2	2	7	6	10	92	666
N				1	1	40	151
O						9	21
P					3	1	4
Q							

FIGURE 4.11 – Tableau croisé des modalités de la classe de prix SRA

On accepte une marge d'erreur de 4 classes, pour permettre aux autres variables de contribuer à l'explication du voisinage. Le critère de liens aberrants engendre une suppression d'environ $2,65\%$ des voisins (dont $1,9\%$ liés à la classe de prix) et donc au total de $7,65\%$ de liens sont coupés.

L'AFDM est tracée avec la triangulation corrigée par les critères précédents et la table d'adjacence est créée, appelée carte des voisins, formalisant les véhicules voisins. Pour chaque code SRA, les codes SRA voisins ainsi que leur nombre sont listés. Avant l'application des critères de voisinage, un type de véhicule caractérisé par un code SRA avait en moyenne 8 voisins et un maximum de 34. Après, un code SRA est associé en moyenne à 7 autres codes et à un maximum de 31.

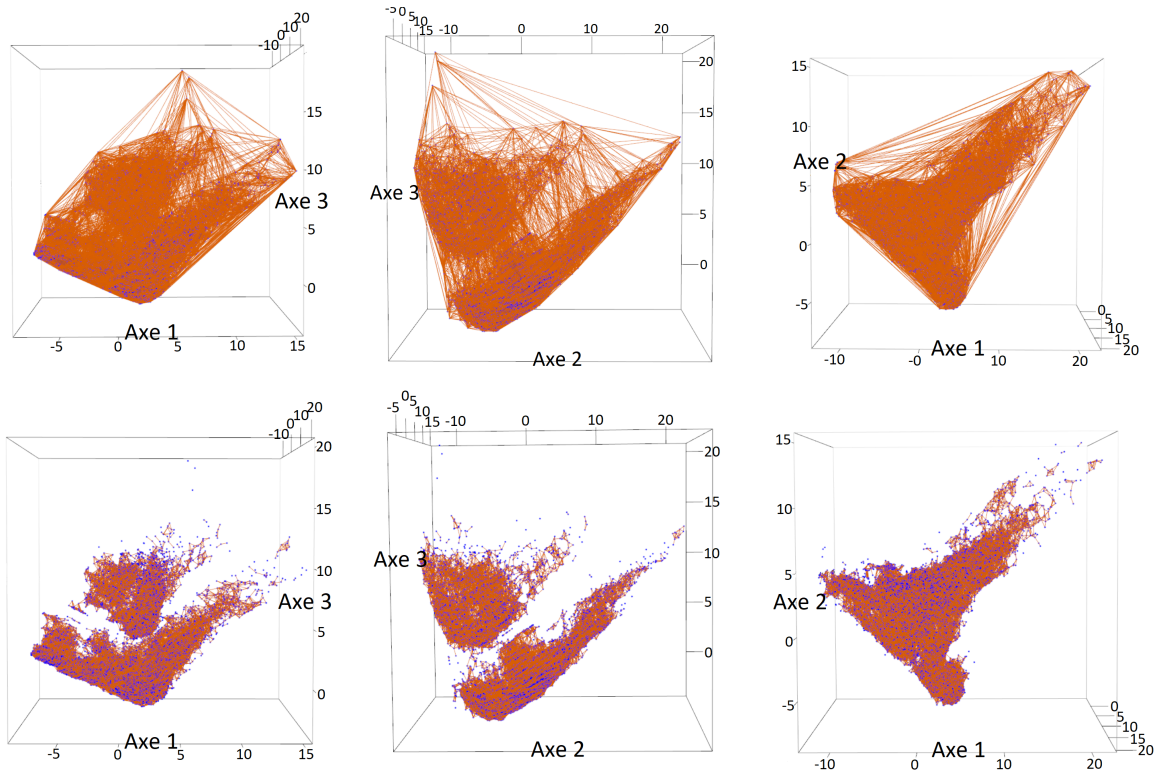


FIGURE 4.12 – Triangulation Delaunay sur l'AFDM avant et après application des critères de voisinage

4.3.2 Lissage par crédibilité

Les résidus à la maille SRA des modèles GLM captent l'effet véhicule, une partie de l'effet non-véhicule qui n'a pas été décrite par les modèles, et un bruit. À cause du nombre important de codes différents, l'exposition d'un code donnée est faible, il en est de même de la sinistralité. L'objectif est de retirer au maximum la part d'effet non-véhicule, le bruit présent et que chaque code ait une pertinence dans la

description du risque.

Analogue aux méthodes développées pour le zonier, des techniques de lissage spatiales sont couramment utilisées pour lisser les risques et les rendre plus fiables. La théorie de la crédibilité permet de lisser les résidus en crédibilisant les données par d'autres informations. Après lissage spatial des résidus, une classification des variables véhicules pourra être constituée à partir de l'algorithme CART, vu précédemment. En fait, par rapport au premier véhiculier produit, l'étape de lissage *a priori* permet de rendre les arbres plus pertinents.

Chaque code SRA sera crédibilisé à partir des codes voisins afin d'améliorer l'estimation du risque. En effet, la carte des voisins lie de manière assez prudente chaque code SRA aux codes adjacents corrélés par les informations véhicules. On suppose que toute chose égale par ailleurs, les véhicules présentant des caractéristiques semblables présentent une même probabilité de risque.

Le modèle de lissage spatial dans le cadre de cette étude est tiré du modèle de Bühlmann-Straub. Il existe diverses méthodes possibles. Les voisins dont la distance est proche ou dont l'exposition [4] est élevée, auront un poids plus élevé pour crédibiliser l'information.

Définition

Le modèle de Bühlmann-Straub est une généralisation du modèle de Bühlmann intégrant une pondération ou poids w_{it} à chaque observation notée X_{it}.

(BS1) Soient $(\theta_i, X_i)_{i=1,\dots,I}$ des couples indépendants où les variables aléatoires $\theta_1, \dots, \theta_I$ sont identiquement distribuées et les variables aléatoires X_{it} ont une variance finie.
(BS2) Les variables aléatoires X_{it} sont telles que,

$$\mathbb{E}[X_{it}|\theta_i] = \mu(\theta_i),\ i = 1, \dots, I$$

$$Cov(X_{it}, X_{iu}|\theta_i) = \delta_{tu}\frac{\sigma^2(\theta_i)}{w_{it}},\ t,\ u = 1, \dots, n$$

Les variables X_{it} sont des ratios. En général, $X_{it} = \frac{S_{it}}{w_{it}}$. La meilleure approximation linéaire non-homogène de la prime de risque $\mu(\theta_i)$, ou de $X_{i,n+1}$ est,

$$\pi^{BS}_{i,n+1} = z_i X_{iw} + (1 - z_i)m,$$

où $z_i = \frac{w_{i\Sigma}}{w_{i\Sigma}+K}$, $K = \frac{s^2}{a}$ et m est le risque collectif.

Le meilleur estimateur, avec la plus faible variance, de l'estimateur collectif $m = \mathbb{E}[\mathbb{E}(X_{kj}|\theta_k)] = \mathbb{E}(X_{kj})$ est,

$$\hat{m} = X_{zw} = \sum_{i=1}^{I} \frac{Z_i}{Z_\Sigma} X_{iw}.$$

Par analogie avec l'estimateur du modèle de Bühlmann, un estimateur sans biais de $s^2 = \mathbb{E}[\sigma^2(\theta_i)] = \mathbb{E}[w_{it}\mathbb{V}[X_{it}|\theta_i]]$ la variance intra mesure des fluctuations des observations,

$$s^2 = \frac{1}{\sum_{i=1}^{I}(n_i - 1)} \sum_{i=1}^{I}\sum_{t=1}^{n_i} w_{it}(X_{it} - X_{iw})^2$$

L'estimateur intuitif rendu sans biais de la variance inter $a = \mathbb{V}[\mu(\theta_k)] = \mathbb{V}[\mathbb{E}(X_{kj}|\theta_k)]$, mesure de l'hétérogénéité au sein du groupe, est,

$$\hat{a} = \frac{w_{\Sigma\Sigma}}{w^2_{\Sigma\Sigma} - \sum_{i=1}^{I} w^2_{i\Sigma}}\left(\sum_{i=1}^{I} w_{i\Sigma}(X_{iw} - X_{ww})^2 - (I-1)\hat{s}^2\right)$$

4. L'exposition est la durée d'exposition pour le modèle de fréquence et le nombre de sinistres pour le modèle de coût moyen.

Remarque : $\hat{a}$ peut être négatif. Un pseudo-estimateur de Bichsel-Straub sans biais et toujours positif, $\tilde{a} = \frac{1}{I-1}\sum_{i=1}^{I} z_i(X_{iw} - X_{zw})^2$ est un résultat de calcul de point fixe. Il est conseillé d'utiliser ce second estimateur.

Application

L'application de la théorie de la crédibilité sur les données prédit les résidus lissés R_i avec $i = 1, \ldots, I$ le type de véhicule caractérisé par le i^{eme} code SRA,

$$R_i = Z_i r_i + (1 - Z_i)\bar{r}_i,$$

où r_i est le risque individuel du i^{eme} code SRA, $\bar{r}_i$ est le risque collectif et Z_i est le facteur de crédibilité défini comme $Z_i = \frac{w_i}{w_i + w_0}$, w_i est l'exposition du i^{eme} code SRA et w_0 est l'analogie du rapport entre la variance intra et inter explicitée précédemment.

Au risque collectif $\bar{r}_i$, un poids supplémentaire est ajouté aux types de véhicules qui ont les caractéristiques les plus proches, c'est-à-dire dont la distance euclidienne dans l'espace est la plus faible,

$$\bar{r}_i = \frac{\sum_{k \neq i}^{n_i} r_k d_{ik}^{-P} w_k}{\sum_{k \neq i}^{n_i} d_{ik}^{-P} w_k},$$

où n_i est le nombre de voisins pour le véhicule caractérisé par le $i^{ème}$ code SRA, d_{ik} est la distance euclidienne entre le principal concerné et son voisin et P est un paramètre.

P est un paramètre dont la valeur reflète l'importance accordée à la distance et donc à la proximité des voisins. Plus P est grand, moins le lissage est important, car un cercle plus étroit des plus proches voisins est considéré.

En raison du faible nombre de voisins par code SRA, P sera fixé à 1, ainsi autant d'importance est accordée aux voisins éloignés et aux voisins proches. Le lissage sera tiré à son maximum.

Ce modèle sera réalisé sous **R** en reconsidérant toutes les étapes de calcul. L'une des principales limites du modèle est que le partitionnement des véhicules favorise l'homogénéité des classes vis-à-vis du nombre de véhicules par groupe, aux dépens de l'homogénéité des variables véhicules au sein de chaque groupe. Une seconde limite est que l'information n'est crédibilisée que sur un nombre faible de voisins, d'ailleurs, les points isolés ne sont pas groupés. Ces deux limites mènent à des résultats corrects, mais dont le lissage n'est pas autant visible que le zonier. Le paramètre P ne peut pas être exploité comme paramètre de lissage.

Une ouverture à ce mémoire serait de rechercher des méthodes encore plus pénalisantes. Une idée serait de lisser le risque associé à un type de véhicule à partir de la moyenne collective de tous les autres, et de ne plus se limiter aux voisins. Ensuite, le poids sur la distance prendrait un sens plus important. Un voisin serait caractérisé par un poids élevé, car il est proche dans l'espace par rapport aux autres points. Pour pousser l'étude encore plus loin, il serait peut-être intéressant de ne pas faire la moyenne sur tous les autres codes, mais sur des cantons, c'est-à-dire ajouter des critères comme ceux définis pour le lissage *"Adjency-based spatial smoothing"* et ainsi ne plus considérer les points trop éloignés en distance et/ou en caractéristique véhicule.

En ajoutant une quatrième dimension au graphique 3D, le lissage des résidus peut être visualisé. Plus les résidus appartiennent aux déciles les plus bas, plus la couleur sur le graphique se rapproche du blanc ou du bleu clair, et à l'inverse, plus la couleur se rapproche du bleu foncé, plus le résidu est élevé par rapport au reste.

Aucune tendance sur le lissage ne peut être relevée graphiquement. Le lissage n'est donc pas suffisamment appuyé. Cependant, cette problématique pourrait faire l'effet d'un autre mémoire. Ici, l'objectif est

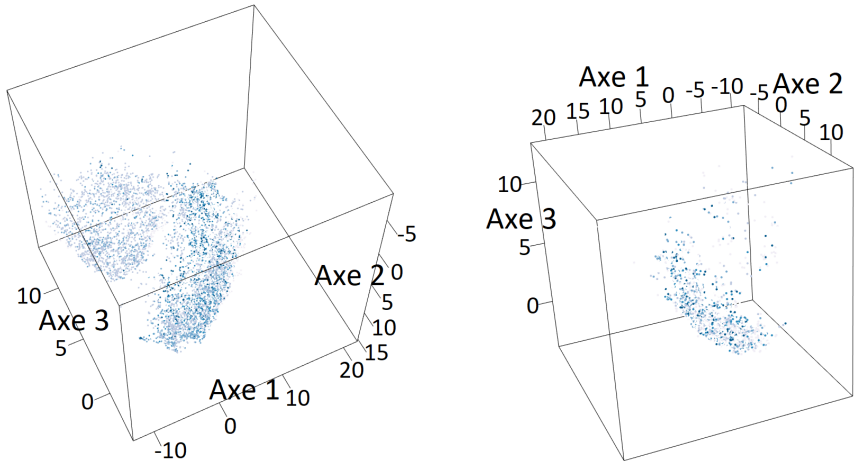


FIGURE 4.13 – Carte des voisins après lissage, à gauche résidus du modèle de fréquence et à droite résidus du modèle de coût moyen

d'explorer plusieurs méthodes et de choisir la méthode qui fonctionne le mieux au sens des critères de performance afin de construire le véhiculier, ou un classement de risque des variables véhicules.

Pour classer les variables véhicules à partir des résidus lissés, l'algorithme CART sera de nouveau exploité. Chaque feuille représente au minimum 2% de l'exposition et le paramètre de complexité, *cp*, est optimisé pour le modèle de coût moyen. Pour le modèle de fréquence, l'erreur résiduelle est toujours croissante en fonction du paramètre *cp*. D'ailleurs, la variance inter estimée $\hat{a}$ est négative, les résidus lissés sont entièrement portés par le risque collectif.

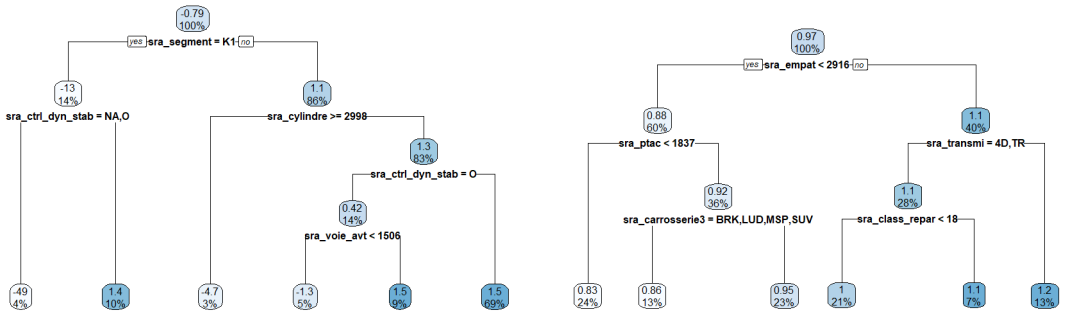


FIGURE 4.14 – Arbres CART, à gauche sur le modèle de fréquence et à droite sur le modèle de coût moyen

4.4 Performance des véhiculiers

Chaque véhiculier est représenté par un arbre. Une variable est conçue à partir de cet arbre en définissant des modalités pour chaque feuille formée. Ainsi, les 3 véhiculiers forment 3 variables. Les modèles GLM suivants seront conservés :

— modèles GLM avec les variables retenues, que l'on notera **glm_retenu**,
— modèles GLM avec les variables retenues dont les variables véhicule sont substituées par le véhiculier définit par un arbre. Ces modèles se nommeront **glm_vehiculier**,
— modèles GLM avec les variables retenues dont les variables véhicule sont corrigées par le véhiculier construit à partir d'un arbre. La variable à expliquer est l'effet véhicule soustrait de l'effet non crédible. Ces modèles seront appelés par la suite **glm_vehiculier_cred**,
— modèles GLM avec les variables retenues dont les variables véhicules sont remplacées par le véhiculier construit à partir de l'effet véhicule lissé par crédibilité et carte des voisins, notées **glm_vehiculier_lisse**.

Il semblerait que les meilleurs modèles du coût moyen et de la fréquence au sens des critères de performances (Gini, MSE, MAE, ...), excepté la courbe lift, sont ceux formés par le véhiculier *glm_vehiculier_cred*. Les performances entre les bases de validation et les bases d'apprentissage sont plus proches l'une de l'autre qu'avec le modèle *glm_retenu*. Les modèles captent plus de signaux et moins de bruit. Pour la courbe lift, graphiquement il semblerait que les modèles *glm_retenu* présentent des courbes de prédiction et d'observé moyen, plus proches.

	K-fold entrainement	K-fold test	
Gini normalisé	41,09%	39,90%	
	41,23%	39,41%	F1
	41,44%	38,95%	F2
	40,82%	40,49%	F3
	40,86%	40,75%	F4

	K-fold entrainement	K-fold test	
Gini normalisé	41,03%	40,32%	
	41,25%	39,65%	F1
	41,47%	38,87%	F2
	40,65%	41,52%	F3
	40,77%	41,25%	F4

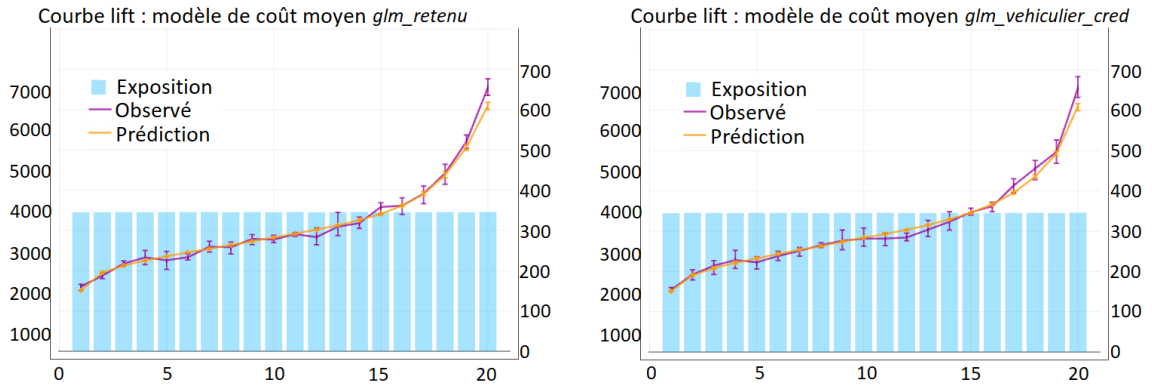


FIGURE 4.15 – Comparaison des modèles de coût moyen, à gauche *glm_retenu* et à droite *glm_vehiculier_cred*

Le modèle *glm_retenu* sera préféré aux modèles avec véhiculier pour plusieurs raisons.

Le véhiculier appliqué dans *glm_vehiculier_lisse* a plusieurs limites. Le lissage par la carte des voisins est faible à cause d'un nombre faible de voisins définis par code SRA et les véhicules isolés ne sont pas crédibilisés. Ces modèles restent toutefois plus performants que *glm_vehiculier* où les résidus sont préservés. Il est difficile de décrire l'effet véhicule qui semble masqué par le bruit et l'effet non-véhicule restant.

L'avantage du véhiculier supprimant l'effet véhicule considéré non crédible est le classement de cet effet sur une minorité de codes SRA, mais qui représentent une grande majorité de véhicules. Cependant, une partie de l'information est mise de côté, la performance gagnée est faible, la définition du seuil de crédibilité est difficile à déterminer et le véhiculier du modèle de fréquence retenu n'est pas le modèle optimisé au sens de l'erreur relative où seulement une feuille aurait été considérée.

Finalement, dans le cadre de l'étude, les véhiculiers ne semblent pas suffisamment pertinents pour décrire l'effet véhicule dans le modèle GLM. Cependant, ils permettent de tracer un cadre clair du risque sur les variables véhicules. Les véhiculiers seront conservés en tant qu'analyse des données des variables véhicules, de leur niveau de significativité et de leur importance. Des pistes plus poussées sur la création du véhiculier sont à envisager. Un lissage par exemple plus développé, permettrait de mieux capter l'effet véhicule.

Chapitre 5

Méthodes de machine learning

Le risque RC automobile sur les véhicules à 4 roues de moins de $3,5$ tonnes a été segmenté à l'aide d'un modèle type GLM. Cependant, ce type de modèle n'est peut-être pas le plus adapté. De plus, les nombreux retraitements de la charge et les nombreuses méthodes de segmentation du risque rendent peu claires les prédictions du modèle.

Finalement, d'autres modèles d'apprentissage seront étudiés. Ensuite, l'objectif sera de comparer le tarif actuel au tarif sans contrainte informatique afin de connaître le gain potentiel d'un tel modèle.

5.1 Comparaison à des méthodes de machine learning

Le modèle utilisé précédemment est le modèle de régression linéaire généralisé. Cependant, d'autres modèles de machine learning existent. Il s'agit donc de savoir si l'ajustement GLM prédit aussi bien que les autres modèles. Deux machines seront étudiées :

— Random Forest
— XGBoost

Ces deux modèles sont plus performants que des arbres de décision de construction complexe demandant une grande puissance de calcul.

5.1.1 Random Forest

Les forêts aléatoires sont construites sur les arbres de décision. Les arbres de décision sont faciles à construire, à utiliser ou à interpréter. Les arbres de décision sont fortement limités par leur propre construction. Les forêts aléatoires combinent simplicité des arbres de décision et flexibilité et diminuent fortement l'imprécision.

Une forêt aléatoire est construite en créant des échantillons *bootstrapped* du jeu de données, c'est-à-dire des bases de données comprenant des individus tirés au hasard dans l'ensemble de données d'origine avec remise. Un arbre de décision est ensuite construit sur le jeu de données *bootstrapped*, en utilisant un nombre limité de variables à chaque étape pour apprendre à diviser une branche. Puis, un second arbre de décision est construit sur un autre jeu de données *bootstrapped*, ainsi de suite. En utilisant un échantillon *bootstrap* et en ne considérant qu'un sous-ensemble des variables à chaque étape, la forêt est composée d'une grande variété d'arbres. C'est la moyenne des résultats de chaque arbre qui correspond à la prédiction.

À l'aide du paquet *sklearn* sur **Python**, l'algorithme du *Random Forest* sera éprouvé. La fonction *RandomForestRegressor* contient plusieurs hyperparamètres utilisés pour éviter le surapprentissage et de chercher de meilleures performances au sens des critères définis. La méthode *Random Search* permet une optimisation non exhaustive des hyperparamètres. Elle est adaptée à l'étude, car un trop grand nombre de combinaisons est possible dans l'hyperparamétrage. En fixant un nombre de 50 combinaisons aléatoires, il est déjà possible de sur-performer par rapport à un modèle sans hyperparamétrage.

La forêt aléatoire possède les mêmes hyperparamètres que les arbres de décision. Le nombre d'arbres composant la forêt est un hyperparamètre supplémentaire. Seuls deux hyperparamètres sont optimisés par la suite :

— **n_estimators** : le nombre d'arbres. Plus ce nombre est élevé, plus le modèle apprend. Un trop faible nombre d'arbres ne permettrait pas *a priori* de décrire suffisamment la variable réponse. À l'inverse, un nombre trop élevé décrirait trop bien la variable réponse et conduirait à des performances faibles sur un nouveau jeu de données,
— **max_depth** : la profondeur maximale des arbres.

Remarque : seul l'hyperparamètre de la profondeur maximale de l'arbre de décision est utilisé, car l'objectif est d'obtenir une forêt dont les arbres soient différents en termes de division en branche, pour, à chaque étape, essayer de capter des effets différents.

Un modèle de forêt aléatoire est lancé sur les variables explicatives sélectionnées pour le comparer au modèle GLM. Cette étude est réalisée sur la variable réponse de coût moyen et sur le nombre de sinistres annualisé (un poids est accordé sur l'exposition).

La recherche aléatoire est réalisée en testant un nombre d'arbres variant de 10 à 400 avec un pas de 20 et une profondeur maximale de 4 à 10 avec un pas de 2 sans distinction des modèles. La cross-validation à 5 folds est ajoutée afin de mieux capter l'effet non aléatoire.

Pour classer les variables discriminantes du modèle, une mesure de l'importance de celles-ci est calculée à l'aide de l'erreur *Out Of Bag* (OOB), c'est-à-dire l'erreur évaluée sur l'ensemble des observations hors échantillon *bootstrap*. En régression, cette erreur est définie comme,

$$errOOB = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

,

L'importance d'une variable donnée est l'accroissement moyen de l'erreur d'un arbre de la forêt lorsque les valeurs observées de cette variable sont permutées au hasard dans les échantillons OOB.

Fixons $j \in \{1, \ldots, p\}$ et X^j la variable étudiée. Soient $\mathcal{L}_n^{\Theta_l}$ un échantillon *bootstrap* et OOB_l l'échantillon associé (l'ensemble des observations qui n'apparaissent pas dans $\mathcal{L}_n^{\Theta_l}$. $errOOB_l$ est donc l'erreur commise sur OOB_l par l'arbre construit sur $\mathcal{L}_n^{\Theta_l}$. Notons $O\tilde{O}B_l^j$, un échantillon perturbé par la permutation aléatoire des valeurs de la j-ième variable dans l'échantillon OOB_l, et $errO\tilde{O}B_l^j$, l'erreur sur l'échantillon $O\tilde{O}B_l^j$.

Alors, l'importance de la variable X^j, est la différence entre l'erreur moyenne d'un arbre sur l'échantillon OOB perturbé et celle sur l'échantillon OOB,

$$VI(X^j) = \frac{1}{q}\sum_{l=1}^{q}(errO\tilde{O}B_l^j - errOOB_l),$$

Une variable est plus importante lorsque les permutations aléatoires de la j-ième variable engendrent une forte augmentation de l'erreur.

Pour le modèle de fréquence et de charge, c'est respectivement l'information géographique et le mode d'achat qui contribue le plus à la prédiction suivie respectivement de la classe de prix et de la note financière. Cette ordre d'importance est très différent du modèle GLM où l'activité de l'entreprise, la zone et l'ancienneté du véhicule ont la plus forte contribution. En effet, plusieurs méthodes permettent de mesurer l'importance des variables pour le XGBoost ou pour le Random Forest, présentant des résultats différents. Il est donc difficile de réellement déterminer un classement de la significativité des variables pour ces méthodes.

Finalement, ce sont les critères suivants qui sont retenus :

— **n_estimators** : 410 pour le modèle de fréquence et 190 pour le modèle de charge,
— **max_depth** : 8 pour les deux modèles.

5.1.2 XGBoost

Extreme gradient boosting (XGBoost) est un algorithme d'apprentissage décomposé en plusieurs sections. XGBoost a été conçu pour être utilisé sur un grand nombre de données complexes.

L'algorithme commence à estimer une variable continue par la moyenne (arbre à une feuille). Puis, il crée un arbre sur les erreurs faites par l'estimation initiale finissant par une mise à l'échelle. Ensuite,

un second arbre est ajusté sur les erreurs produites par l'arbre précédent et ainsi de suite. Les pseudos résidus sont définis ici comme la différence entre l'observé et la prédiction. La prédiction sur une feuille d'un arbre est la moyenne des pseudos résidus des individus inclus dans la feuille. Les prédictions de chaque arbre sont additionnées. Hormis l'estimation initiale, chaque estimateur est multiplié par un taux d'apprentissage. La mise à l'échelle par le taux d'apprentissage se traduit par un petit pas dans la bonne direction.

Traduction mathématique

Soient $(x_i, y_i)_{i=1}^n$ les données d'apprentissage et une fonction différentiable dite de perte $L((y_i, F(x)) = 1/2(y_i - \hat{y}_i)^2$, pour évaluer la performance du modèle où $\hat{y}$ est la prédiction de y.

L'algorithme s'initialise avec la constante $F_0(x) = argmin_\gamma \sum_{i=1}^n L(y_i, \gamma)$ où γ se réfère à la prédiction. Pour $m = 1, \ldots, M$, on réalise les étapes suivantes,

1. $r_{im} = -[\frac{\partial(y_i, F(x_i))}{\partial F(x_i)}]_{F(x)=F_{m-1}(x)}$ pour $i = 1, \ldots, n$,
2. ajustement d'un arbre de régression sur les valeurs r_{im} et création des régions terminales R_{jm} pour $j = 1, \ldots, J_m$,
3. pour $j = 1, \ldots, J_m$, $\gamma_{jm} = argmin_\gamma \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$,
4. $F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_m I(x \in R_{jm})$ où ν est le taux d'apprentissage.

La sortie finale de l'algorithme est $F_m(x)$.

La différence entre XGBoost et Gradient Boosting repose sur l'utilisation des arbres. Le Gradient Boosting est construit à partir de l'arbre CART contrairement à XGBoost qui utilise un arbre de régression unique. À chaque étape de construction de l'arbre d'un XGBoost, un score est calculé, le score de similarité qui est égale à la somme des résidus au carré divisée par le nombre de résidus additionné d'un paramètre $\lambda \leqslant 0$ de régularisation. Il s'agit d'une méthode ascendante appelée *greedy algorithm*. La division d'une racine d'un arbre est réalisée à partir d'une notion de gain. L'algorithme cherche à optimiser ce gain (le plus grand possible) :

$$Gain = Gauche_{similarité} + Droite_{similarité} - Racine_{similarité}.$$

L'arbre peut être élagué en fixant un seuil γ. Si $Gain < \gamma$, alors l'algorithme supprime la branche. L'élagage par la notion de gain n'est pas réalisé sur la totalité des données. L'algorithme utilise une approximation du *greedy algorithm* en séparant les variables continues par des quantiles afin d'éviter de calculer le gain pour toutes les combinaisons possibles. Plus le nombre de quantiles définis est élevé, plus l'algorithme sera long et plus il y a de risque de surapprentissage. Par défaut, le nombre de quantiles est fixé à environ 33 quantiles.

La prédiction d'une feuille est égale à la somme des résidus liés aux individus concernés par la feuille divisée par le nombre de résidus additionné de λ. La différence notable entre Gradient boosting et XGBoost se fait donc sur la manière de diviser les données (notion de gain) et l'ajout d'un paramètre de régularisation en général fixé à 0. Des optimisations sont apportées sur l'algorithme XGBoost afin que celui-ci donne un résultat plus rapide notamment sur la gestion de la mémoire cache.

Afin d'optimiser un modèle XGBoost, les hyperparamètres suivants seront fixés à l'aide d'une recherche aléatoire :

— **min_child_weight** : somme minimale de poids d'instance nécessaire dans une feuille prenant la valeur 0,
— **gamma** : paramètre d'élagage à partir de la notion de gain prenant les valeurs $[0, 0.5, 1, 1.5, 2, 5]$,
— **subsample** : fraction des données utilisées par jeu d'entraînement prenant les valeurs $[0.6, 0.8, 1]$,
— **colsample_bytree** : fraction des variables tarifaires conservées par jeu de données d'apprentissage prenant les valeurs $[0.6, 0.8, 1]$,
— **max_depth** : profondeur maximum de l'arbre prenant les valeurs $[4, 6, 8, 10]$,
— **learning_rate** : taux d'apprentissage prenant les valeurs $[0, 0.1, 0.01, 0.05, 0.2]$.

Plus l'importance des variables dans un XGBoost fournit un score élevé plus un attribut est utilisé pour prendre des décisions clés avec des arbres de décision. Il est calculé, pour un seul arbre de décision,

par le montant que chaque point de division d'attribut améliore la mesure de performance, pondéré par le nombre d'observations dont le nœud est responsable. La mesure de performance peut être la pureté, indice de Gini. Les importances sont ensuite moyennées sur tous les arbres de décision du modèle.

Pour le modèle de fréquence et de coût moyen, c'est l'information géographique qui contribue le plus à la prédiction, suivie respectivement de l'ancienneté et du fractionnement.

Pour vérifier un éventuel phénomène de surapprentissage, l'évolution du RMSE en fonction du nombre d'arbres est tracée graphiquement. En effet, le nombre d'arbres sélectionné grâce à la recherche aléatoire peut-être jugé trop grand ou trop petit en fonction de la représentation graphique du tracé. L'objectif est de trouver un compromis entre performances au sens du RMSE et nombre d'arbres. Un paramètre d'arrêt permet de stopper automatiquement le développement de la forêt lorsque le RMSE augmente au lieu de diminuer.

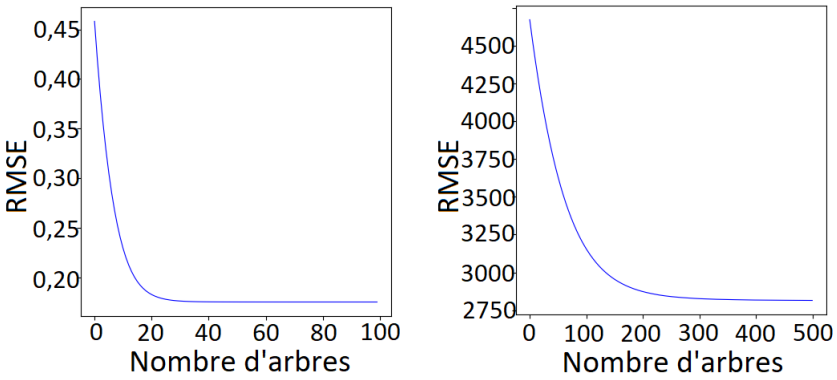


FIGURE 5.1 – Développement des estimateurs XGBoost en fonction de l'erreur quadratique moyenne, à gauche sur le modèle de fréquence et à droite sur le modèle de coût moyen

Le développement des estimateurs du modèle xgboost sur la fréquence s'arrête à partir de 87 au lieu de 100 et sur le coût moyen reste à 500. Ainsi, les paramètres retenus du modèle sont :

— **min_child_weight** : 5,
— **gamma** : $0,5$ pour le modèle de fréquence et 0 pour le modèle de coût moyen,
— **subsample** : 1 pour le modèle de fréquence et $0,6$ pour le modèle de coût moyen,
— **colsample_bytree** : $0,6$ pour le modèle de fréquence et $0,8$ pour le modèle de coût moyen,
— **max_depth** : 8 pour le modèle de fréquence et 6 pour le modèle de coût moyen,
— **learning_rate** : $0,1$ pour le modèle de fréquence et $0,01$ pour le modèle de coût moyen.

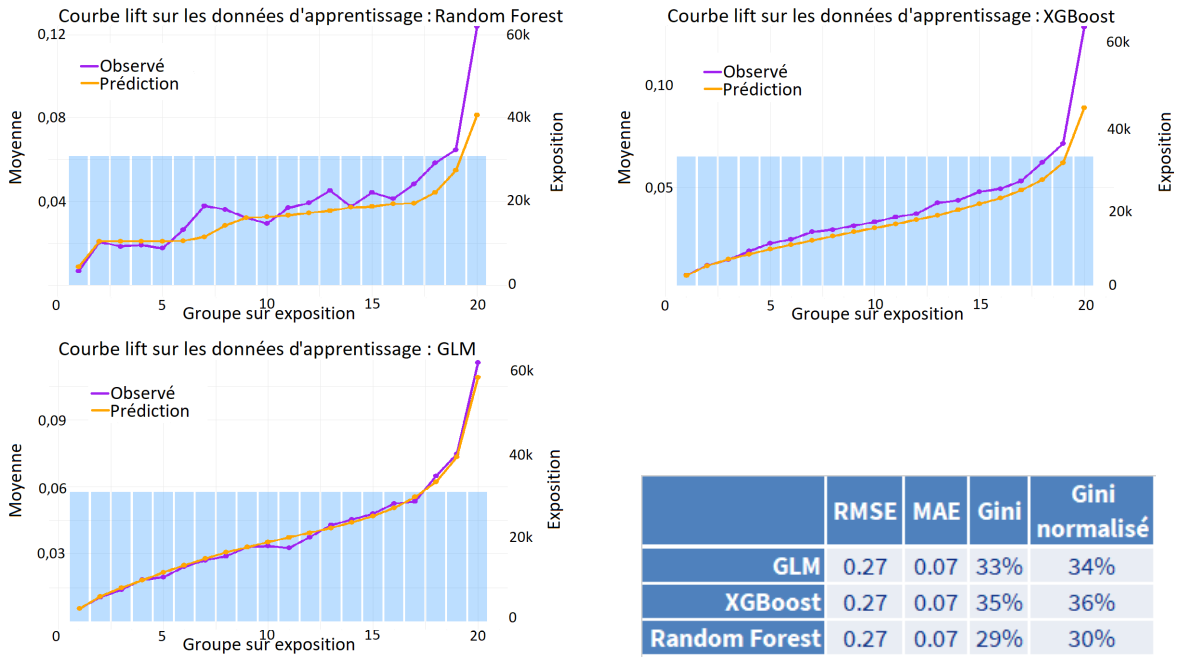
5.2 Pertinence des modèles retenus

Les forêts aléatoires et le XGBoost sont couramment utilisés en assurance. En revanche, ils sont peu employés dans la tarification, à cause de leur complexité. Ainsi, en plus d'une comparaison avec le modèle GLM, une critique des méthodes de machine learning sera apportée afin de juger de la difficulté d'implémentation et d'interprétation.

5.2.1 Pertinence sur la base d'apprentissage

Les méthodes de machine learning sont entraînées à partir des variables sélectionnées pour ajuster le GLM. En réalité, ces méthodes s'adaptent bien avec un nombre élevé de variables. La sélection de variables n'est pas nécessaire. En revanche, l'intégration opérationnelle d'un nombre élevé de variables décourage les clients de créer un devis. D'un point de vue commercial, une tarification adaptée au client est une description du risque à partir d'un minimum de variables répondant à la fois au besoin de segmentation et à la rapidité de souscription.

Sur la base d'apprentissage, les critères de performance des modèles de fréquence octroient :

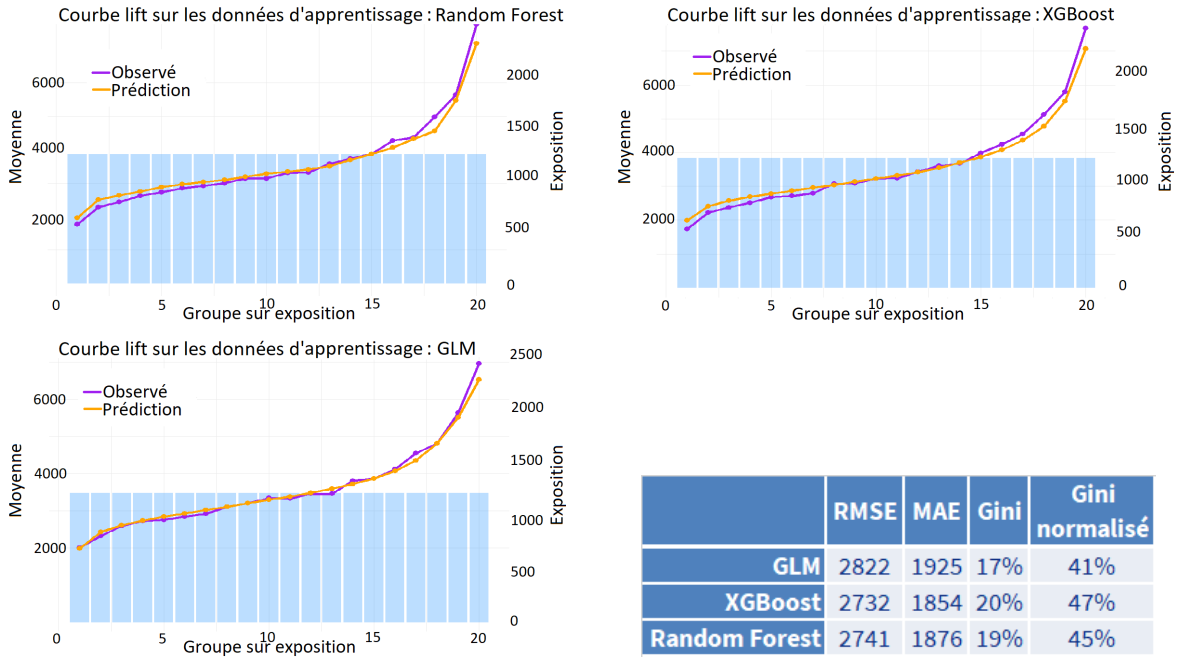


	RMSE	MAE	Gini	Gini normalisé
GLM	0.27	0.07	33%	34%
XGBoost	0.27	0.07	35%	36%
Random Forest	0.27	0.07	29%	30%

FIGURE 5.2 – Critères de performance des modèles de fréquence sur la base d'apprentissage

Les forêts aléatoires ne semblent pas être adaptées pour décrire la fréquence de sinistralité malgré l'utilisation d'un nombre plus grand d'estimateurs que le XGBoost. Une partie de l'explication est la sinistralité rare, la grande fréquence de 0. En revanche, le modèle XGBoost présente de meilleures performances au sens du Gini que les autres modèles.

Pour le modèle de coût moyen, le Random Forest restitue des résultats plus performants que les scores du GLM.



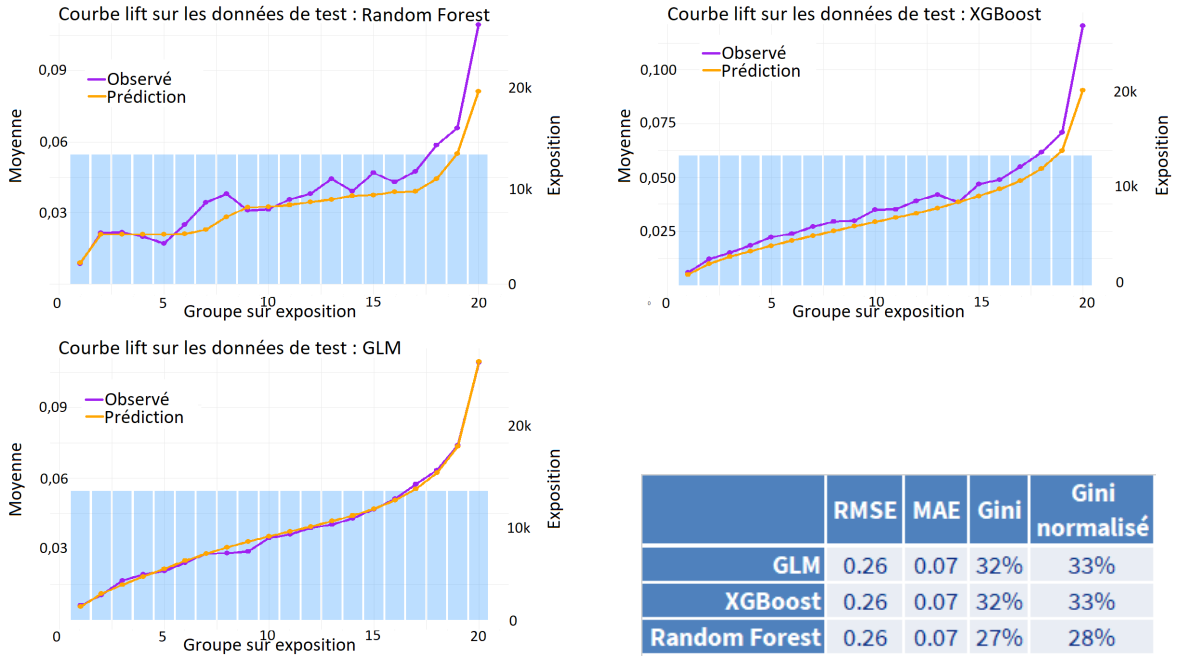
	RMSE	MAE	Gini	Gini normalisé
GLM	2822	1925	17%	41%
XGBoost	2732	1854	20%	47%
Random Forest	2741	1876	19%	45%

FIGURE 5.3 – Critères de performance des modèles de coût moyen sur la base d'apprentissage

Cependant, ces résultats reflètent la performance des modèles sur la base d'apprentissage. Sur un nouveau jeu de données, ces critères changent et peuvent amener à rejeter un modèle. En effet, un bon modèle de prédiction est un modèle qui s'adapte à un nouveau jeu de données. Les résultats précédents informent seulement sur le pouvoir d'ajustement des machines.

5.2.2 Pertinence sur la base test

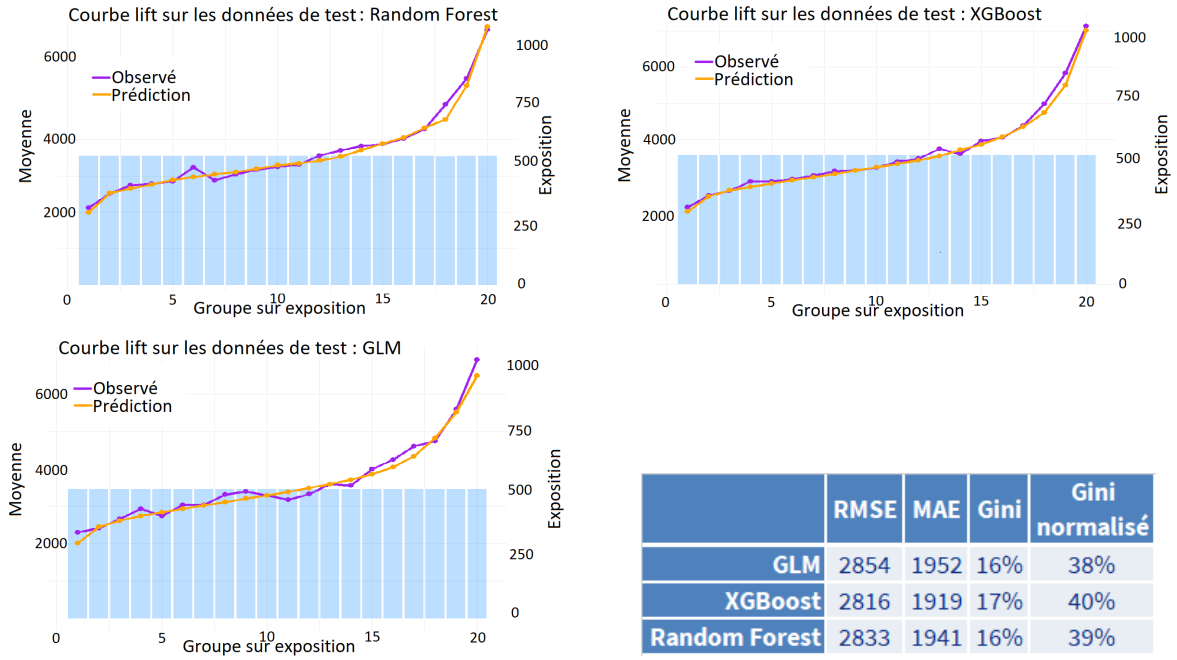
La base test est un jeu de données qui n'a pas servi à l'ajustement des modèles. Pour rappel, elle représente 20% de l'ensemble du jeu de données.



	RMSE	MAE	Gini	Gini normalisé
GLM	0.26	0.07	32%	33%
XGBoost	0.26	0.07	32%	33%
Random Forest	0.26	0.07	27%	28%

FIGURE 5.4 – Critères de performance des modèles de fréquence sur la base test

Pour le modèle de fréquence, les performances des modèles sur la base test sont légèrement inférieures par rapport aux performances sur la base d'ajustement. Cela prouve que nos trois méthodes s'adaptent au jeu de données. La baisse de performance est certainement liée à une faible captation du bruit, caractère aléatoire de la sinistralité. La cross-validation, la sélection de variables, la crédibilisation, l'hyperparamétrage, les méthodes de bootstrap ou encore la régularisation sont des méthodes qui ont permis d'augmenter le pouvoir prédictif des modèles.



	RMSE	MAE	Gini	Gini normalisé
GLM	2854	1952	16%	38%
XGBoost	2816	1919	17%	40%
Random Forest	2833	1941	16%	39%

FIGURE 5.5 – Critères de performance des modèles de coût moyen sur la base test

Finalement, le XGBoost obtient une performance meilleure que les autres modèles, pour la fréquence et le coût moyen. Il s'agit donc de l'ajustement qui devrait être retenu. Cependant, l'implémentation d'une machine XGBoost ou Random Forest est complexe en assurance, car la définition de cases tarifaires est moins claire. De plus, la description de l'influence des variables sur le modèle est également floue. La compréhension de l'effet des variables étant un point indispensable dans la segmentation du risque, ces modèles ne sont en général pas employés.

Chapitre 6

Résultats de la modélisation

Les assureurs ont pour but de mieux segmenter leur tarif afin d'être plus compétitifs sur le marché. Un assuré va préférer prendre un assureur proposant le tarif le plus bas pour une même prestation de service. Cependant, une segmentation élevée du risque peut impliquer une sous-performance des modèles. Il s'agit donc d'optimiser la segmentation. De plus, d'un point de vue opérationnel, il est préférable de segmenter le risque en faisant appel à un minimum de variables. En effet, cela permet de faciliter la souscription d'un contrat d'assurance en réclamant un minimum d'informations à l'assuré. Le tarif actuellement mis en place est soumis à plusieurs contraintes opérationnelles. Il s'agira donc de comparer cette prime à la prime modélisée sans contrainte afin de vérifier deux points :

— la prime actuelle est-elle bien ajustée malgré les contraintes informatiques ?
— l'ajout de nouvelles variables, notamment à partir des informations de la base SRA, est-il pertinent ?

Enfin, en raison des nombreux retraitements de la charge, la prime modélisée sera réconciliée avec la charge brute tout en confrontant celle-ci avec la prime avec contrainte.

6.1 Comparaison des primes actuelles et modélisées

Le tarif actuellement commercialisé en auto-entreprise répond à des besoins opérationnels pour faciliter la souscription, mais également à des stratégies commerciales. Par exemple, le dirigeant de l'entreprise, selon la catégorie du véhicule et la garantie, peut compter sur une réduction d'environ 20% sur la prime de son véhicule.

Le tarif appliqué est également soumis à des contraintes informatiques. Lors de la création du tarif, des données relatives aux véhicules ou à la santé financière de l'entreprise assurée n'étaient pas disponibles. La dernière refonte majeure de tarification a eu lieu en 2016 avec l'introduction du fichier SIV. Depuis, environ tous les 3 ans, des mises à jour régulières sont produites afin d'actualiser les coefficients de la modélisation et la constante. Cela permet de prendre en compte notamment l'inflation du coût des pièces, des véhicules ou encore des salaires en cas de responsabilité civile menant à l'incapacité de travail d'une victime d'un accident. Au bout de 3 ans, le risque peut avoir dévié et donc la segmentation peut être amenée à être modifiée. Cependant, en raison des contraintes informatiques, cette segmentation ne peut pas être modifiée à moins de mobiliser un budget important. Le sujet de ce mémoire est donc de construire un modèle sans contrainte axé sur la segmentation rentrant en compétitivité avec le tarif actuel. Puis, l'objectif est de déduire le gain d'un tel modèle. Dans ce modèle sans contrainte est introduit de nouvelles données. Lors de la création de la base de données, le but était d'être le plus exhaustif possible en recensant l'ensemble des variables à disposition.

Les deux ajustements, avec et sans contrainte, ont été construits sous un modèle collectif, fréquence $\times$ coût moyen. De plus, les coefficients du tarif actuel ont été mis à jour à l'aide d'un modèle de coût moyen dont la charge est privée des sinistres graves (charge supérieure au seuil de 30 000), des sinistres dont le coût est forfaitaire (forfaits IDA et coûts d'ouverture) et des recours. Ce traitement est relativement similaire au modèle sans contrainte retenu. Les points qui diffèrent entre les modèles sont inscrits dans le tableau suivant :

	Tarif avec contrainte	Tarif sans contrainte
Sinistres graves	hors grave	graves mutualisés
Seuil des graves	30000	18000
Forfait IDA	sans	sans
Coûts d'ouverture	sans	sans

FIGURE 6.1 – Tableau des différences de traitement entre le tarif actuel et le tarif modélisé

Ces dissimilarités ont un impact infime sur les modèles de fréquence et de coût moyen. Il est donc possible de comparer les performances statistiques de chacun des ajustements. Il s'agit de mesurer le gain de performance du modèle sans contrainte. Évidemment, les bases d'apprentissage et de test sont différentes.

		Tarif avec contrainte			Tarif sans contrainte		
	Modèles	Fréquence	Coût moyen	Prime Pure	Fréquence	Coût moyen	Prime Pure
Base d'apprentissage	RMSE	0,40	3426	1604	0,27	2822	1209
	MAE	0,16	2089	335	0,07	1925	267
	Gini	27%	7%	28%	33%	17%	43%
	Gini normalisé	28%	11%	29%	34%	41%	42%
Base de test	RMSE	0,40	3537	1597	0,26	2854	1231
	MAE	0,16	2129	334	0,07	1952	268
	Gini	27%	7%	28%	32%	16%	44%
	Gini normalisé	28%	12%	29%	33%	38%	42%

FIGURE 6.2 – Tableau des performances du tarif actuel et du tarif modélisé

Le gain de performance est relativement important qu'il s'agisse du modèle de fréquence ou du modèle de coût moyen. L'augmentation de +5 points de Gini sur le modèle de fréquence et de +10 points sur le modèle de coût moyen montre que le modèle sans contrainte est meilleur au sens des critères de performance. La segmentation est également améliorée. Après l'agrégation des modèles de fréquence et de coût moyen, le gain de gini est de plus de +10 points suggérant grandement l'implémentation de ce nouveau modèle. La différence est plus notable sur le modèle de coût moyen à l'aide de l'introduction des variables SRA.

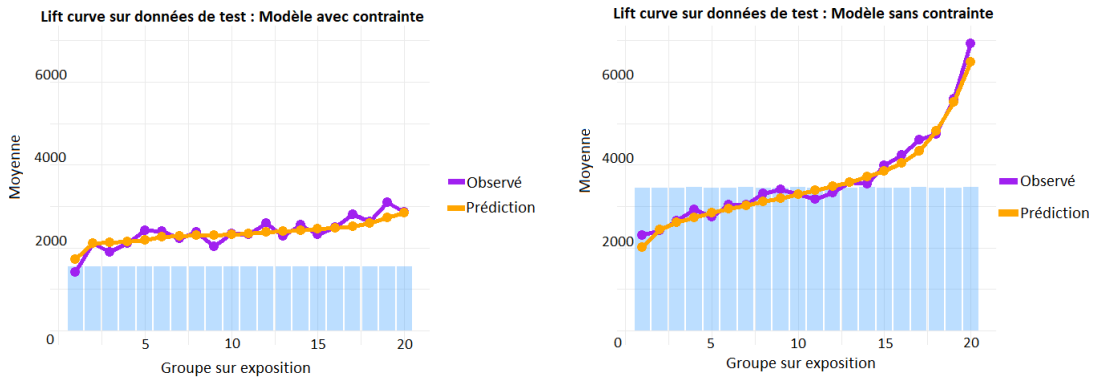
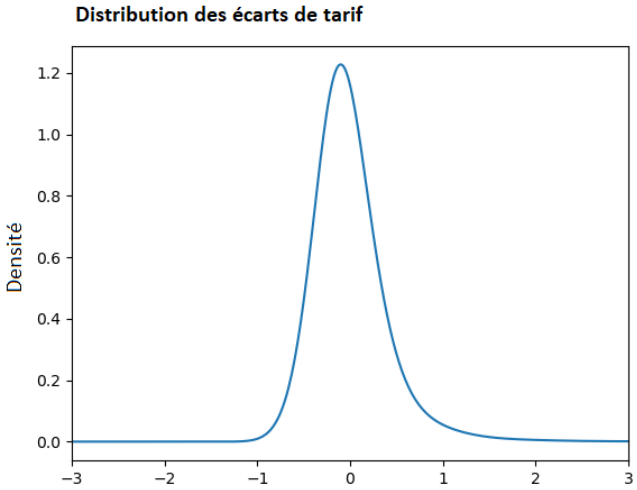


FIGURE 6.3 – Courbe lift des modèles de coût moyen sur les données de test

En matière de courbe lift, on constate que le modèle sans contrainte semble mieux capter la variabilité en regardant les échelles des graphiques. En effet, on remarque une faible croissance de la lift curve de la prédiction du modèle de coût moyen avec contrainte.

Le modèle sans contrainte est mieux ajusté à la sinistralité. Cependant, il n'est pas clair que ce soit l'effet des variables véhicules qui explique cette différence. Une analyse des coefficients de tarification ne suffit pas pour déterminer les réelles différences entre les primes à cause des corrélations qui subsistent entre les variables. La comparaison entre le tarif actuel et le tarif modélisé s'effectue alors par une analyse des écarts. Cette analyse a pour but de définir l'impact chiffré sur le portefeuille et d'identifier les profils les plus impactés par des hausses ou des baisses tarifaires. La mesure de l'impact de l'évolution tarifaire peut se mesurer en représentant la distribution des écarts de tarifs et la dispersion des primes. Par la suite, le tarif actuel et le tarif modélisé correspondent à la prime pure y compris forfaits, recours et sinistres graves. La prime actuelle est privée de tout chargement et d'éventuelles majorations annuelles. L'écart de tarif est défini comme le rapport entre le tarif actuel et le tarif modélisé [1].



	Exposition	Nombre de véhicules	Ecarts
]0,5;+∞[	7%	7%	27%
]0,2;0,5]	12%	12%	16%
]0;0,2]	20%	20%	7%
]-0,2;0]	32%	31%	13%
]-0,5;-0,2]	26%	26%	31%
]-∞;-0,5]	3%	3%	7%

FIGURE 6.4 – Distribution des écarts entre le tarif actuel et le tarif modélisé

La distribution des écarts de tarif est quasi-centrée en zéro ce qui signifie que la prime totale est presque identique dans les deux cas. D'ailleurs, cette information est retranscrite à l'aide du tableau sur la dispersion des écarts. La différence de -1% entre le tarif actuel et le tarif modélisé peut provenir de la mise en as-if (inflation) ou de la différence tarifaire sur la base test. De plus, la constante du tarif actuel a été mise à jour à partir d'une étude de la sinistralité. L'écart est donc rassurant, car il permet de considérer une petite marge d'erreur liée notamment à la perte d'information de la sinistralité sur le tarif modélisé (sous-section 2.2.4, page 36). De plus, la courbe de distribution suit à peu près une distribution d'une loi normale ce qui est également réconfortant. En effet, cela permet de s'assurer que les déviations de tarif sont minimes et que globalement, il n'y a pas d'informations évidentes qui n'auraient pas été captées par le tarif actuel. Cependant, la courbe présente une légère asymétrie. D'après le tableau (Figure 6.4), 20% des cas étudiés ont des tarifs plus hauts avec les primes modélisées. Cela peut s'expliquer en partie par la baisse relative aux véhicules dirigeants représentant un faible nombre de véhicules, mais qui implique l'augmentation des primes des autres véhicules par le principe de mutualisation.

Pour comprendre comment se lit le tableau, prenons un exemple : l'intervalle $]0,2;0,5]$ représente l'ensemble des écarts $\frac{modèle\ sans\ contrainte}{modèle\ avec\ contrainte} - 1$ tel que 16% de la valeur absolue de ces écarts, c'est à dire 16% de la différence de charge, y est expliquée représentant 12% des individus et de l'exposition. L'asymétrie est donc détectée par la présence d'une fréquence élevée d'écarts entre $]-0,5;0]$ expliquant une partie importante des écarts négatifs. Ensuite, 1/3 des écarts se situent dans l'intervalle $]0,5;+\infty[$. Par la suite, on essaiera de capter la variabilité des écarts.

Les profils à l'origine des écarts de tarifs peuvent être analysés en expliquant le rapport de primes entre l'actuel et le modélisé par un arbre de décision. Cet arbre est élagué à l'aide du paramètre de complexité. De plus, le nombre de feuilles a été limité afin de le rendre lisible.

1. $\frac{tarif\ actuel}{modèle} - 1$

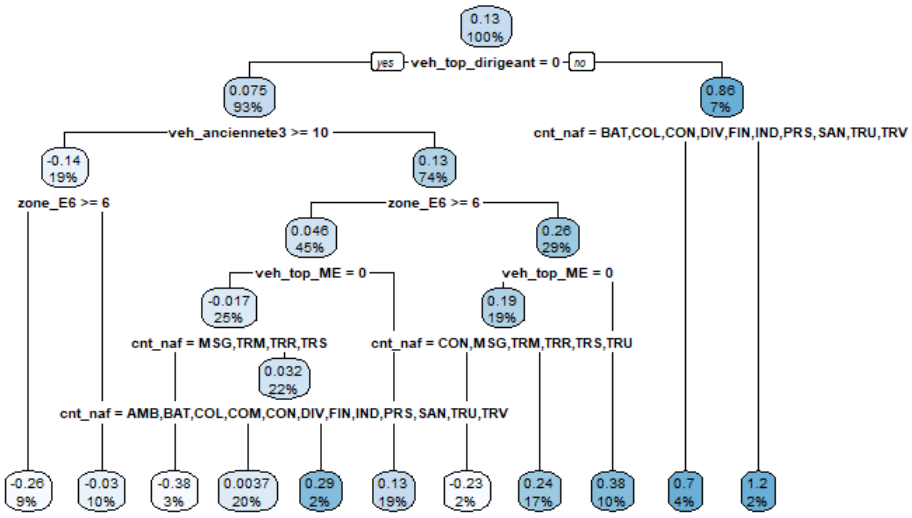


FIGURE 6.5 – Arbre de décision sur les écarts de tarif

D'après l'arbre de décision, la variable qui introduit le plus d'écart est le topage du véhicule dirigeant. Il s'agit d'une contrainte commerciale qui n'est donc pas intéressante d'étudier, car elle sera de toute manière introduite dans le modèle commercialisé. Une grande partie de la variabilité des écarts est expliquée par le code d'activité : *cnt_naf*. Cela s'explique par une volonté d'appliquer une forte augmentation de tarif sur certaines activités comme le transport de voyageurs ou le transport de marchandises (TRV, TRM) où le coût moyen a radicalement augmenté ces dernières années. Les écarts ne semblent pas être expliqués par les variables SRA. En réalité, il est normal de ne pas observer ces variables influencées la segmentation de l'arbre. En effet, pour mieux visualiser l'arbre, il a fallu limiter le nombre de feuilles. De plus, ce sont les variables les plus significatives de la modélisation comme la zone, l'activité de l'entreprise ou l'ancienneté qui sont également les plus significatives pour décrire les écarts. Afin de mieux visualiser si les variables véhicules décrivent les écarts, les trois variables citées précédemment ainsi que les variables commerciales sont mises de côté. La corrélation entre les variables explicatives restantes et les variables écartées ne sera donc pas prise en compte dans l'arbre suivant :

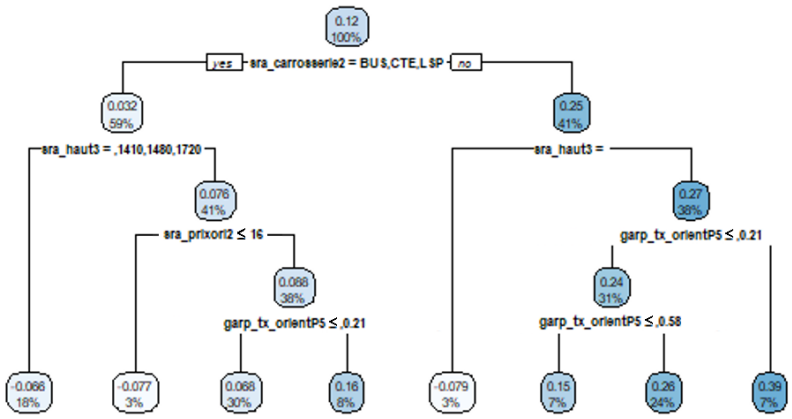


FIGURE 6.6 – Arbre de décision sur les écarts de tarif en écartant les variables commerciales et les plus significatives

Les variables introduites dans le modèle sans contrainte semblent expliquer une part de l'écart de tarif. Le modèle avec contrainte et le modèle sans contrainte sont ajustés à l'aide des informations géographiques, activité de l'entreprise et ancienneté du véhicule. De plus, les variables commerciales concernent des cas de véhicules peu fréquents. On peut donc émettre l'idée que ce sont les variables véhicules et le taux d'orientation garage partenaire qui décrivent le mieux le gain de performance du modèle sans contrainte. Les arbres mettent en avant les cases tarifaires sur lesquelles on va créer de la segmentation

grâce aux variables SRA.

6.2 Réconciliation de la prédiction à la charge

La prédiction établie par le modèle retenu est l'addition de deux primes. Les charges négatives, forfaits IDA et les coûts d'ouvertures sont traités hors modélisation et sont implémentés par la suite à l'aide d'une prime forfaitaire[2]. De plus, la charge grave, définie par le seuil de la TVE, a été mutualisée au prorata de la prime des sinistrés. L'ensemble de ces retraitements de la charge rend la prime finale peu claire. Ainsi, il s'agit de contrôler que la prédiction sur une base non entraînée converge vers la charge brute. La prime modélisée est inflatée à fin septembre 2021 au même titre que la prime actuelle à l'aide d'un indice composite entre la RC CORP et la RC MAT. Cependant, la prime est annuelle contrairement à la charge. En effet, la prime a été annualisée à l'aide du poids accordé par la durée d'exposition au risque. De ce fait, malgré un redimensionnement de la charge par l'exposition, la moyenne de la prime ne sera pas égale à la moyenne de la charge. Pour réconcilier la prime et la charge, les moyennes des deux primes seront posées égales à la moyenne de la charge suite à la pondération des primes par l'exposition. C'est la variation par case tarifaire qui sera contrôlée. De plus, pour valider la construction du modèle, la prime actuelle sera représentée. Des tendances similaires et une prime totale comparable sont attendues. Dans le cas contraire, l'un des tarifs n'a pas été construit de façon correcte. Des différences de tarifs seront toutefois constatées, dues à la limitation du tarif actuel, lui-même dû à des contraintes informatiques.

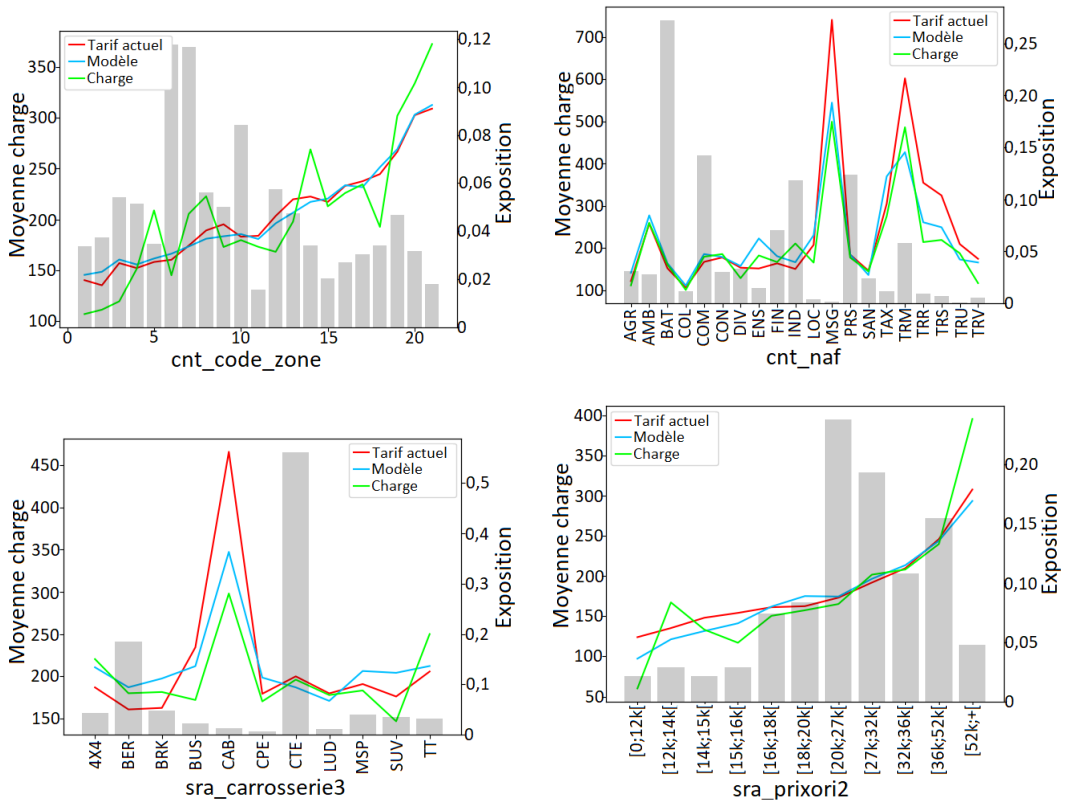
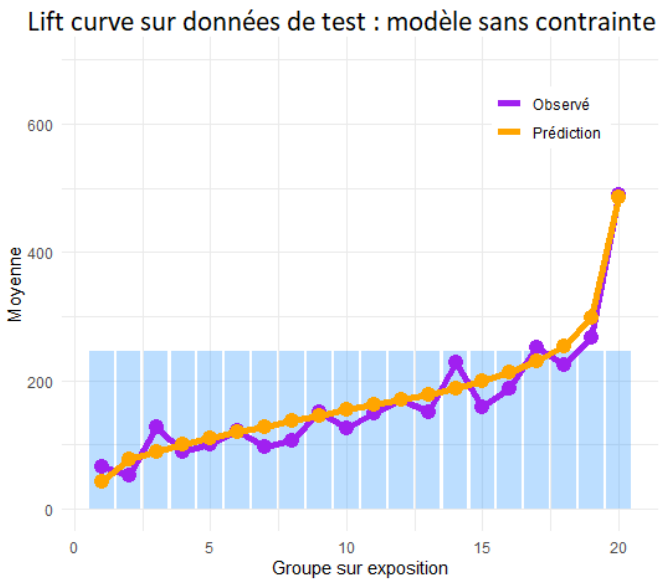
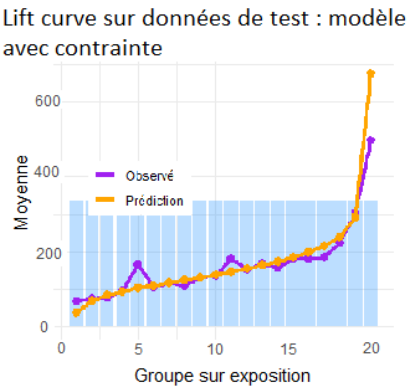


FIGURE 6.7 – Réconciliation des coefficients entre la charge et la prédiction

Ces quelques graphiques permettent de représenter les variations de primes ou charge moyenne selon des cases tarifaires. Il est intéressant de remarquer que les prédictions semblent proches et les tendances sont similaires avec la charge malgré une variation plus importante de la charge brute. Cette variation s'explique par une part de bruit et une sous-exposition pour certaines cases tarifaires rendant peu crédible la sinistralité.

2. La prime forfaitaire sur les charges négatives, les forfaits IDA et les coûts d'ouvertures est introduite par la formule, $\frac{\sum charge\ des\ sinitres\ forfaitaires}{\sum exposition}$





	RMSE	MAE	Gini	Gini normalisé
Tarif actuel	2514	309	26%	27%
Modèle	2514	311	27%	28%

FIGURE 6.8 – Réconciliation entre la charge et la prédiction

En raison de la mise en offset de l'exposition, la courbe lift présente une variation relativement importante par quantile. Les performances du modèle avec contrainte et du modèle sans contrainte pour prédire la charge brute sont similaires. D'après le Gini, le modèle sans contrainte informatique semble avoir une meilleure segmentation par rapport au tarif actuel. À présent, le modèle sans contrainte est réconcilié avec la charge brute.

Contrairement à la partie précédente, le gain de performance est léger. Cela s'explique par la présence d'un nombre important de sinistres forfaitaires biaisant les performances. En effet, les sinistres dont le montant correspond à un forfait IDA, un recours ou un coût d'ouverture représentent 64% de la fréquence et 26% de la charge. Finalement, le modèle sans contrainte obtient de meilleures performances que le modèle actuellement mis en place, notamment sur le coût moyen, mais ce gain est moindre. Ce n'est donc pas sur la garantie RC que l'on peut vraiment observer un gain de performances. En revanche, il serait intéressant d'étudier l'impact de l'inclusion des variables véhicules sur des garanties privées des conventions IRSA et IRCA comme la garantie dommage. Cette dernière indemnise les frais de réparation ou le coût du véhicule de l'assuré et représente la deuxième prime la plus importante en assurance automobile.

Conclusion

Le produit *parc dénommé* représente une part importante de l'assurance automobile entreprise. Plusieurs garanties y sont proposées dont la garantie obligatoire Responsabilité Civile. Le modèle de tarification actuellement utilisé par la branche est segmenté par catégories de véhicules. 60% des véhicules composant l'ensemble des parcs dénommés correspondent à des véhicules légers à 4 roues, en général utilitaires (catégorie 1 de véhicule). De nouvelles données sont disponibles sur ce genre de véhicule. Il est donc judicieux de réaliser une étude sur l'impact d'un tarif prenant en compte ces informations. Ce sujet n'a pas été exploité lors de la refonte tarifaire 2021 à cause de fortes contraintes budgétaires. De plus, cette révision n'a pas donné lieu à un changement de la segmentation du tarif. *In fine*, l'objectif de ce mémoire est de comparer les performances d'un modèle réalisé sans contrainte informatique avec les performances du tarif actuel. En effet, la performance d'un modèle est étroitement liée à la pertinence de la segmentation et donc, plus le modèle est performant, plus le tarif est adapté au client.

Cette étude permet de prendre une décision sur le changement de la segmentation du risque en comparant performance et retour sur investissement. Dans le cas où la réponse serait positive, le modèle sans contrainte remplacerait le tarif actuel sur les affaires nouvelles et, dans le futur, il serait implémenté dans l'outil de tarification des contrats renouvelés. Dans le cas contraire, il serait envisageable d'introduire un nouveau modèle de majoration des contrats au moment du renouvellement, rapport entre la prime modélisée et le tarif actuel (ELR). Il permettrait une meilleure segmentation du risque pour appliquer des majorations plus proches de la réalité.

Afin d'arriver à un tarif le plus performant que possible, la construction du jeu de données a fait l'effet d'un traitement approfondi. L'un des objectifs était d'introduire de nouvelles données dont les données véhicule et de les fiabiliser. La complétion des données a été réalisé à l'aide d'algorithmes qui ont permis de multiplier par $3,68$ le taux de complétude de l'information véhicule provenant de la base SRA et de diviser par 6 le taux de sinistres non captés. La modélisation de la prime pure s'est déroulée en 4 étapes.

Une première étape consistait à déterminer quelle part de la charge pourrait venir biaiser le modèle linéaire généralisé. Il s'est avéré que les forfaits IDA provenant de conventions accélérant le versement des prestations, les recours et les coûts d'ouverture se retrouvaient finalement modélisés séparément du reste de la sinistralité. De plus, la sinistralité grave biaise également le modèle. Un seuil a été déterminé pour définir un sinistre grave à partir de la théorie des valeurs extrêmes.

Ensuite, une seconde étape reposait sur le calibrage du modèle. En effet, un modèle fréquence $\times$ coût moyen a été retenu pour mieux capter des effets inverses entre les deux modèles. De plus, il n'a pas été jugé nécessaire de diviser la modélisation de la garantie RC en deux selon le type de dommage qu'implique le sinistre (matériel ou corporel) en raison de la forte volatilité du modèle de la garantie RC due à des dommages corporels.

En troisième lieu, un travail approfondi sur le regroupement et la classification des variables, complété d'une sélection parcimonieuse de variables, ont permis d'augmenter la performance du modèle. En supplément, des méthodes de regroupement tel que le zonier ou le véhiculier ont permis d'examiner plus en détail respectivement l'information géographique et l'information véhicule. En effet, ces deux informations posaient questions, par exemple comment expliquer les zones à risques, quels sont les interactions entre les variables véhicules. Finalement, les zones à risques n'ont pas pu être concrètement expliquées par des données externes comme l'état de la route ou la météo, et malheureusement le véhiculier n'a été que peu efficace en matière de gain de performance.

Enfin, en quatrième étape, une comparaison du modèle linéaire généralisé et d'autres machines d'apprentissages dont le XGBoost et le Random Forest, a été réalisée. Le GLM mène à des résultats facilement exploitables et interprétables, ainsi, malgré une faible perte de performance, ce modèle est conservé pour décrire le risque.

Finalement, le modèle sans contrainte se révèle être plus performant que le tarif actuel sur la charge hors forfait IDA, recours et hors coûts d'ouverture. Un gain notable de plus de 10 points de Gini après agrégation des modèles de fréquence et de coût moyen, soit un rapport de 145%, prouve l'importance d'ajouter les variables de la base SRA dans le modèle, et de reconsidérer la segmentation du tarif actuel. En fait, le gain de performance est lié à trois facteurs, l'un est l'introduction de nouvelles informations :

les variables véhicules SRA, la santé financière des entreprises et le taux d'orientation vers les garages partenaires. Le second facteur est que la segmentation n'a pas été révisée depuis 2016 et peut s'avérer pour une part obsolète, notamment sur l'information géographique où un zonier spécifique au modèle de coût moyena été construit. Ce zonier participe à $+1,5$ points de Gini sur le modèle de coût moyen. Enfin, des variables commerciales et une dérive de la segmentation du risque par rapport à 2016 peut biaiser le tarif actuel.

Cependant, le gain de performance, en considérant la charge des sinistres forfaitaires, les recours et les coûts d'ouverture, est de seulement +1 point. En effet, cette sinistralité particulière représente 26% de la charge totale et 64% du nombre de sinistres, d'où ce faible gain. Ainsi, le retour sur investissement n'est pas bon. Trois raisons nous amènent à cette conclusion : le coût d'implémentation d'un tarif sans contrainte, c'est-à-dire le fait d'aller récupérer les informations véhicules de la base SRA, le faible gain tarifaire à l'affaire nouvelle et le fait que les affaires nouvelles représentent chaque année 10% du portefeuille. Il est préférable d'utiliser le tarif sans contrainte en tant qu'outil de majoration des contrats au moment du renouvellement, car on impacte tout le portefeuille sans coût informatique supplémentaire.

Quelques améliorations pourraient être faites dans la continuité de cette étude. Dans un premier temps à reconsidérer l'étude sur le véhiculier : En effet, le voisinage n'est pas assez étendu. D'autres méthodes sont à envisager comme utiliser la distance euclidienne au lieu de la triangulation de Delaunay. Dans un second temps, les résultats sur la comparaison du tarif sans contrainte et du modèle actuel démontrent un potentiel de gain de performance relativement élevé, mais qui sont fortement détériorés par les conventions IRSA et IRCA. Ainsi, l'étude devrait être élargie à d'autres garanties comme la garantie dommage. Il s'agit d'une couverture des dommages matériels du véhicule de l'assuré. De plus, la base SRA est composée également de véhicules de catégorie 6, les véhicules à deux roues additionnées des quads, ce qui peut faire l'effet d'un examen approfondi du modèle et de l'impact des nouvelles données véhicules.

*

Bibliographie

BERARD, J. (2020). Modèles linéaires généralisés. *Cours master 1.*

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

BORCHANI, A. (2010). Statistiques des valeurs extrêmes dans le cas de lois discretes.

DATA.GOUV. (2020). *Bases de données annuelles des accidents corporels de la circulation routière - Années de 2005 à 2019.* https://www.data.gouv.fr/fr/datasets/bases-de-donnees-annuelles-des-accidents-corporels-de-la-circulation-routiere-annees-de-2005-a-2019/

DATA.GOUV. (2021). *Classes d'état des chaussées du réseau routier national non concédé à partir de 2019.* https://www.data.gouv.fr/fr/datasets/classes-detat-des-chaussees-du-reseau-routier-national-non-concede-a-partir-de-2019/

FERRIER, C. (2016). Le zonier en tarification IARD : approche comparative de deux techniques de construction d'un critère de segmentation géographique en assurance habitation. *Mémoire d'actuariat.*

GARDES, L. (2020). Théorie des valeurs extrêmes. *Cours master 2.*

GENUER, R. & POGGI, J.-M. (2017). Arbres CART et Forêts aléatoires, Importance et sélection de variables. *hal-01387654v2f.*

HENCKAERTS, R., ANTONIO, K., CLIJSTERS, M. & VERBELEN, R. (2017). A data driven binning strategy for the construction of insurance tariff classes.

KHARROUBI, I. (2014). *Actuariat Introduction.*

LANCELOT, R. & LESNOFF, M. (2005). Sélection de modèles avec l'AIC et critères d'information dérivés. *CIRAD, Montpellier.*

LAVENU, J. (2016). Les méthodes de Machine Learning peuvent-elles être plus performantes que l'avis d'experts pour classer les véhicules par risque homogène ? *Mémoire d'actuaire.*

LINTERNAUTE. (2021). *Les départements les plus ensoleillés de France en 2020.* https://www.linternaute.com/voyage/climat/classement/departements/soleil/2020

METHNI, J. E. (2013). *Contributions à l'estimation de quantiles extrêmes. Applications à des données environnementales* (thèse de doct.). Université de Grenoble.

MINISTÈRE-DE-L'INTÉRIEUR. (2018). *Création du fichier des véhicules assurés (FVA).* https://mobile.interieur.gouv.fr/Archives/Archives-des-communiques-de-presse/2018-Communiques/Creation-du-fichier-des-vehicules-assures-FVA

MONNIER, D. (2015). Modèles linéaires généralisés et assurance santé individuelle : Tarification et évaluation des engagements sous solvabilité II. *Mémoire d'actuariat.*

NASRI, B. & REMILLARD, B. (2017). Miracles et coïncidences : évènements rares et théorie des valeurs extrêmes.

OLIVEAU, S. (2010). Autocorrélation spatiale : leçons du changement d'échelle. *L'Espace geographique*, *39*(1), 51-64.

QUILFEN, M. (2020). Tarification Non-Vie. *Cours master 1.*

RUIMY, M. (2017). Calcul de la valeur contrat sur la branche Multirisque Immeuble comme aide opérationnelle à la relation client. *Mémoire d'actuaire.*

THUILLIER, M. (2020). Calcul de la valeur contrat sur la branche Multirisque Immeuble comme aide opérationnelle à la relation client. *Mémoire d'actuaire.*

TREMBLAY, C. (2016). Prédire les sinistres graves en assurance : les apports de l'apprentissage statistique aux modèles linéaires. *Mémoire d'actuariat.*

TUTZ, G. & BERGER, M. (2018a). Tree-structured clustering in fixed effects models. *Journal of computational and graphical statistics*, *27*(2), 380-392.

TUTZ, G. & BERGER, M. (2018b). Tree-structured modelling of categorical predictors in regression. *Advances in Data Analysis and Classification*, *12*(3), 737-758.

ZOUGGAGH, F.-Z. (2018). Tarification automobile à l'aide de modèles de machine learning et apport des données télématiques.

Table des figures

Annexes

A Traitements de la charge

A.1 Forfaits

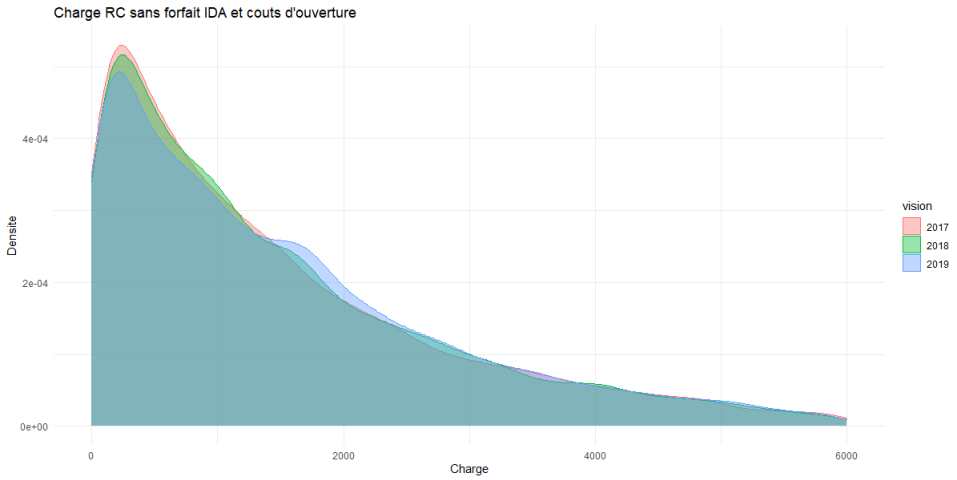


FIGURE A.1 – Densité de la charge RC

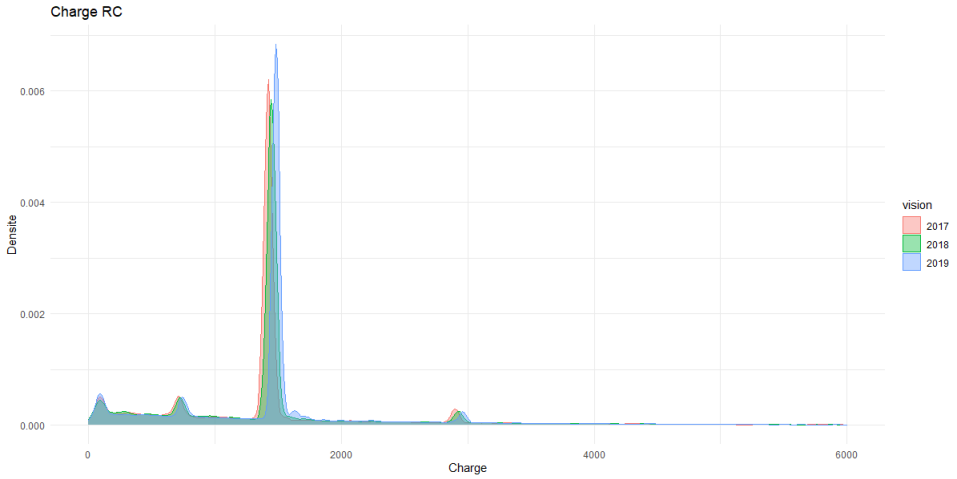


FIGURE A.2 – Densité de la charge RC exclue des forfaits IDA et AXA

A.2 Petits sinistres

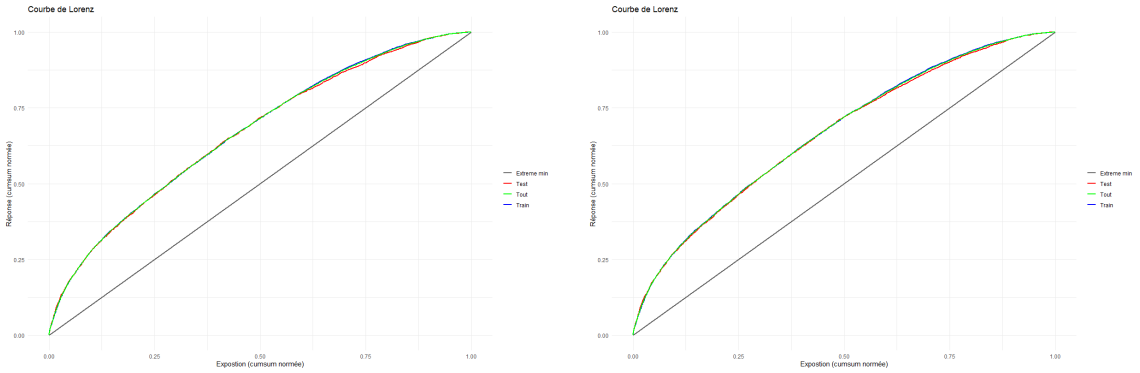


FIGURE A.3 – Courbe de Lorenz entre le modèle hors grave et hors sinistre > 120 (à gauche) et le modèle hors grave (à droite)

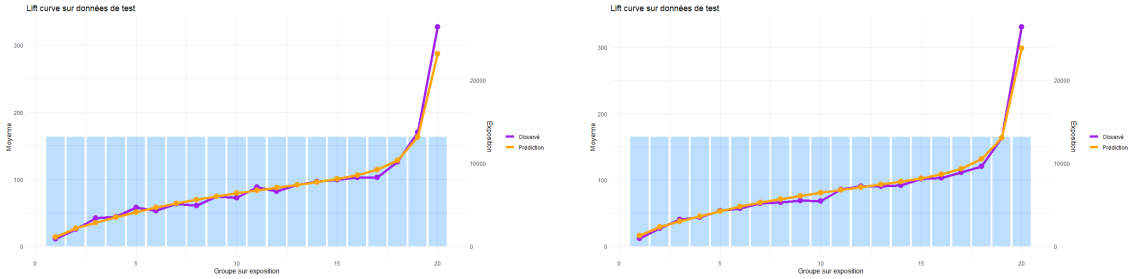


FIGURE A.4 – Courbe lift entre le modèle hors grave et hors sinistre > 120 (à gauche) et le modèle hors grave (à droite)

B Théorie des valeurs extrêmes

B.1 RCM

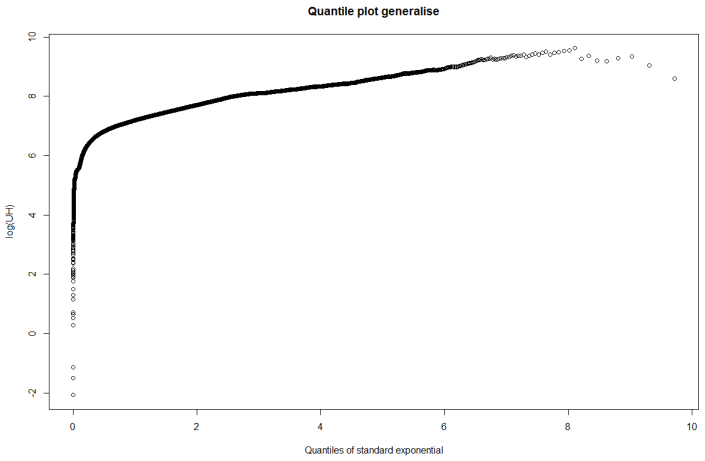


FIGURE B.5 – Quantile plot généralisé de la RCM

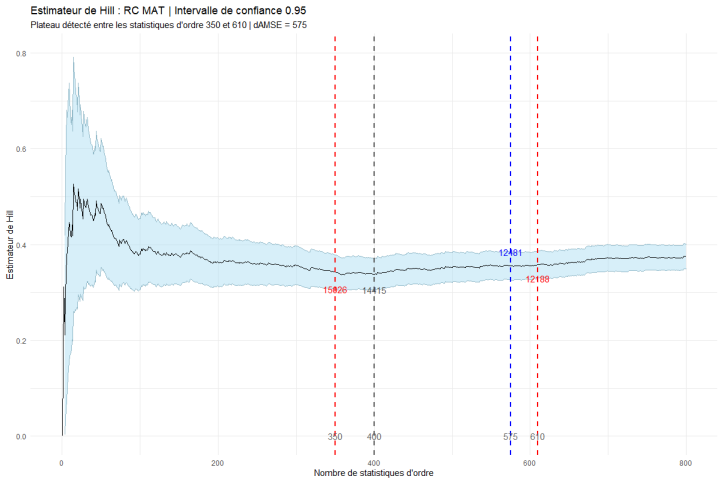


FIGURE B.6 – Évolution de l'estimateur de Hill selon k_n

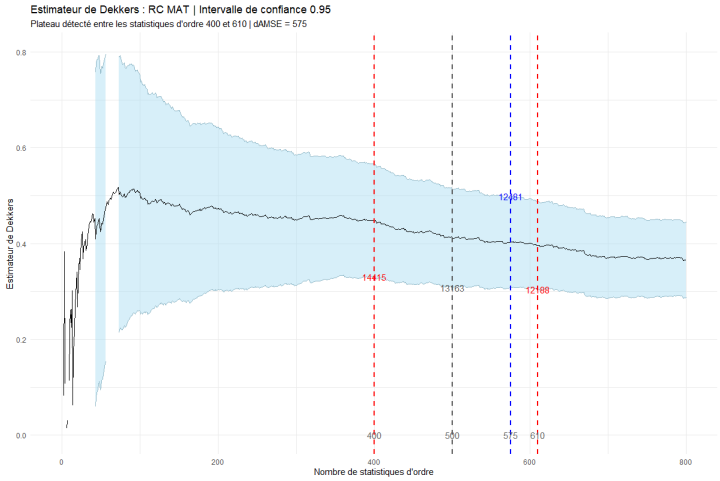


FIGURE B.7 – Évolution de l'estimateur de DEdH selon k_n

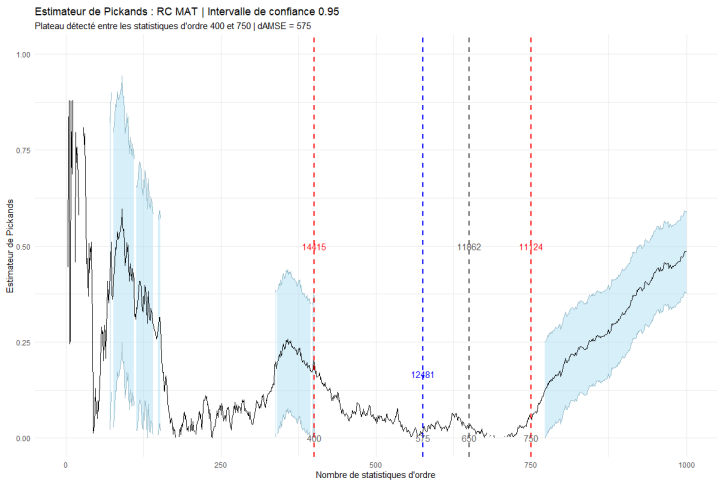


FIGURE B.8 – Évolution de l'estimateur de Pickands selon k_n

B.2 RCC

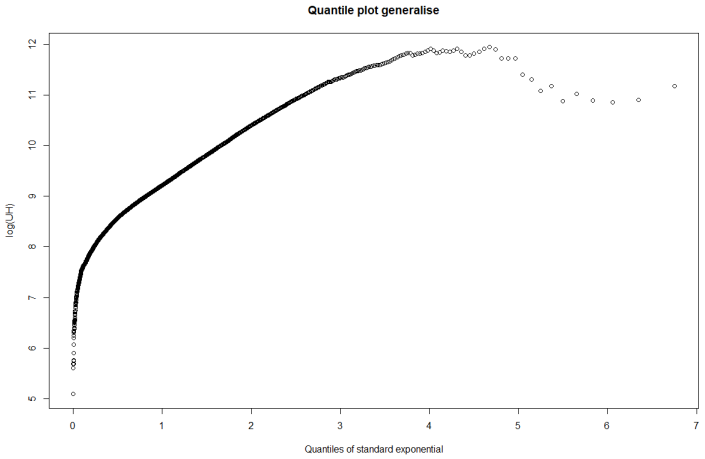


FIGURE B.9 – Quantile plot généralisé de la RCC

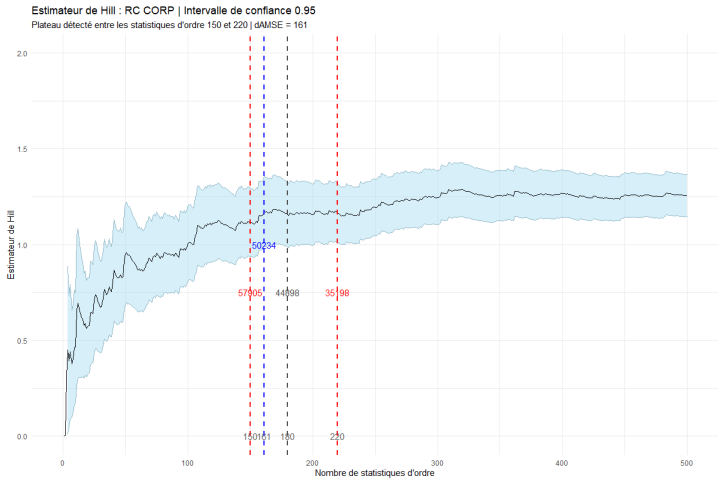


FIGURE B.10 – Évolution de l'estimateur de Hill selon k_n

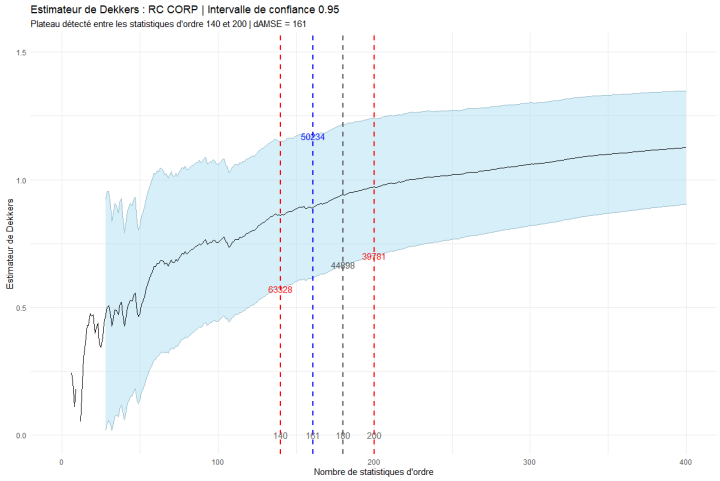


FIGURE B.11 – Évolution de l'estimateur de DEdH selon k_n

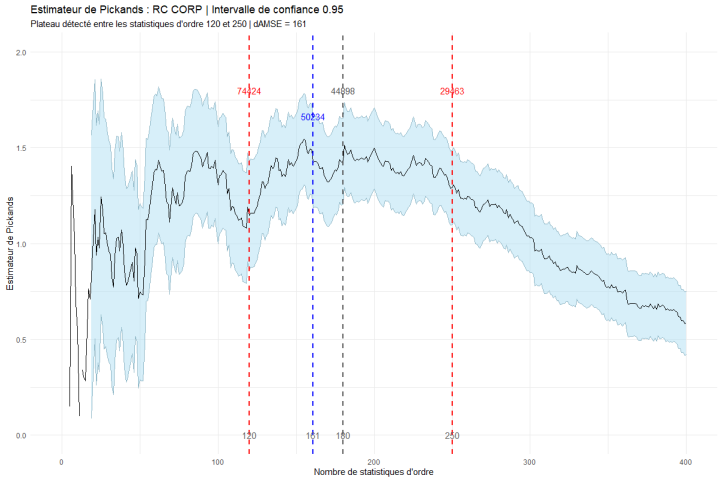


FIGURE B.12 – Évolution de l'estimateur de Pickands selon k_n

C Méthodes de regroupement

C.1 Véhiculier

FIGURE C.13 – Influence de la classe de prix sur l'AFDM

C.2 zonier

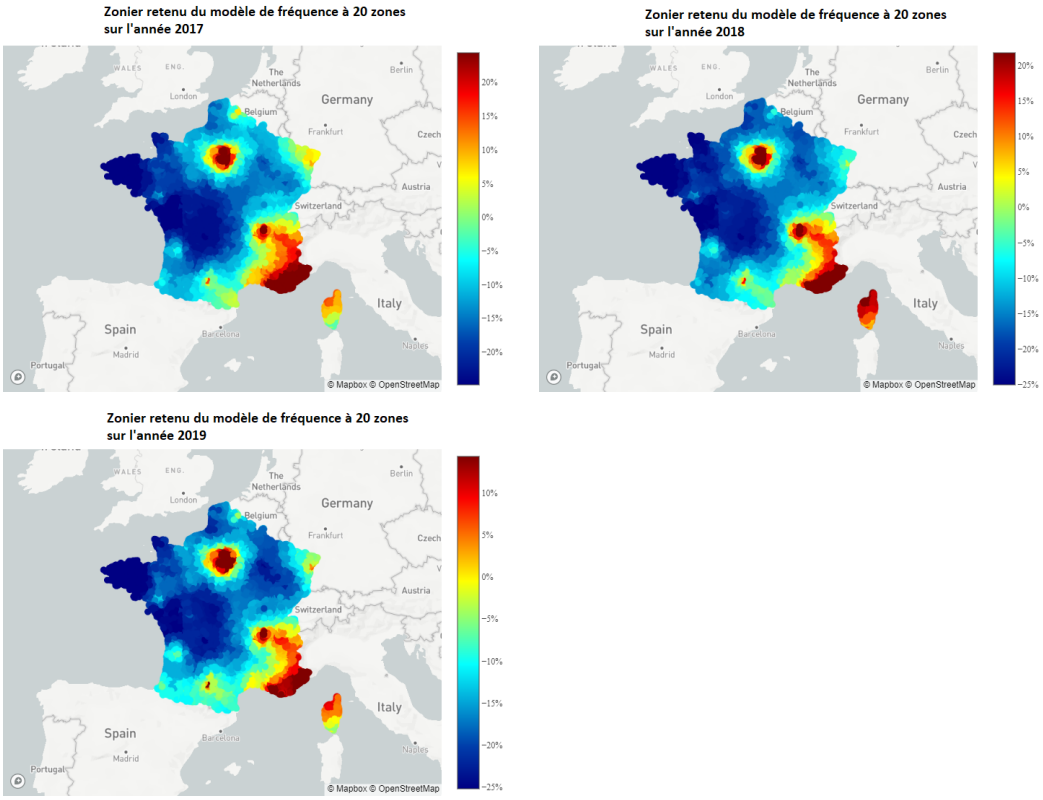


FIGURE C.14 – *Time consistency* du zonier obtenant les meilleures performances

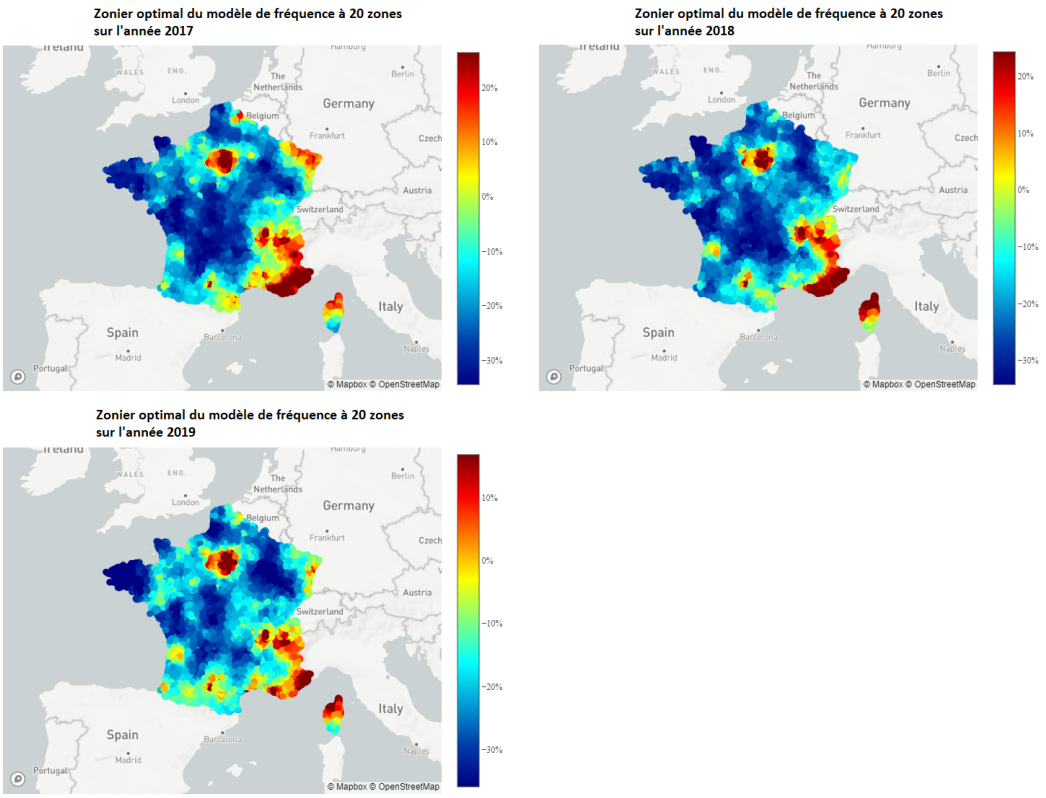
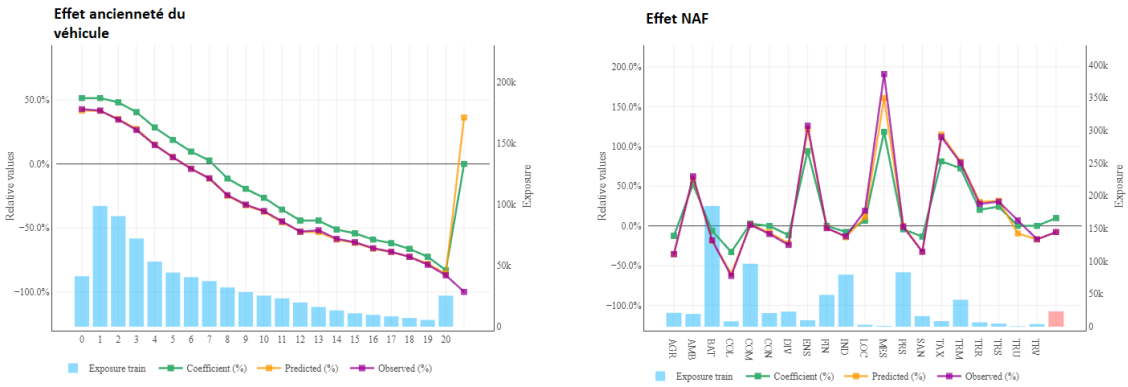


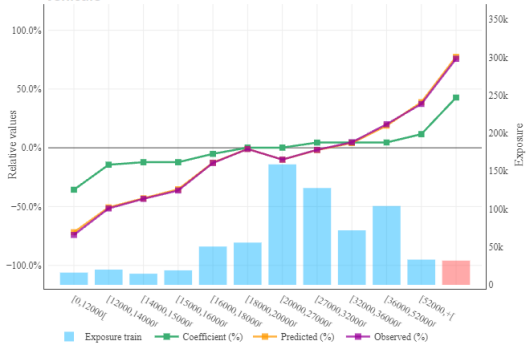
FIGURE C.15 – *Time consistency* du zonier retenu

D Statistiques univariés des variables retenus

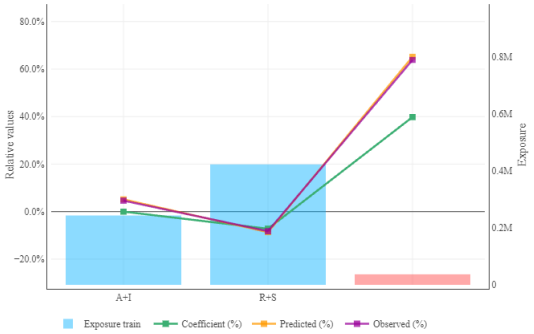
D.1 Modèle de fréquence



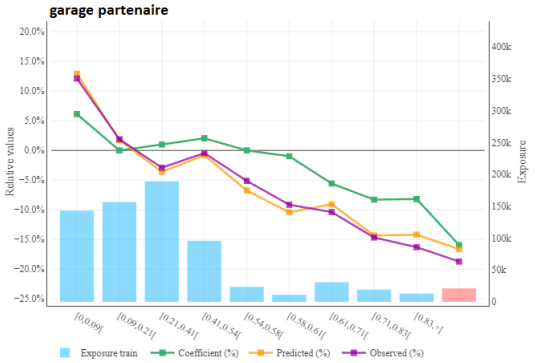
Effet prix à l'origine du véhicule



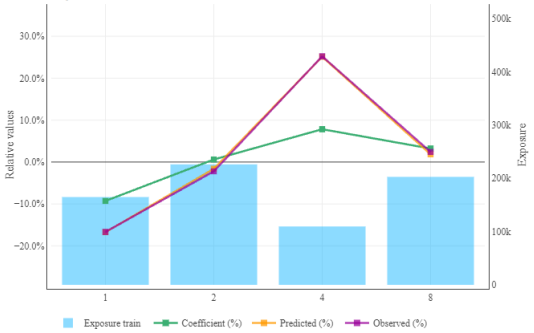
Effet suspension du véhicule



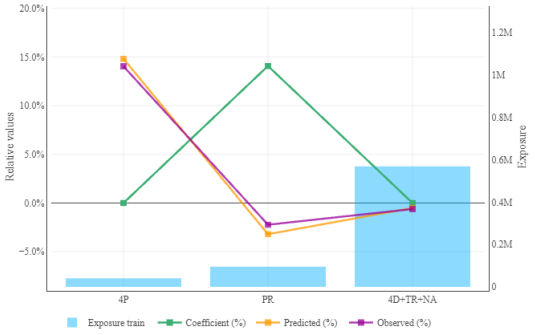
Effet taux d'orientation garage partenaire



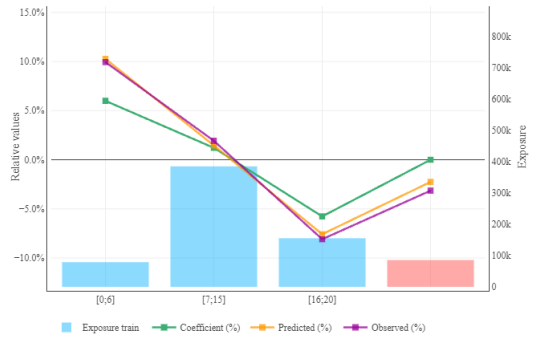
Effet fractionnement de la prime



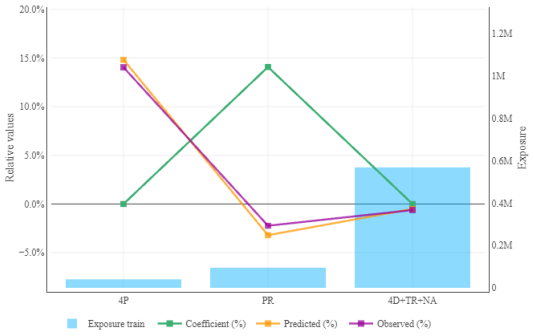
Effet transmission du véhicule



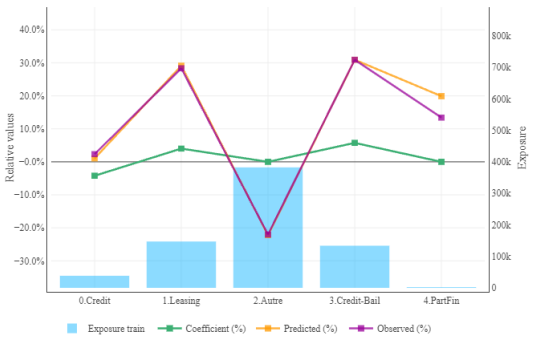
Effet note financière



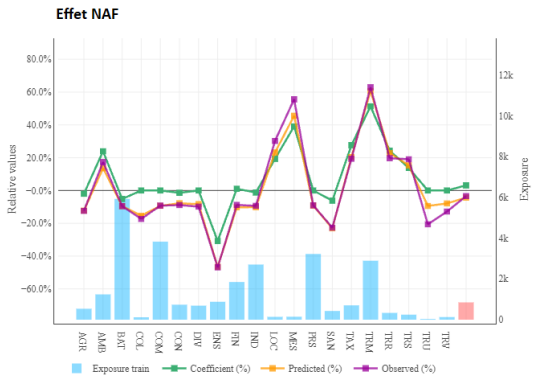
Effet transmission du véhicule

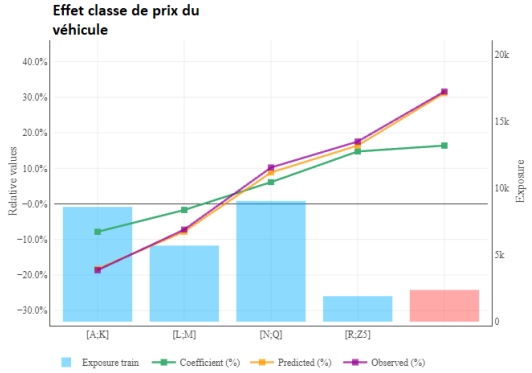


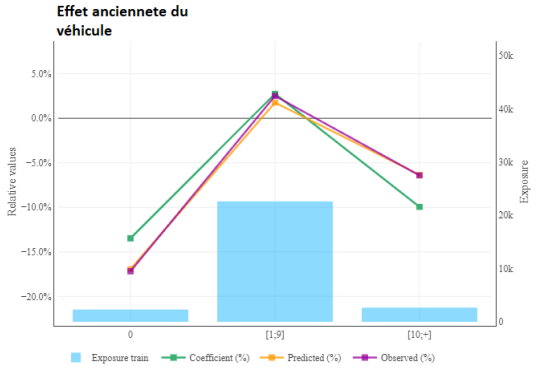
Effet mode de financement du véhicule

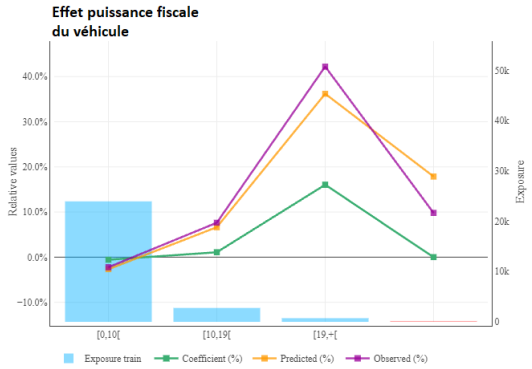


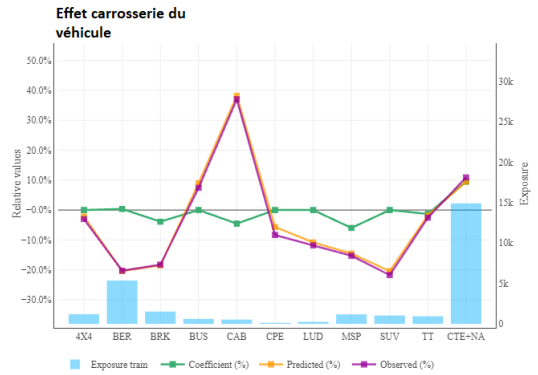
D.2 Modèle de coût moyen

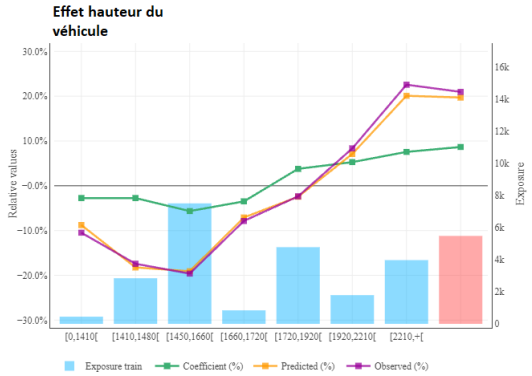


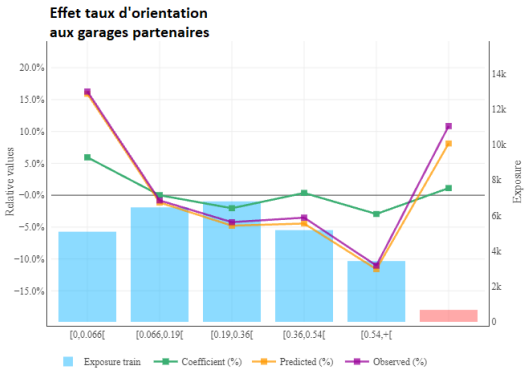


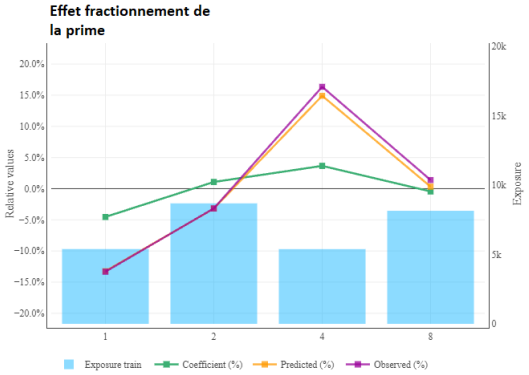









Effet taux d'orientation aux garages partenaires
Relative values
Exposure
20.0%
15.0%
10.0%
5.0%
0.0%
−5.0%
−10.0%
−15.0%
14k
12k
10k
8k
6k
4k
2k
0
[0,0.066[
[0.066,0.19[
[0.19,0.36[
[0.36,0.54[
[0.54,+[
Exposure train
Coefficient (%)
Predicted (%)
Observed (%)


Effet fractionnement de la prime
Relative values
Exposure
20.0%
15.0%
10.0%
5.0%
0.0%
−5.0%
−10.0%
−15.0%
−20.0%
20k
15k
10k
5k
0
1
2
4
8
Exposure train
Coefficient (%)
Predicted (%)
Observed (%)