

**Mémoire présenté devant l'Institut du Risk Management
pour la validation du cursus à la Formation d'Actuaire
de l'Institut du Risk Management
et l'admission à l'Institut des actuaires
le**

Par : Christopher CHAPUIS

Titre : Modélisation de la prime pure RC omnicanale en assurance automobile

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

Membres présents du jury de l'Institut des
actuaires :

Entreprise : AVIVA ASSURANCES

Nom : Anne-Sophie VERSCHAVE

Signature et Cachet :

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

Directeur de mémoire en entreprise :

Nom : Brice CARLES

Signature :

Invité :

Nom :

Signature :

**Autorisation de publication et de mise en
ligne sur un site de diffusion de documents
actuariels**

(après expiration de l'éventuelle confidentialité)

abeille
ASSURANCES

ABEILLE IARD & SANTE
Siège social : 13 rue du Moulin Bailly
92270 Bois-Colombes
Tél. 01 76 62 50 00

Signature du responsable entreprise

Société Anonyme d'Assurances Incendie Accidents et
Risques Divers
Entreprise régie par le Code des assurances
au capital de 176 771 906,38 €
306 522 655 RCS Nanterre

Signature(s) du candidat(s)

Secrétariat :

Bibliothèque :

Résumé

Face à une pression de la concurrence toujours plus forte et à la transformation digitale à l'origine des nouvelles attentes des clients de bénéficier d'une expérience client la plus complète, les différents acteurs du marché sont amenés à revoir leurs stratégies en proposant des offres plus modulables et un service client omnicanal.

Ce mémoire s'inscrit dans le cadre de la nouvelle stratégie « Client Unique » d'Aviva France destinée à renforcer son attractivité sur le marché des particuliers. Il propose, à travers l'étude de la tarification de la garantie responsabilité civile automobile, un nouveau modèle de prime pure capable de s'ajuster sur tous les canaux de distribution.

L'approche présentée dans cette étude consiste à unifier l'expérience du digital et des agents généraux au sein d'une même base de données étendue et à modéliser séparément la fréquence et le coût moyen des sinistres. Cette modélisation s'appuie sur la méthode traditionnelle des modèles linéaires généralisés (GLM) en distinguant les composantes matérielles et corporelles. Ce mémoire aborde également la problématique de l'influence de la zone géographique dans la tarification avec l'élaboration d'un zonier permettant de couvrir la totalité du territoire.

Mots clés : Tarification, Assurance automobile, Responsabilité civile, Omnicanalité, Modèles linéaires généralisés (GLM), Théorie des valeurs extrêmes, Zonier.

Abstract

The increasing competitive pressure and the digital transformation made the customer demand more difficult to satisfy in order to have the best purchaser experience. This is the reason why the various market insurers had to review their strategies by offering more flexible deals and propose an omnichannel environments to better serve their clients.

The purpose of this note is to introduce Aviva's new strategy « Unique customer » intended to strengthen the attractiveness of the compagnie on the individual market and allow a better evaluation of the risks to which the company is exposed through the study of the pricing of automobile civil liability cover. This new premium model is capable of adjusting to all distribution channels.

We will focus on this study on this new approach which consists in unifying the experience of digital and general agents within the same extended database and in processing separately the frequency and the average cost of claims. For this we will use the generalized linear models method (GLM) widely applied in insurance by distinguishing each of the two material and bodily injury claims components. This thesis also addresses the influence of the geographical area in pricing with the development of a zoning system to cover the entire territory.

Key words : Pricing, Auto Insurance, Liability, Omnichannel, Generalized Linear Models (GLM), Extreme Value Theory, Zoning.

Note de synthèse

Principal marché des assurances de biens et de responsabilités et produit d'appel par excellence, l'assurance automobile des particuliers est incontournable dans le paysage assurantiel français. Fortement concurrentiel, le marché de l'assurance automobile constitue de ce fait un véritable enjeu stratégique pour bon nombre d'assureurs souhaitant attirer de nouveaux clients. Par ailleurs, le marché de l'assurance doit également s'adapter à la transformation digitale qui impose de nouveaux standards en matière de personnalisation des offres et de relation avec le client.

C'est dans un tel contexte qu'Aviva France a décidé de refondre fondamentalement son offre d'assurance automobile à destination des particuliers en regroupant l'ensemble des offres existantes sous une bannière unique. Ce projet s'inscrit dans la stratégie omnicanale « **Client Unique** » d'Aviva France destinée à renforcer son attractivité sur le marché des particuliers par une convergence de ses offres en assurance dommages sous une marque unique accessible via tous les canaux de distribution (web, agence, mobile, call center).

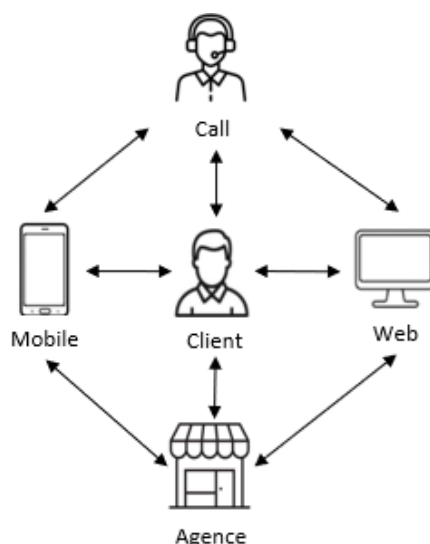


Figure 1 – La relation client dans une stratégie omnicanale

L'objectif de ce mémoire est, à travers l'étude de la tarification de la garantie responsabilité civile automobile, de proposer un modèle de prime pure qui s'ajusterait sur chacun des deux canaux de distribution (direct et intermédié). Cette nouvelle modélisation vise également une simplification du modèle opérationnel par une réconciliation des deux offres existantes de la gamme auto possédant chacune leur propre tarification avec d'une part *Eurofil by Aviva*, produit du canal direct et de l'autre *Vectura*, produit des agents généraux. Concernant le choix de la garantie RC, il n'est pas anodin car au-delà d'être la seule garantie obligatoire (article L.211-1 du Code des assurances), il s'agit d'une garantie très complexe et difficile à appréhender du fait de la présence de sinistres corporels rares et très coûteux à l'origine de la volatilité des résultats.

La première étape de cette étude a consisté à unifier les deux portefeuilles existants au sein d'une unique base de données étendue et fiabilisée afin de disposer de la vision la plus complète du risque de la branche auto. La nécessité de détenir une telle base de données a été préalablement établie par

une analyse comparative des portefeuilles mettant en évidence des divergences structurelles importantes et reflétant des différences d'appétit de souscription, de niveau de couverture des garanties ou encore d'exposition géographique. Concernant la profondeur d'historique retenue, elle a été fixée à 6 ans (exercices 2013-2018) afin de garantir un volume de données suffisant dont les ¾ proviennent de l'expérience des intermédies (produit *Vectura*), ce qui est fidèle à la pondération initiale des deux portefeuilles.

Le datamart omnicanal a ensuite fait l'objet de divers retraitements des données et notamment d'un écrêtement des valeurs extrêmes dont le seuil adéquat a pu être déterminé par différentes méthodes issues de la théorie des valeurs extrêmes (fonction moyenne des excès, estimateur de Hill et fonction GPD). Cette étude a été réalisée sur chacune des composantes de la garantie RC (matérielle et corporelle) et a permis d'écarter un très faible nombre de sinistres représentant une part importante de la charge.

	Matérielle	Corporelle
Seuil d'écrêtement retenu	50 000 €	1 500 000 €
Nombre de sinistres > seuil	27	22
Poids de la charge en sucète	1,6%	27,9%

Table 1 – Résultat de l'écrêtement par composante

Cette étape préliminaire des données s'est conclue par une analyse univariée du portefeuille d'étude qui constitue un premier outil de détection des variables les plus influentes ainsi qu'une étude des corrélations afin de quantifier les liaisons entre les différentes variables explicatives. Cette dernière analyse est importante à mener en amont de la modélisation car, dans le cadre d'une régression, une forte multicollinéarité peut poser problème lors de l'estimation des paramètres les rendant instables et difficilement interprétables. Pour cela, deux méthodes complémentaires ont été utilisées, la première par une approche classique qui utilise la statistique du V de Cramer et la seconde moins courante par une classification de variables qui fournit un ensemble de classes de variables homogènes. Suite à l'étude du dendrogramme, une classification à 9 classes de variables a été retenue, ce qui facilite grandement la lisibilité des corrélations entre les différentes variables.

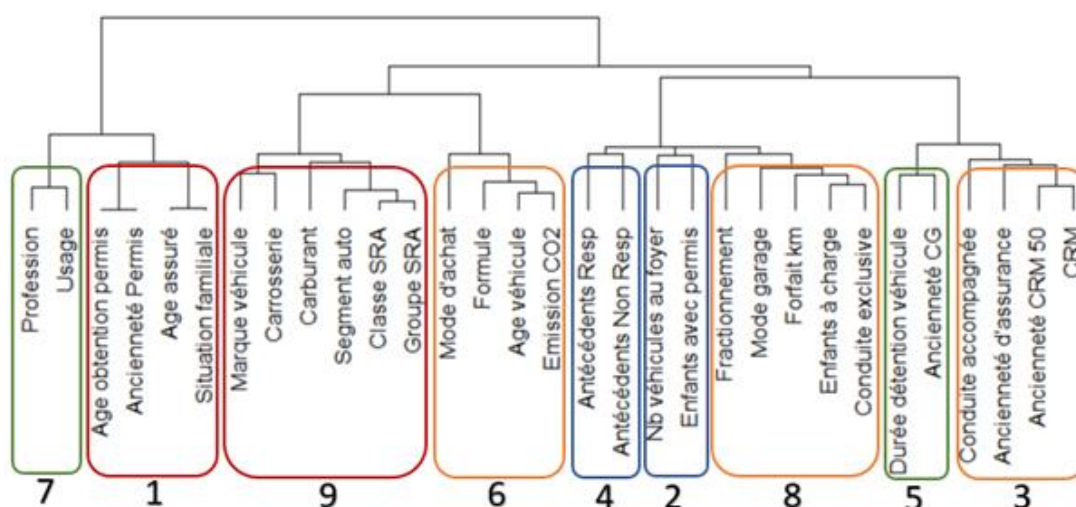


Figure 2 – Dendrogramme de la classification de variables

La suite de l'étude s'est portée sur la modélisation de la prime pure de la garantie RC selon une approche classique consistant à modéliser séparément la fréquence et le coût moyen des sinistres à l'aide des modèles linéaires généralisés (GLM). Cette méthode a été privilégiée à raison des nombreux avantages qu'elle offre tels qu'une excellente lisibilité du tarif, une performance reconnue et éprouvée ou encore une structure multiplicative facilitant sa mise en œuvre. Afin de tenir compte des spécificités de chaque nature de risques, les composantes matérielles et corporelles ont également été traitées séparément avec une modélisation de la partie attritionnelle des sinistres, c'est-à-dire celle en-dessous du seuil d'écrêtement estimé pour chaque composante et ceci afin d'éviter que les sinistres extrêmes ne biaisent les modèles de coût moyen. La charge résiduelle a quant à elle été mutualisée entre les différents profils de risque et suivant chaque composante du risque.

Pour chaque sous-modèle considéré, les principales étapes de la modélisation sont résumées ci-dessous :

1. **Détermination de la loi théorique la plus appropriée** : Loi de Poisson pour les modèles de fréquence et loi Gamma pour les modèles de coût moyen.
2. **Prise en compte des contraintes** : Offset de l'exposition et du CRM (contrainte réglementaire) dans les modèles de fréquence.
3. **Sélection des variables tarifaires et détection des interactions** : Mise en place d'une procédure couplant deux méthodes différentes (Stepwise et LASSO) et recherche du meilleur sous-ensemble de variables par une sélection par le vote.
4. **Retraitements des variables** : Réduction de la complexité des modèles par un regroupement des modalités non significatives et par l'utilisation de splines pour les variables de type ordinal permettant de gommer les fluctuations trop brutales entre deux modalités consécutives.
5. **Vérification de la stabilité temporelle des variables du modèle** : Etape qui consiste à vérifier la stabilité des estimateurs sur chaque exercice de survenance et pour chaque variable.
6. **Validation du modèle** : Déploiement d'une batterie d'analyses visant à s'assurer de la pertinence, du pouvoir de discrimination et de la robustesse du modèle (test de significativité du modèle, qualité d'ajustement global, indice de Gini, courbe de lift, étude des résidus).

Par ailleurs, la composante matérielle, dont la distribution des coûts moyens des sinistres responsables est fortement impactée par la présence de sinistres à coût forfaitaire (forfaits IDA) indemnisés suivant la convention IRSA, a fait l'objet d'une modélisation suivant deux approches. La première approche dite « exhaustive » qui croise le mode d'indemnisation (IDA / non IDA) et le niveau de responsabilité (responsable / non responsable) et la seconde approche dite « simplifiée » qui ne distingue que le niveau de responsabilité. Et bien que l'approche exhaustive ait permis une meilleure adéquation des hypothèses du modèle (meilleur ajustement de la distribution des sinistres à une loi théorique), les deux approches ont finalement abouti à la même segmentation tarifaire avec un avantage sur le critère de la performance pour l'approche simplifiée qui a donc été retenue pour modéliser la composante matérielle.

Au final, le modèle de prime pure de la garantie RC est obtenu en combinant 3 sous-modèles de prime pure :

$$\text{Modèle de PP RC} = \left(PP_{mat}^{resp} \times (1 + \alpha) + PP_{mat}^{non\ resp} \times (1 + \beta) + PP_{corp} \times (1 + \gamma) \right) \times (1 + \delta)$$

Avec :

- $PP = Fréq \times CM$ pour chaque sous-modèle, soit un total de 6 sous-modèles déployés.
- α, β, γ le coefficient de redressement de chaque sous-modèle égal au rapport de la charge totale observée sur la charge attritionnelle modélisée.
- δ le coefficient d'ajustement global égal au rapport de la charge totale observée sur la charge totale prédite. Ce coefficient est calculé sur l'ensemble des données et a pour but de corriger les écarts générés par la combinaison des sous-modèles. Sa valeur est inférieure à 1%, ce qui indique une bonne calibration du modèle de prime pure.

Le modèle de prime pure a été complétée par l'intégration d'un nouveau zonier élaborée partir des résidus de fréquence matérielle et suivant la méthode de zonage par lissage spatial qui a été privilégiée pour sa capacité à s'affranchir des problèmes d'exposition et à réduire les variations entre zones voisines. Ce zonier omnicanal établi à partir de l'expérience complémentaire des deux portefeuilles existants s'est avéré particulièrement pertinent et d'une importance majeure dans la tarification comme en témoigne le classement des variables les plus influentes donné ci-dessous.

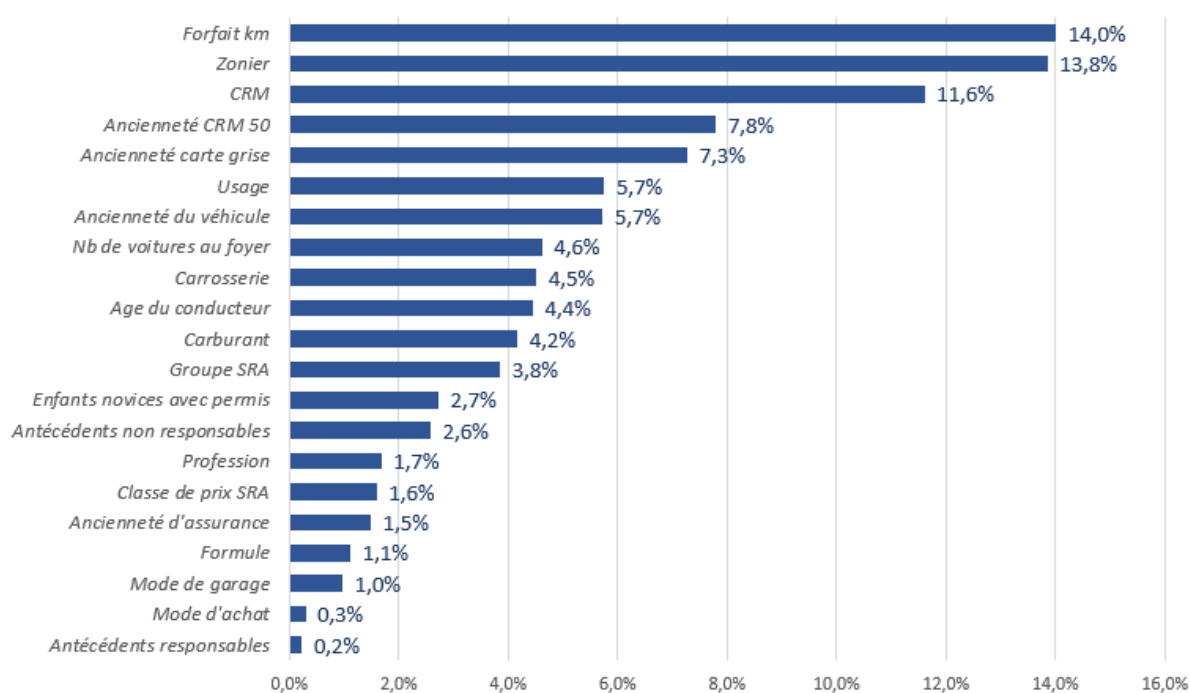


Figure 3 – Contributions des variables au modèle de prime pure

La dernière étape de ce mémoire s'est attachée à l'étude de la performance du modèle de prime pure et à sa pertinence selon le canal de distribution afin de juger de sa capacité à s'ajuster sur chacun des deux principaux canaux. Ces différentes analyses ont été menées au global et par critère tarifaire en se basant sur plusieurs indicateurs synthétiques (ratio charge observée sur charge prédite, RMSE, MAE, indice de Gini). L'analyse des résultats a montré une bonne qualité prédictive du modèle et en particulier sur la composante matérielle avec un ratio charge observée sur charge prédite à l'équilibre

sur les deux jeux de données (apprentissage et test) validant au passage le choix de l'approche simplifiée. L'ajustement sur la composante corporelle est apparu plus délicat à obtenir et de qualité plus modeste mais tout de même très acceptable compte-tenu de la nature très volatile de la sinistralité corporelle.

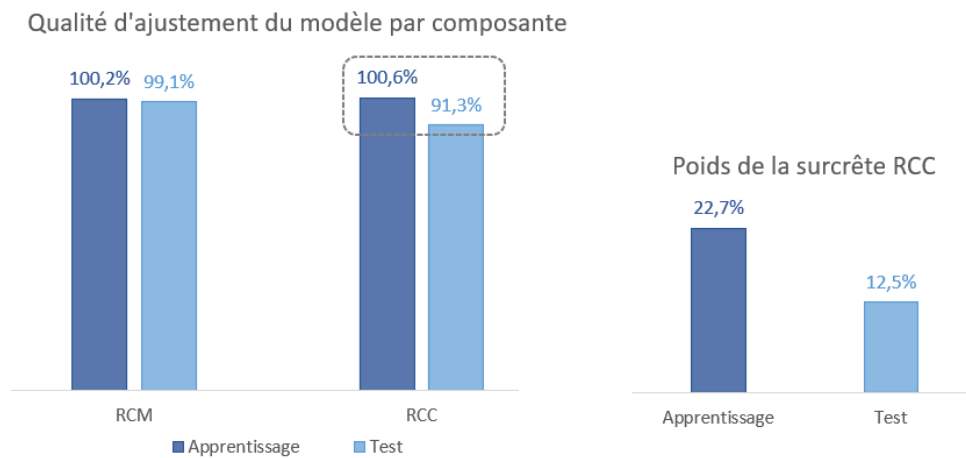


Figure 4 – Performance du modèle de prime pure par composante du risque

Enfin, concernant l'objectif premier de ce mémoire qui était de proposer une nouvelle modélisation de la garantie RC capable de s'ajuster sur chacun des modes de distribution, il a été pleinement atteint et a permis de gommer l'erreur d'ajustement sur le canal direct qui avait été constatée avec le modèle existant (surestimation de 16% avec le modèle existant vs 1,1% avec le nouveau modèle).

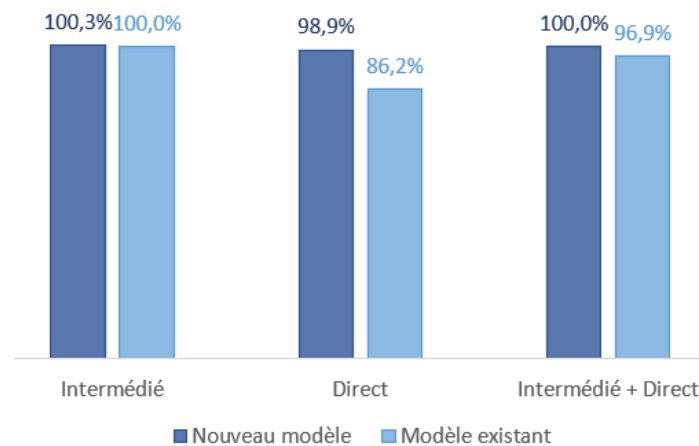


Figure 5 – Qualité d'ajustement des modèles selon le canal de distribution

Executive summary

The individual car insurance is the main insurance market for property and casualty in France. Highly competitive, it is a real strategic challenge for many insurers wishing to attract new customers. Therefore, the insurance companies must adapt to the digital transformation which imposes new standards in terms of offer's personalization of and customer's relation.

It is in such a context that Aviva France has decided to fundamentally overhaul its car insurance offer for individuals by unifying the existing products under a single banner. This project is part of Aviva's new strategy « Unique customer » intended to strengthen the attractiveness of the compagnie on the individual market by converging its property and casualty insurance deals under a single brand accessible via all distribution channels (web, agency, mobile, call center).

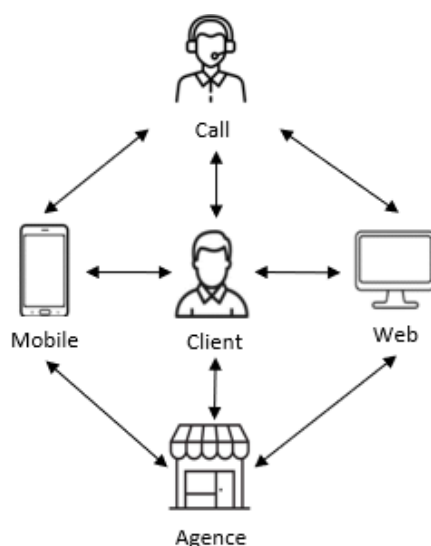


Figure 1 – The customer relationship in an omnichannel strategy

The purpose of this note is to introduce, through the study of the pricing of automobile civil liability cover, the pure premium model that would fit on each of the two distribution channels. This new modeling process aims to simplify the current model by reconciling the Eurofil by Aviva (product of the direct) and Vectura (product of general agents) and offer the same price on each of them. Choosing civil liability is not insignificant because beyond being the only compulsory guarantee (article L.211-1 of the Insurance Code), it is a very complex and difficult guarantee to apprehend due to the presence of rare and very costly bodily claims, which are the source of the volatility of the results.

In order to reach an excellent technical pricing, it is important to have good data and it is necessary to control every aspect which is involved in the process. In accordance to this statement, the first stage of this study consisted in unifying the two existing portfolios within a single extended and reliable database intending to have the most complete view of the automobile risk. The need to hold such a database was previously established by a comparative analysis of the portfolios reflecting major structural differences and highlighting various divergence in the subscription methods, guarantee's coverage and geographic exposure. The historical depth retained for our study was set at 6 years (fiscal

years 2013-2018) in order to guarantee a sufficient volume of data, $\frac{3}{4}$ of which come from the experience of intermediaries (Vectura product) and confirms the initial distribution of the portfolio.

Once the database has been restructured, it is important to choose the right threshold level for identifying large losses. Thus, we relied on the theory of extreme values to determine to adequate capping level (Excess function, Hill estimator, GPD function). This study was carried out on each of the components of the RC guarantee (material and bodily) and made it possible to rule out a very small number of claims representing a significant part of the expense.

	Matérielle	Corporelle
Seuil d'écrêtement retenu	50 000 €	1 500 000 €
Nombre de sinistres > seuil	27	22
Poids de la charge en sucrête	1,6%	27,9%

Board 1 – Result of the capping by component

Furthermore, we started running additional checks and format the variables on the final merged car database. This last part is called the preliminary analysis. In fact, it is important to detect the most influential variables as well as a study of correlations in order to quantify the links between the different factors. This last analysis is substantial to carry out before of the modeling stage because the possible correlation between the explanatory variables can pose a problem estimating the parameters, making them unstable and difficult to interpret. Two complementary methods were used. The first one is Cramer's V statistic classic approach. The second method is less common which consist in classifying the rating factors and providing a set of homogeneous's classes variables. In fact, the study of the dendrogram shows 9 groups of variables retained for modeling, which greatly facilitates the readability of the correlations between these explanatory variables.

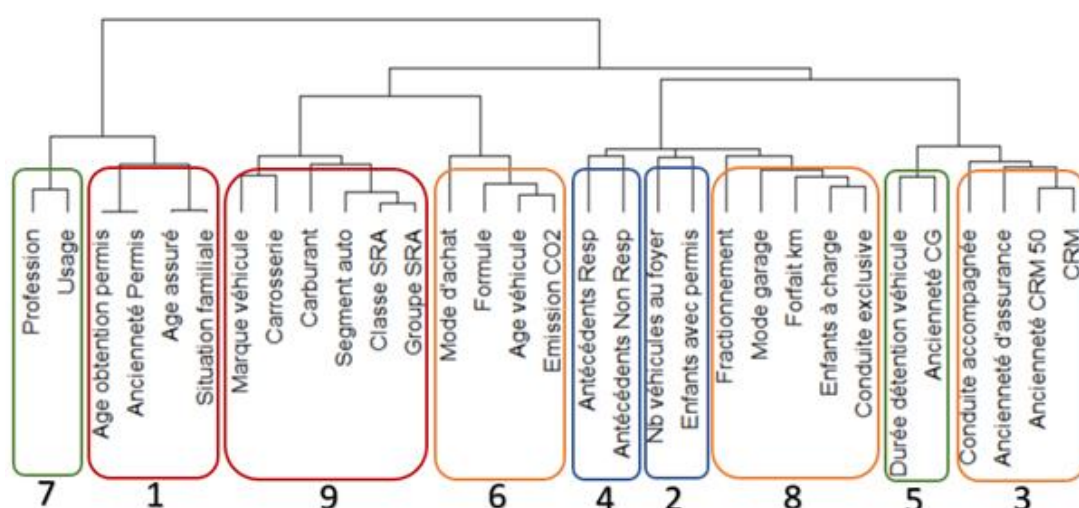


Figure 2 – Variable classification Dendrogram

The second part of the study focused on the pricing of the RC guarantee's pure premium by separately modeling the frequency and average cost of claims using the classical method of generalized linear model (GLM). This method has been favored because of the many advantages it offers such as excellent pricing readability, good model performance and a multiplicative structure facilitating its

implementation. In order to take into account the specificities of each type of risk, we have treated separately material and bodily components with modeling of the attritional part of claims, i.e. the claim load below the capping threshold defined above in order to avoid lowering the pricing of the average costs. The residual charge was distributed between the different risk profiles according to each risk component.

For each sub-model, the following list summarizes the main modeling steps :

1. **Determination of the most appropriate theoretical distribution:** Poisson distribution for frequency models and Gamma distribution for average cost models.
2. **Taking restrictions into account:** offset of exposure and CRM (regulatory constraint) in frequency models.
3. **Selection of pricing variables and detection of interactions:** Using two different methods (stepwise and Lasso regression) and finding the best subset's variables by voting selection.
4. **Reprocessing of variables:** reduction of the complexity of the models by grouping non-significant modalities and smoothing spline for ordinal-type variables allowing erase the sudden fluctuations between two consecutive modalities.
5. **Verification of the temporal stability of the variables of the model:** Consists in verifying the stability of the estimators on each occurrence exercise and for each rating factor.
6. **Validation of the model:** Performing analyzes aimed at ensuring the relevance, the power of discrimination and the robustness of the model (significance test of the model, overall quality of adjustment, Gini index, lift curve, study of residuals).

The material component, whose distribution of the average costs of responsible claims is strongly impacted by the presence of fixed cost claims (IDA packages) compensated according to the IRSA agreement, was modeled using two approaches. The first so-called "exhaustive" approach which crosses the compensation method (IDA / non-IDA) with the claim's responsibility level. The second so-called "simplified" approach only distinguishes the claim's responsibility level. Although the exhaustive approach allowed a better adequacy of the models hypotheses (better adjustment of the distribution of claims to a theoretical distributions), the two approaches finally resulted in the same pricing. This is why we retained the simplified approach for modeling the material component.

The pure premium model of the RC guarantee is obtained by combining 3 pure premium sub-models :

$$PP\ RC\ Model = \left(PP_{mat}^{resp} \times (1 + \alpha) + PP_{mat}^{non\ resp} \times (1 + \beta) + PP_{corp} \times (1 + \gamma) \right) \times (1 + \delta)$$

Caption :

- $PP = Frequency \times Average\ cost\ of\ claims$ for each sub-model, i.e. a total of 6 deployed sub-models.
- α, β, γ the adjustment parameters of each sub-model equal to the ratio of the total observed load per the modeled attritional load.

- δ the global adjustment parameter equal to the ratio of the total load observed on the total load predicted. This factor is calculated on all the data and is intended to correct the differences generated by the combination of the sub-models. Its value is less than 1%, which indicates a good calibration of the pure premium model.

The pure premium model has been completed with the integration of a new zoning system, developed from material frequency residuals and following the spatial smoothing zoning method, in order to overcome exposure problems and reduce variations between neighboring areas. This omnichannel territorial classification was established from the complementary experience of the two existing portfolios. In effect, zoning analysis has been proven to be particularly relevant and a major importance in pricing, as evidenced by the ranking of the most influential variables given below.

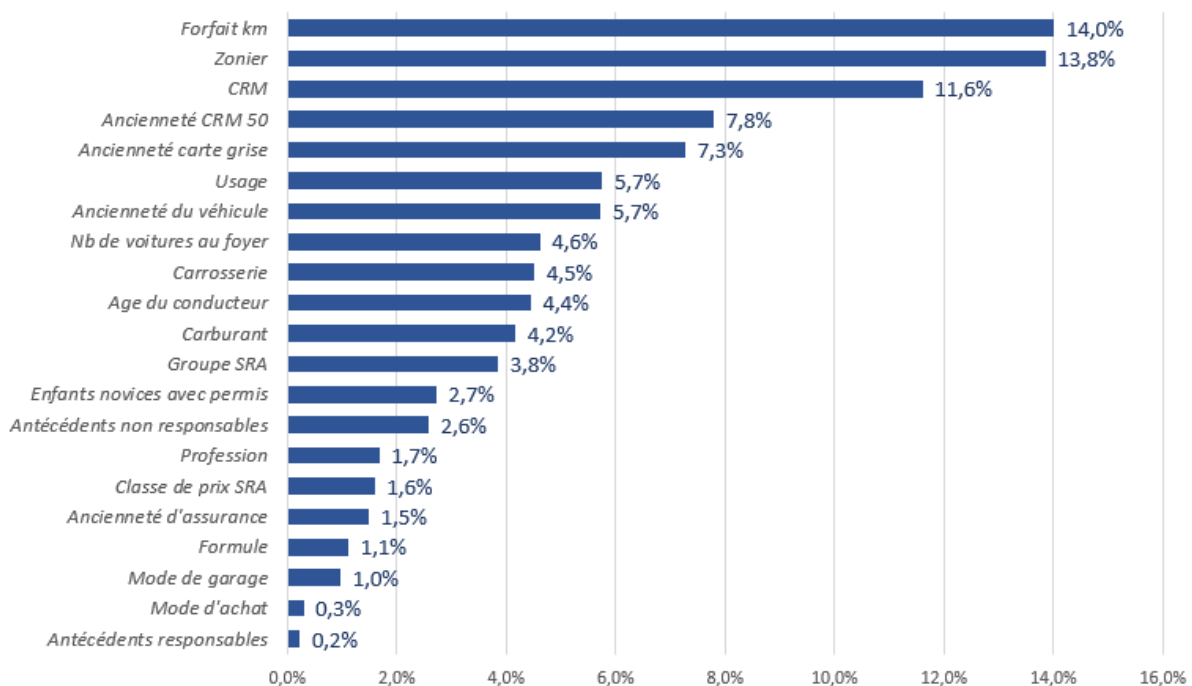


Figure 3 – Features importance of the pure premium model

The last stage of this thesis focused on studying the performance of the pure premium model and its relevance in order to judge its ability to adjust to each of the two main channels. These various studies were based on several synthetic indicators (ratio of observed load per predicted load, RMSE, MAE, Gini index). The analysis of the results showed a good predictive quality of the model on the material component with a good ratio of observed load per predicted load on the two sets of data (learning and test) which confirms the choice of the simplified approach. The adjustment on the bodily component appeared to be more difficult to obtain and of more modest quality, but still very acceptable given the highly risk's volatile nature.

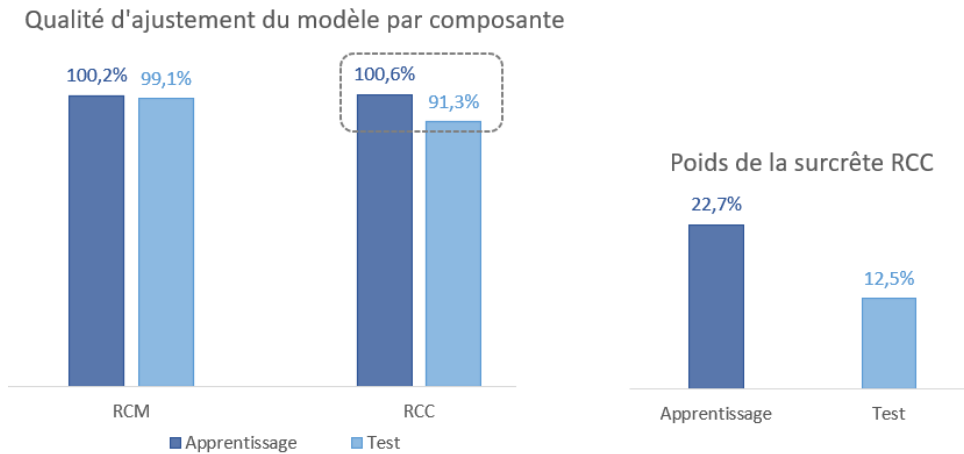


Figure 4 – Performance of the pure premium model by risk component

Finally, concerning the primary objective of this project, which was to propose a new modeling of the RC guarantee capable of adjusting to each of the distribution modes, it has been fully achieved and has made it possible to erase the adjustment error on the direct channel which had been observed with the existing model (overestimation of 16% with the existing model vs 1.1% with the new model).

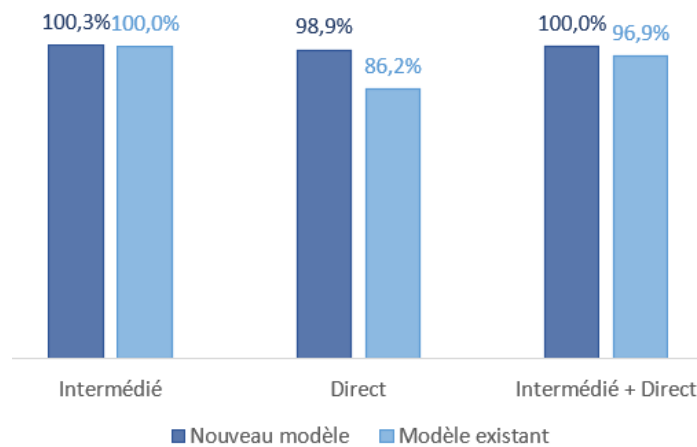


Figure 5 – Adjustment of the models according to the distribution channel

Remerciements

Je tiens tout d'abord à remercier très chaleureusement Fabienne GOURINCHAS, directrice de la Data Factory, qui m'a permis d'intégrer la formation du Centre d'Etudes Actuarielles.

Un grand merci également à Romain LAILY, consultant chez Sia Partners, pour sa disponibilité et ses réponses à mes nombreuses questions.

Mes remerciements vont également à Brice CARLES pour son accompagnement, ses conseils avisés et sa relecture attentive du mémoire.

Enfin, je souhaite remercier tout mon entourage et ma famille pour leur soutien et leurs encouragements tout au long de la rédaction de ce mémoire.

Table des matières

Résumé	3
Abstract	4
Note de synthèse	5
Executive summary	10
Remerciements	15
Introduction	19
Contexte et objectif du mémoire	21
1.1 Présentation du contexte	21
1.1.1 Le marché de l'assurance automobile.....	21
1.1.2 Structure du portefeuille du marché	21
1.1.3 Les nouvelles tendances du marché automobile.....	22
1.1.4 Cadre de l'étude : Le projet « Client Unique »	23
1.2 L'offre d'assurance auto d'Aviva Assurances.....	24
1.2.1 Présentation des produits existants.....	24
1.2.2 Structure des portefeuilles existants	25
1.3 Le nouveau produit Auto Unique et ses attentes	28
1.3.1 Principes de conception	28
1.3.2 Les attentes du produit Auto Unique.....	28
1.4 Motivations et objectif du mémoire	29
1.4.1 Motivations du mémoire.....	29
1.4.2 Objectif du mémoire	30
Analyse préliminaire des données	31
2.1 Constitution de la base de modélisation	31
2.1.1 Définition du périmètre.....	31
2.1.2 Exclusion de périmètre.....	32
2.1.3 Validation et réconciliation des données	32
2.1.4 Présentation des données	32
2.2 Retraitement des données	34
2.2.1 Gestion de l'inflation.....	34
2.2.2 Les honoraires d'expert.....	34
2.2.3 Ecrêtement des valeurs extrêmes.....	34
2.3 Analyse exploratoire des données	38
2.3.1 Résultats généraux.....	38
2.3.2 Décomposition de la sinistralité	39
2.3.3 Analyse univariée	43
2.4 Etude des corrélations entre variables explicatives.....	46
2.4.1 Méthode du V de Cramer.....	46
2.4.2 Classification automatique de variables.....	47

Modélisation de la prime pure	53
3.1 Motivation de l'approche	53
3.1.1 Le cadre usuel de la tarification IARD	53
3.1.2 Tests d'indépendance	55
3.2 Rappels théoriques.....	57
3.2.1 Les modèles linéaires généralisés.....	57
3.2.2 Estimation des paramètres du modèle	58
3.2.3 Significativité des paramètres du modèle	59
3.2.4 Qualité d'ajustement et validation du modèle.....	59
3.2.5 Critères de sélection de modèles.....	60
3.2.6 Stratégie de sélection de variables	62
3.3 Modélisation de la fréquence	62
3.3.1 Détermination de la loi.....	63
3.3.2 Prise en compte des contraintes	64
3.3.3 Sélection de variables.....	64
3.3.4 Retraitements des variables dans les modèles.....	66
3.3.5 Stabilité des variables dans le temps	67
3.3.6 Elaboration du zonier	68
3.3.7 Résultats du modèle.....	70
3.4 Modélisation du coût moyen	75
3.4.1 Détermination de la loi.....	75
3.4.2 Présentation du modèle	76
3.4.3 Performance du modèle.....	77
3.4.4 Validation du modèle	79
3.5 Modèle de prime pure.....	81
Analyse des résultats et comparaison avec l'existant.....	83
4.1 Ajustement du modèle de prime pure	83
4.1.1 Les antécédents de sinistres à la souscription.....	83
4.1.2 La clause de conduite exclusive	84
4.2 Analyse du modèle de prime pure	84
4.2.1 Contributions des variables au modèle.....	84
4.2.2 Pouvoir de segmentation des variables tarifaires.....	86
4.2.3 Distribution de la prime pure	87
4.3 Performance du modèle de prime pure	90
4.3.1 Qualité prédictive du modèle	90
4.3.2 Analyse des résultats par critère tarifaire	91
4.4 Analyse comparative des modèles.....	94
4.4.1 Comparaison des structures tarifaires	95
4.4.2 Comparaison par critères tarifaires	96
4.4.3 Comparaison des zoniers.....	97
4.4.4 Comparaison des distributions de prime pure	99
4.5 Performance des modèles selon le canal de distribution.....	101
4.5.1 Qualité d'ajustement global des modèles.....	101
4.5.2 Autres indicateurs de performance : MAE, RMSE et indice de Gini	102
4.5.3 Analyse des résultats par critère et selon le canal de distribution	103

Conclusion	109
Bibliographie.....	111
Annexes	112

Introduction

Incontournable dans le paysage assurantiel français et produit d'appel par excellence, l'assurance automobile occupe une place particulièrement importante sur le marché des assurances de biens et de responsabilité des particuliers. Face à une pression concurrentielle toujours plus forte et aux nouvelles attentes des clients de bénéficier d'une expérience client omnicanale, les différents acteurs du marché sont amenés à revoir leurs stratégies et à repenser leur manière d'interagir avec leurs clients.

C'est dans ce contexte de transformation digitale du secteur de l'assurance qu'Aviva France a décidé de refondre fondamentalement son offre d'assurance automobile en regroupant l'ensemble de ses offres existantes sous une bannière unique et une distribution omnicanale de son nouveau produit. Mais pour être en mesure proposer le même tarif unique quel que soit le canal de distribution sollicité, il est primordial que le modèle de prime pure sous-jacent soit parfaitement ajusté sur chacun d'eux, ce qui n'est pas chose aisée tant les populations assurées diffèrent à bien des égards.

Aussi, l'objectif de ce mémoire est, à travers l'étude de la tarification de la garantie responsabilité civile automobile, de proposer un modèle de prime pure omnicanal qui s'ajusterait sur chacun des deux principaux canaux de distribution (direct et intermédié). Par ailleurs, le choix de la garantie RC n'est pas anodin car au-delà d'être la seule garantie obligatoire (article L.211-1 du Code des assurances), il s'agit d'une garantie très complexe et difficile à appréhender du fait de la présence de sinistres corporels rares et très coûteux à l'origine de la volatilité des résultats

Ce mémoire se compose de quatre grands chapitres. Le premier chapitre débutera par une présentation synthétique du marché de l'assurance automobile et des nouvelles tendances du secteur automobile qui n'a cessé d'évoluer pour s'adapter aux besoins des consommateurs et aux nouvelles réglementations environnementales. Ce chapitre sera également l'occasion de préciser la vision stratégique d'Aviva France sur le marché des particuliers et de présenter la gamme des produits existants ainsi que les grandes lignes du nouveau produit Auto Unique qui a vocation à être distribué sur tous les canaux.

Le second chapitre sera consacré à l'analyse préliminaire des données, étape importante de l'étude qui permet de s'approprier les données. Il constitue un premier outil de détection des variables les plus influentes. Ce chapitre abordera donc différents thèmes centrés autour de la donnée tels que la constitution de la base de données, les retraitements des données ou encore l'analyse descriptive du portefeuille d'étude qui a la particularité d'être constitué à partir de l'expérience de deux produits très différents ayant chacun leur propre modèle opérationnel et leur mode de distribution dédié.

Le troisième chapitre portera sur la modélisation de la prime pure de la garantie RC par une méthode largement éprouvée dans le secteur de l'assurance, à savoir les modèles linéaires généralisés (GLM) qui reste aujourd'hui encore la plus utilisée dans ce domaine. Afin de tenir compte des spécificités de chaque nature de risques, les composantes matérielles et corporelles seront traitées séparément selon une approche classique fréquence-coût moyen. Facteur considérablement influent dans la tarification, la zone géographique fera l'objet d'une attention toute particulière avec l'élaboration d'un zonier permettant de couvrir la totalité du territoire.

Pour finir, Le quatrième chapitre portera sur l'analyse des résultats et une comparaison avec le modèle existant sera réalisée. Une étude de la performance du modèle de prime pure sera menée en s'appuyant sur plusieurs indicateurs synthétiques qui permettront d'avoir un avis objectif sur les forces et faiblesses du nouveau modèle. Enfin, la pertinence du nouveau modèle selon le canal de distribution sera vérifiée par une évaluation de la qualité d'ajustement au global et par segment, ce qui permettra de juger de la capacité du modèle à tenir la promesse de l'omnicanalité et d'un tarif unique quel que soit le canal sollicité.

Chapitre 1

Contexte et objectif du mémoire

1.1 Présentation du contexte

1.1.1 Le marché de l'assurance automobile

Avec un chiffre d'affaires s'élevant à 21,1 milliards d'euros en 2020 et un parc de véhicules assurés de 43,5 millions d'après le rapport annuel de la fédération française des assurances (FFA), l'assurance automobile est le principal marché des assurances de biens et de responsabilité des particuliers. Fortement concurrentielle avec près de 100 acteurs qui y opèrent, l'assurance automobile n'est pourtant pas une activité rentable pour la plupart des acteurs du marché comme en témoigne le ratio combiné net de réassurance de ces 10 dernières années qui affiche un niveau supérieur ou égale à 100% (L'année 2020 étant particulièrement atypique du fait de la crise sanitaire du Covid-19 marquée par la chute du trafic routier durant les périodes de confinement engendrant une baisse considérable du nombre d'accidents de la route).

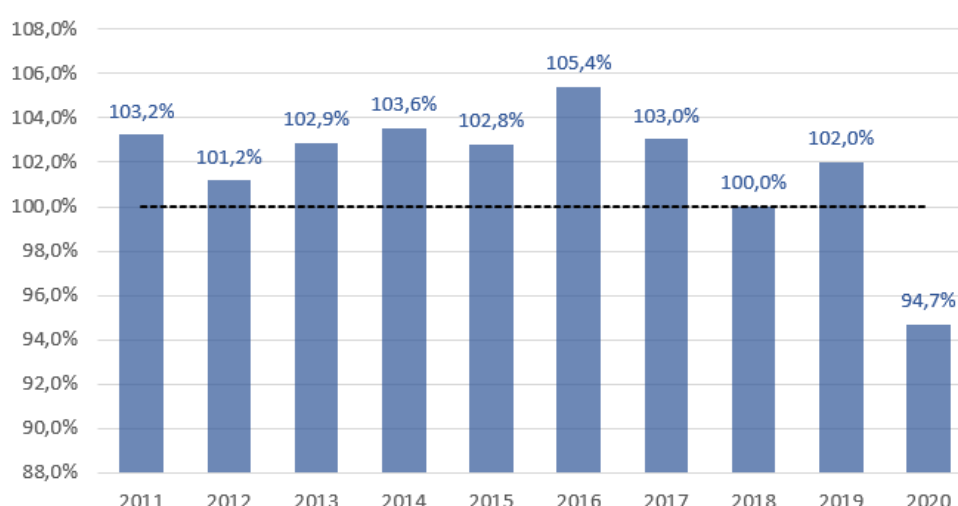


Figure 1.1 – Ratio combiné net de réassurance en automobile (FFA, 2021)

1.1.2 Structure du portefeuille du marché

Toujours selon le rapport annuel de 2020 de la FFA, le parc automobile français se caractérise par un âge moyen de 10,9 ans avec une part importante de véhicules anciens de 15 ans et plus (26,7%). Du côté du niveau de couverture, la formule tous risques est de loin la plus plébiscitée dans 65,3% des cas contre 13,4% pour la formule minimale. Concernant le type d'énergie, le diesel continue de marquer le pas mais reste toujours majoritaire représentant 59,4% des carburants suivi de l'essence (36,6%). Quant au mix selon le coefficient de réduction-majoration (CRM) résumant l'expérience de sinistralité de chaque assuré, il est très favorable avec 96,5% d'assurés avec un bonus (CRM < 100) dont 70,4% avec un bonus max (CRM 50).

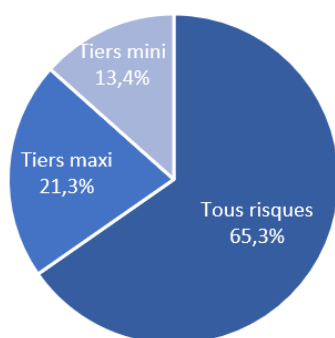


Figure 1.2 – Répartition selon la formule (FFA, 2021)

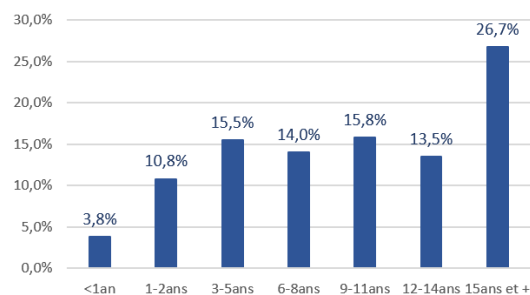


Figure 1.3 – Répartition selon l'ancienneté du véhicule (FFA, 2021)

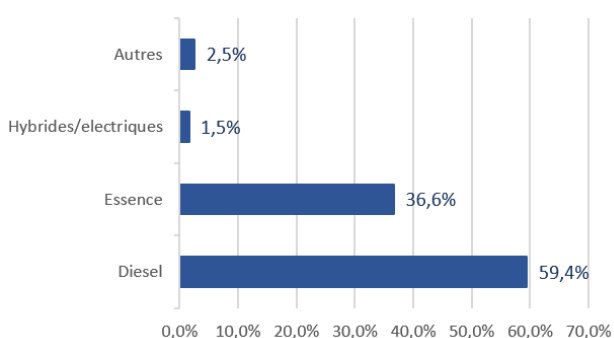


Figure 1.4 – Répartition selon le type d'énergie (FFA, 2021)

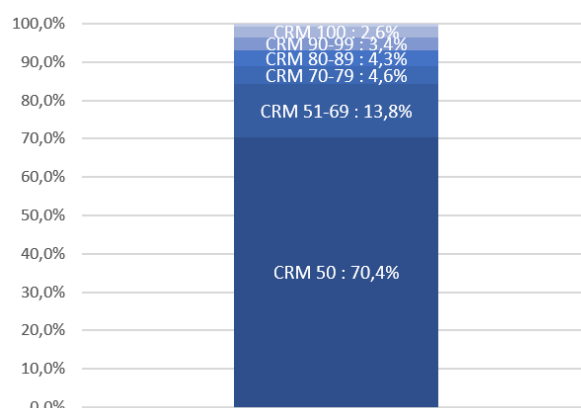


Figure 1.5 – Répartition selon le CRM (FFA, 2021)

1.1.3 Les nouvelles tendances du marché automobile

Le secteur de l'automobile connaît depuis maintenant quelques années des transformations profondes sous la pression de l'impératif environnemental, des évolutions technologiques et des changements de comportement des consommateurs.

Parmi les grandes tendances du secteur automobile :

- **L'essor des véhicules hybrides et électriques** : Largement encouragé par les nouvelles réglementations visant à réduire les émissions de CO₂, le système de bonus-malus écologique ou encore la planification de la fin des moteurs thermiques d'ici 2035, les ventes de véhicules hybrides et électriques ont récemment explosé représentant 21,5% des immatriculations de voitures neuves en 2020 contre 7,6% en 2019 (CCFA/AAA DATA, 2021). Et cette transformation vers un mode de déplacement plus écologique devrait très certainement s'intensifier dans les années à venir.

Type d'énergie	2019		2020		Evolution
	Unités	%	Unités	%	
Essence	1 281 798	57,9%	774 363	46,9%	-39,6%
Diesel	755 579	34,1%	504 191	30,6%	-33,3%
Hybride	125 436	5,7%	243 651	14,8%	94,2%
Electrique	42 764	1,9%	110 912	6,7%	159,4%
Autre	8 702	0,4%	16 965	1,0%	95,0%
Total	2 214 279	100,0%	1 650 082	100,0%	-25,5%

Table 1.1 – Ventes de voitures neuves selon le type d'énergie (CCFA/AAA DATA, 2021)

- **Le succès de la Location avec Option d'Achat (LOA) / Location Longue Durée (LLD)** : Modes de financement très répandues chez les professionnels, la LOA et la LLD sont de plus en plus plébiscitées par les particuliers. En 2020, d'après une étude du cabinet Autoways, la LOA/LLD représente 45% des immatriculations de véhicules neufs pour les particuliers contre 38% en 2019. Cette nouvelle tendance s'explique notamment par une évolution de la consommation avec un recul progressif de la propriété individuelle en faveur d'un usage plus partagé de la mobilité (Covoiturage, autopartage, ...).
- **La voiture connectée / autonome** : Si la voiture connectée est déjà une réalité (de plus en plus de véhicules neufs sont équipés de systèmes de télécommunications), la voiture autonome n'arrivera quant à elle que très progressivement sur le marché, le temps de faire évoluer la réglementation et les comportements. Les bénéfices attendus de ce nouveau genre de véhicules sont multiples avec par exemple une amélioration du confort et de la sécurité pour le consommateur (optimisation des temps de trajets, réduction de la consommation, réduction du nombre d'accidents et de leur gravité, ...).

1.1.4 Cadre de l'étude : Le projet « Client Unique »

Acteur majeur de l'assurance en France avec plus de 3 millions de clients, Aviva France propose ses produits IARD (Incendie, Accidents et Risques Divers) à travers une approche multicanale via son réseau de distribution implanté sur tout le territoire (1 000 agents généraux et 860 agences) et sa marque d'assurance en ligne *Eurofil by Aviva* et ses 185 conseillers téléphoniques.

Afin de renforcer son attractivité sur le marché des particuliers et de séduire de nouveaux clients, Aviva France a mis en place une nouvelle stratégie visant à regrouper l'ensemble de ses offres en assurance dommages sous une marque unique : **Le projet « Client Unique »**. Un modèle novateur dans son approche du marché des particuliers qui consiste à converger vers une seule proposition permettant de « *proposer les mêmes offres, offrir le même niveau de service, aux mêmes tarifs quel que soit le canal de distribution* » comme l'explique Patrick Dixneuf, Directeur Général d'Aviva France. Cette nouvelle stratégie omnicanale répond au besoin du client d'être partout à la fois (web, agence, applications mobiles, ...) en le positionnant véritablement au centre de la relation client.

Après la récente commercialisation de la nouvelle offre Aviva habitation unique (MRH Unique en 2019) constituant la première brique du programme Client Unique, la prochaine étape dans la poursuite de cette stratégie consiste en une refonte complète de l'offre automobile visant à converger vers une seule et unique solution d'assurance compatible avec une distribution omnicanale.

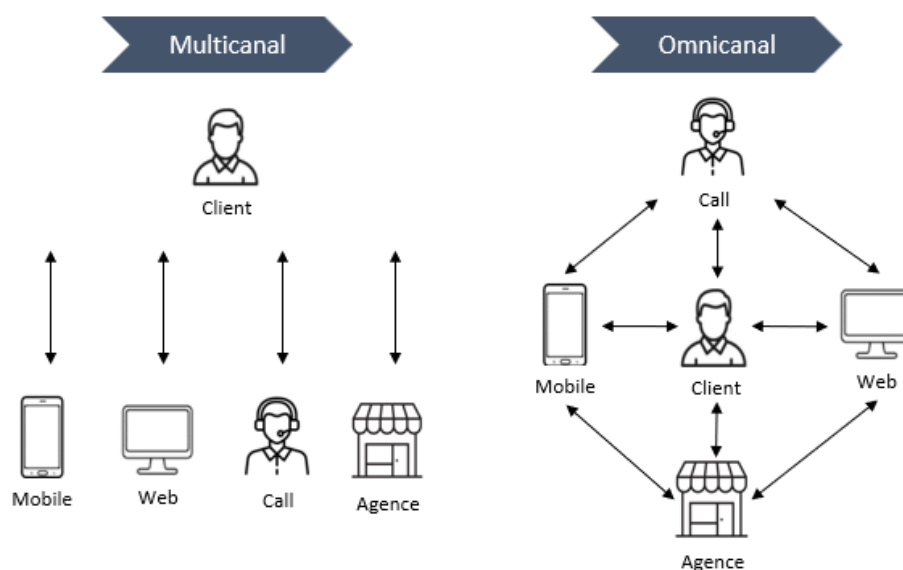


Figure 1.6 – Stratégie multicanale vs Stratégie omnicanal

1.2 L'offre d'assurance auto d'Aviva Assurances

Elle se compose de deux offres d'assurance bien distinctes avec un mode de distribution dédié, visant des segments très spécifiques de clientèle et possédant chacune leur propre modèle opérationnel (tarification, système informatique, ...).

1.2.1 Présentation des produits existants

Vectura – Produit du canal intermédié :

Distribué par le réseau des agents généraux, le produit *Vectura* vise une large gamme de clientèle de particuliers mais aussi de professionnels (artisans, commerçants et TNS) grâce à ses nombreuses options ou packs offrant la possibilité de passer d'un produit standard à du sur-mesure. Le produit *Vectura* a la particularité d'être en conduite libre de base en opposition à la conduite exclusive qui limite la conduite du véhicule aux seuls conducteurs désignés au contrat. Il se distingue également par son haut niveau de couverture des garanties socles (prise en charge des sinistres grêle via la garantie Incendie, indemnisation majorée du véhicule en tous risques, ...) et ses nombreux avantages tarifaires (réductions multi-détention, forfaits kilométriques, mois gratuits, ...).

Eurofil by Aviva – Produit du canal direct :

Réponse d'Aviva à la digitalisation et produit destiné à une clientèle de particuliers, *Eurofil by Aviva* est numéro 2 de l'assurance auto en ligne grâce à un produit extrêmement compétitif (tarif attractif centré sur les garanties essentielles) au script de souscription efficace et particulièrement adapté au digital et aux comparateurs d'assurances. Doté d'une clause de conduite exclusive de base, ce produit est également modulable grâce à une sélection d'options triées sur le volet permettant de

compléter éventuellement les garanties socles et d'un éventail de forfaits kilométriques permettant de l'adapter à tous les usages.

1.2.2 Structure des portefeuilles existants

De par leurs caractéristiques, les produits *Vectura* et *Eurofil by Aviva* présentent des divergences structurelles importantes reflétant des différences d'appétit de souscription, de niveau de couverture des garanties ou encore d'exposition géographique. Une analyse comparative sur quelques critères clés permet de rendre compte de cela.

La taille des portefeuilles

Au 31/12/2020, le portefeuille *Vectura* comptabilise 681 000 risques contre 296 000 pour *Eurofil by Aviva*, soit un volume de contrats deux fois plus important pour le canal intermédiaire qui bénéficie d'un périmètre de souscription plus large et d'une ouverture sur le marché des professionnels. Côté tarif, *Eurofil by Aviva* se démarque par un positionnement tarifaire plus agressif que *Vectura* traduisant la forte concurrence qui règne sur le canal direct et la forte sensibilité au prix de sa clientèle.

Le type de formule

Le graphique 1.7 met en évidence un mix-portefeuille très différent selon le type de formule entre les deux produits. En effet, le portefeuille *Vectura* se distingue par une part importante de contrats en formule tous risques comparativement au marché (73% pour *Vectura* contre 65% pour le marché) tandis que le portefeuille *Eurofil by Aviva* s'illustre davantage sur la formule minimale (dans 23% des cas contre seulement 13% pour le marché).

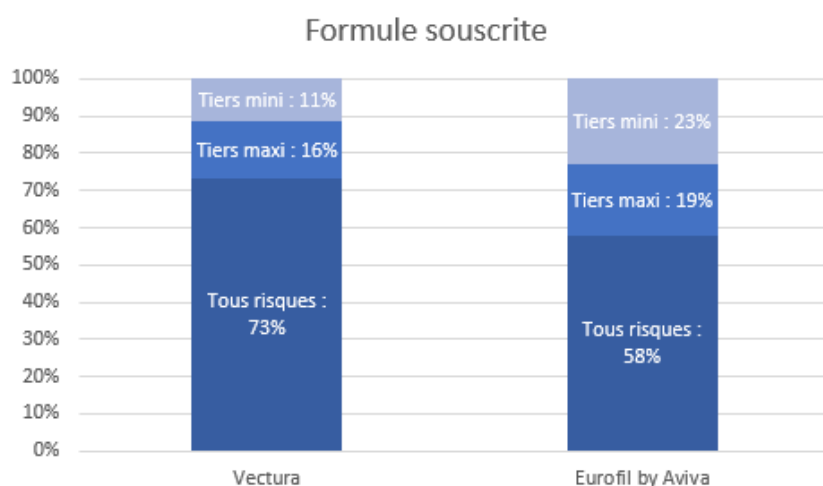


Figure 1.7 – Répartition des portefeuilles selon la formule souscrite

L'ancienneté du véhicule

Le graphique ci-dessous montre une répartition selon l'ancienneté des véhicules assurés similaire entre les deux portefeuilles et en adéquation avec le parc automobile français excepté sur les véhicules neufs légèrement surreprésentés par rapport au marché.

Le portefeuille *Vectura* se différencie néanmoins sur le segment des véhicules plus anciens (9 ans et au-delà) avec un taux encore élevé de contrats en formule tous risques.

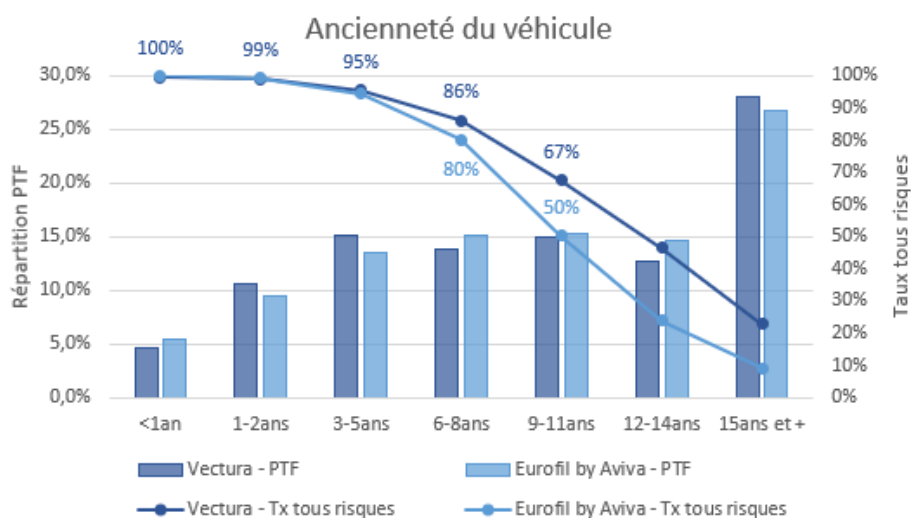


Figure 1.8 – Répartition des portefeuilles et taux de tous risques selon l'ancienneté du véhicule

Les forfaits kilométriques

Les contrats au forfait kilométrique sont largement répandus sur le canal direct, ils sont un bon moyen de diminuer la prime d'assurance. Les forfaits kilométriques concernent 88% du portefeuille *Eurofil by Aviva* dont 53% de plus de 9 000 km contre seulement 46% sur *Vectura* dont 33% inférieur ou égal à 9 000 km.

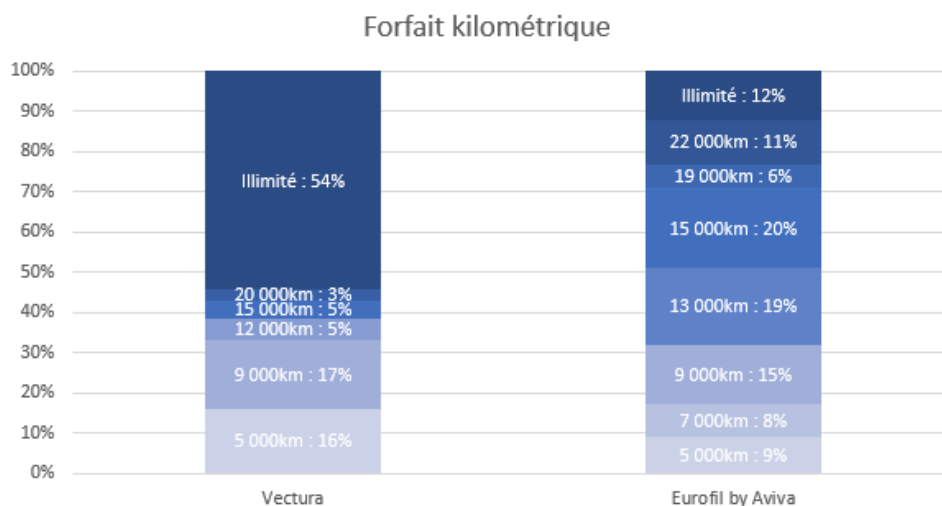


Figure 1.9 – Répartition des portefeuilles selon le forfait kilométrique

Le coefficient de Réduction-Majoration (CRM)

Le graphique 1.10 montre une répartition du portefeuille *Vectura* selon le CRM proche du marché avec malgré tout une part plus importante de très bons risques (CRM 50). Le portefeuille *Eurofil by Aviva* se

distingue davantage sur le segment des CRM 51-69 représentant 20,3% du portefeuille contre 13,8% pour le marché. A noter que l'offre d'assurance proposée par *Eurofil by Aviva* ne permet pas de souscrire les CRM 100 et plus.

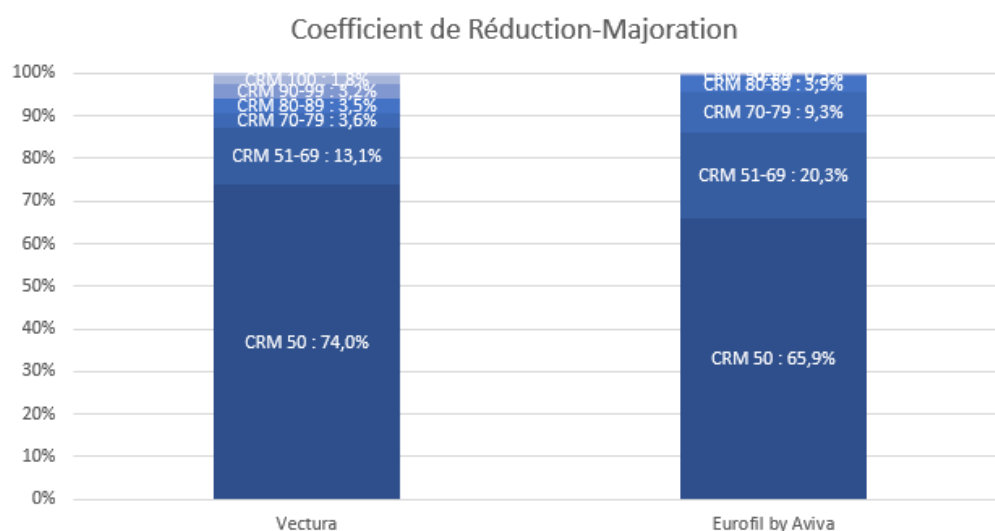


Figure 1.10 – Répartition des portefeuilles selon le CRM

L'implantation géographique

Vectura couvre une large partie du territoire avec une implantation géographique majoritairement en zone rurale et moins marquée dans les grandes agglomérations. *Eurofil by Aviva* est davantage implanté en zone urbaine et périurbaine avec une très forte concentration en banlieue représentant la moitié du portefeuille.

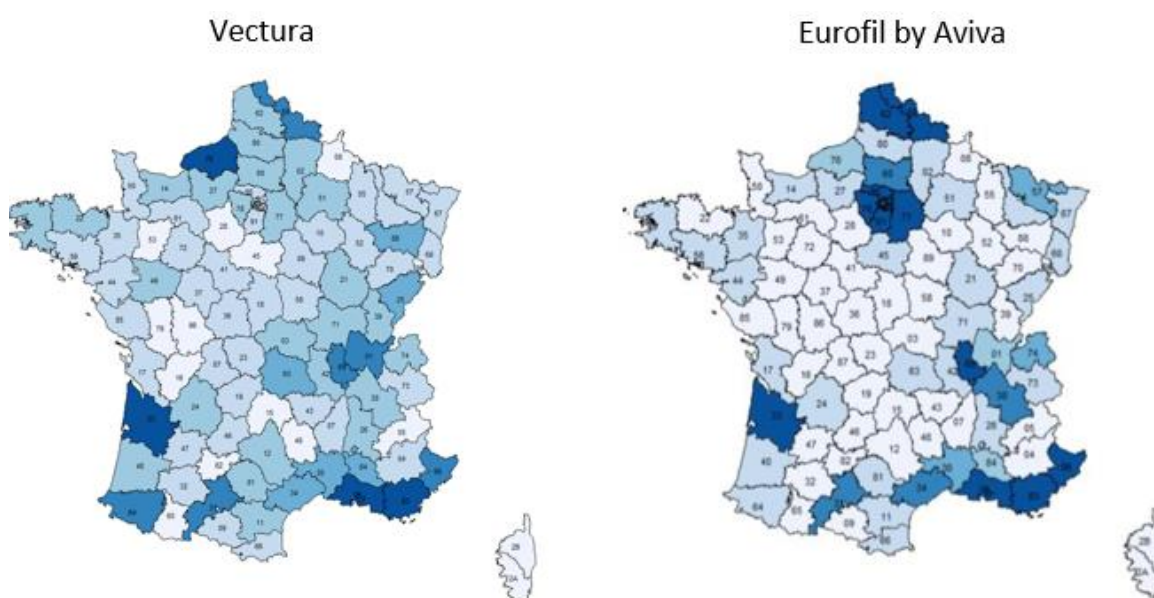


Figure 1.11 – Couverture territoriale des portefeuilles

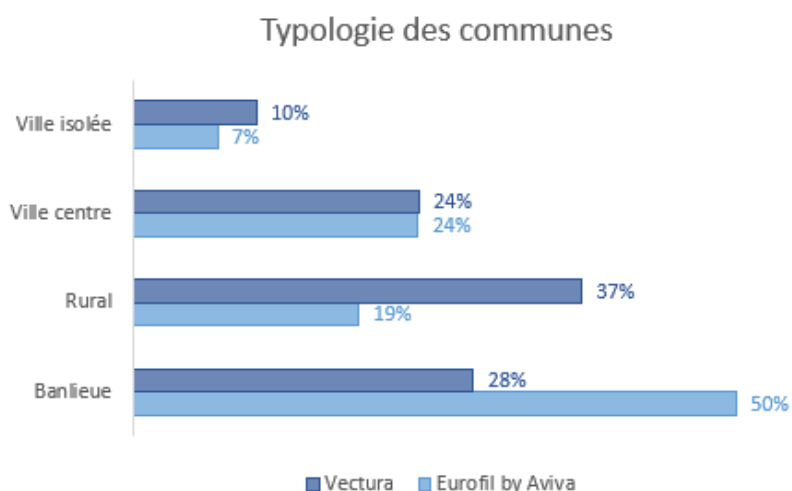


Figure 1.12 – Répartition des portefeuilles Aviva selon la typologie des communes

1.3 Le nouveau produit Auto Unique et ses attentes

1.3.1 Principes de conception

Le produit Auto Unique répond à 3 grands principes que sont la **Rentabilité** qui veille à l'équilibre technique du produit, l'**Acquisition** qui définit le contenu de l'offre, le périmètre de souscription ainsi que le positionnement tarifaire et enfin l'**Omnicanalité** avec la promesse d'un tarif unique et d'une expérience client enrichie quel que soit le canal sollicité.

Rentabilité	- Des garanties de base de qualité - Plus de garanties en option - Tarif plus technique et évolutif dans le temps
Acquisition	- Garanties solides essentielles focalisées sur les attentes des clients - Tarif attractif sur les formules de base - Modularité de l'offre afin de répondre aux besoins de toutes les typologie de clients
Omnicanalité	- Un seul produit pour tous les canaux couvrant dans un continuum l'ensemble des cas - Une expérience et un parcours clients fluides entre les canaux

Figure 1.13 – Principes de conception du produit Auto Unique

1.3.2 Les attentes du produit Auto Unique

Si la gamme de produits auto d'Aviva a su évoluer au fil des années par des ajustements tarifaires ciblés visant à corriger un déséquilibre tarifaire sur certains segments stratégiques, sa structure tarifaire est quant à elle restée inchangée et a progressivement perdu de sa cohérence au fil des années. D'autre part, la présence de deux offres distinctes a certes permis de booster les ventes grâce

à une meilleure couverture du marché mais elle a aussi engendré une confusion et une perte de distinction de la marque vis-à-vis du client ainsi que des conflits entre les canaux en termes d'offre, de services ou encore de tarif.

Aussi, les attentes du produit Auto Unique sont nombreuses :

- **Renforcement de l'attractivité de la marque Aviva** sur la branche auto en regroupant l'ensemble de ses produits sous une bannière unique.
- **Amélioration de la segmentation** par l'élaboration d'un tarif plus technique, mieux ajusté aux différents profils de risques et compatible avec une distribution omnicanale.
- **Amélioration de la compétitivité** par l'affichage d'un tarif attractif en entrée de gamme sur les formules de base.
- **Simplification du modèle opérationnel** par la convergence des outils existants et le rapprochement des équipes du Direct et de la Proximité.

1.4 Motivations et objectif du mémoire

1.4.1 Motivations du mémoire

Ce mémoire repose sur des travaux préalables qui ont permis de proposer une première tarification du produit Auto Unique. Ce modèle de prime pure, dont la conception est résumée ci-après, s'appuie principalement sur l'expérience de sinistralité du portefeuille *Vectura* complétée par la connaissance du portefeuille *Eurofil by Aviva* via des ajustements techniques ciblés afin d'améliorer sa compatibilité avec une distribution omnicanale. Cette approche avait été privilégiée en raison de l'absence d'un datamart global.

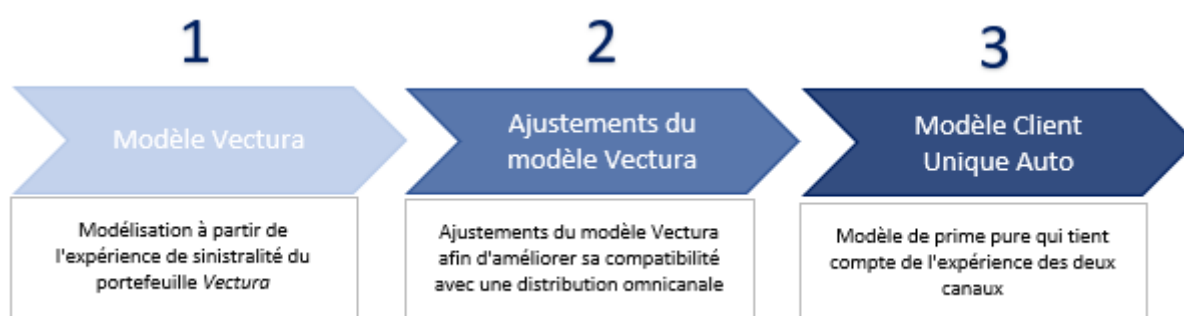


Figure 1.14 – Etapes de construction du modèle Client Unique Auto.

Ce modèle de tarification n'a cependant pas donné pleine satisfaction. En effet, des écarts significatifs ont été constatés sur le canal direct lors de la simulation du modèle rendant le modèle incompatible avec une distribution omnicanale. Ces écarts de prédiction se sont principalement concentrés autour de la garantie RC.

1.4.2 Objectif du mémoire

L'objectif principal de ce mémoire est donc de proposer une nouvelle modélisation de la prime pure de la garantie RC capable de s'ajuster sur les deux principaux canaux de distribution, à savoir le digital et le réseau des intermédiaires.

A cet effet, les grands thèmes ci-dessous seront abordés :

- Création d'un datamart global regroupant l'expérience des portefeuilles *Vectura* et *Eurofil by Aviva* permettant d'avoir une vision la plus complète du risque de la branche automobile du marché des particuliers chez Aviva Assurances.
- Analyse et modélisation séparée des composantes de la garantie RC (matérielle/corporelle)
- Création d'un zonier omnicanal couvrant l'ensemble du territoire (zones rurales et urbaines)
- Performance et pertinence du nouveau modèle selon le canal de distribution et comparaison avec l'existant

Chapitre 2

Analyse préliminaire des données

2.1 Constitution de la base de modélisation

La constitution de la base de données est le point de départ de tout projet de refonte tarifaire, c'est sur celle-ci que s'appuiera l'élaboration du modèle de prime pure et in-fine le tarif. Cette étape revêt donc une importance cruciale, la qualité et la fiabilité des données collectées conditionnent fortement la réussite du projet.

La base de données d'étude (Datamart omnicanal) est constituée à partir des données du portefeuille *Vectura* qui décrit l'expérience du canal intermédié et du portefeuille *Eurofil by Aviva* qui décrit celle du canal direct. La création de cette base commune a nécessité une étape préalable de *mappage de données* qui consiste en une transformation et une harmonisation des données issues de systèmes d'informations différents.



Figure 2.1 – Diagramme de la création de la base de données d'étude

2.1.1 Définition du périmètre

Ce mémoire s'appuie sur l'expérience des deux portefeuilles de la gamme auto du particulier d'Aviva Assurances et portent sur les exercices de 2013 à 2018 (historique de 6 ans) en considérant les contrats ayant été actifs au moins 1 jour durant cette période. Cette dernière condition est vérifiée si et seulement si l'exposition au risque, définie comme la fraction de l'année durant laquelle un assuré a été couvert pour une situation de contrat donnée, est strictement supérieure à 0.

L'étude de la sinistralité repose quant à elle sur l'ensemble des sinistres qui ont sollicité la garantie RC au cours de ces 6 exercices de survenance avec une vision de la sinistralité au 31/12/2020, hors sinistres annulés et clos sans suite.

2.1.2 Exclusion de périmètre

Dans un souci d'homogénéité, certains risques très spécifiques ont été exclus de la base d'étude. Sont concernés par ces exclusions les autocaravanes et remorques, les véhicules de collection, les voiturettes (voitures sans permis) et les véhicules au repos (véhicules non roulants).

Les risques exclus du périmètre sont négligeables, ils représentent 1,88% du portefeuille étudié et moins de 1% de la charge de sinistres de la garantie RC.

Risques exclus	Exposition	Nombre de sinistres RC	Charge de sinistres RC
Autocaravanes / Remorques	0,45%	0,15%	0,28%
Véhicules de collection	0,41%	0,45%	0,41%
Voiturettes	0,27%	0,02%	0,09%
Véhicules au repos	0,75%	0,01%	0,01%
Total	1,88%	0,63%	0,79%

Table 2.1 – Volume des exclusions de la base d'étude

2.1.3 Validation et réconciliation des données

Rappelant l'importance de fiabiliser les données utilisées, la base de données d'étude a naturellement fait l'objet d'un contrôle tout au long de son processus de construction.

La validation des données a donc été réalisée à différents niveaux :

- **Validation globale des indicateurs clés** (Volume des expositions, nombre et charge des sinistres)
- **Contrôle de cohérence des données** (Cohérence de l'évolution des indicateurs clés par modalité de variable)

La réconciliation des données qui consiste en une confrontation des statistiques calculées sur la base d'étude avec une source existante, s'est appuyée sur un ensemble de reportings réalisés par la *Data Factory*, garant de la qualité des données chez Aviva France.

Ces différents contrôles ont permis de s'assurer de la fiabilité et de la robustesse des données de la base de modélisation.

2.1.4 Présentation des données

Les données exploitables au sein de la base de données d'étude sont nombreuses et de natures variées. Elles se composent des informations de type « **Conducteur** » donnant le profil des assurés, les informations de type « **Contrat** » donnant les caractéristiques du contrat, les informations de type « **Véhicule** » fournissant des renseignements sur le bien assuré, les informations de type « **Sinistre** » retraçant l'expérience de sinistralité des assurés et enfin les informations de type « **Externe** ».

La liste exhaustive des données disponibles est détaillée ci-dessous :

Les informations « Conducteur »

- Age du conducteur principal
- Catégorie socio-professionnelle
- Situation familiale
- Ancienneté de permis de conduire
- Age d'obtention du permis de conduire
- Permis acquis dans le cadre de la conduite accompagnée (Oui / non)
- Coefficient de Réduction-Majoration (CRM)
- Nombre d'années à CRM 50
- Antécédents à la souscription (Ancienneté d'assurance, nombre de sinistres, ...)
- Nombre d'enfants à charge
- Nombre d'enfants titulaires du permis et non assurés à titre personnel
- Abonnement aux transports en commun (Oui / non)
- Nombre de véhicules au foyer
- Durée de détention du précédent véhicule
- Commune du lieu de garage

Les informations « Contrat »

- Formule souscrite
- Packs / Options
- Franchise (BDG / Dommage)
- Usage du véhicule
- Forfait kilométrique
- Conduite exclusive (Oui / non)
- Mode de garage du véhicule assuré
- Fractionnement de la cotisation (Annuel, semestriel, trimestriel, mensuel)
- Zoniers actuels
- Canal de distribution (Intermédié / Direct)

Les informations « Véhicule »

- Ancienneté du véhicule
- Ancienneté de la carte grise
- Mode d'achat (Crédit, comptant, LOA, LLD)

Les informations « Sinistres »

- Date de survenance
- Garanties sollicitées
- Nature du sinistre (Matérielle, corporelle)
- Etat du sinistre (Clos, en cours, annulé, sans suite)
- Montant du sinistre (Règlements, réserves, recours, frais d'expert)
- Niveau responsabilité (Totale, partielle, non responsable)

Les données « Externes »

- Données socio-démographiques de l'INSEE (Taux de chômage, niveau des revenus, ...)
- Informations générales sur la commune (Statut, densité de population, nombre de commerces, ...)

- Données véhicules SRA (Classe de prix, groupe, énergie, carrosserie, ...)

2.2 Retraitement des données

2.2.1 Gestion de l'inflation

Etant donné que la sinistralité étudiée s'étale sur une période assez étendue (6 exercices de survenance de 2013 à 2018), les coûts des sinistres ne sont pas directement comparables entre eux et doivent faire l'objet d'une correction par l'inflation, retraitement couramment appelé mise en « as if » des sinistres.

Pour ce faire, un redressement des coûts de la base individuelle des sinistres en amont de la modélisation est généralement réalisé mais c'est une autre méthode de gestion de l'inflation, plus simple à mettre en œuvre, qui a été privilégiée. Elle consiste à intégrer la variable liée à l'exercice de survenance directement dans les modèles afin de capter l'effet inflation et le décorréliser des autres variables tarifaires.

2.2.2 Les honoraires d'expert

Afin d'éliminer tout effet potentiel qui serait lié aux honoraires d'experts, les frais d'expert ont également été exclus de la charge des sinistres et seront réinjectés in-fine via un coefficient de chargement. Les comparaisons entre primes pures prédites et observées se feront hors honoraires d'expert.

Les honoraires d'expert représentent 1,97% de la charge totale RC.

2.2.3 Ecrêtement des valeurs extrêmes

En tarification non-vie et en particulier en modélisation de la garantie responsabilité civile automobile où le risque corporel peut engendrer des coûts considérables, il est nécessaire de segmenter les sinistres en deux catégories :

- Les sinistres dits **attritionnels** correspondant à la masse des sinistres observés.
- Les sinistres dits **extrêmes ou graves** situés en queue de distribution et caractérisés par une faible fréquence de survenance et des coûts très élevés et extrêmement volatiles.

Le tableau des quantiles des coût de la garantie RC ci-dessous montre clairement la présence de sinistres extrêmes avec un coût maximum de près de 870 k€ sur la composante matérielle et de 22 M€ sur la composante corporelle.

Niveau	Matérielle	Corporelle
100,0%	869 936 €	22 063 317 €
99,0%	7 567 €	214 000 €
98,0%	5 386 €	105 777 €
95,0%	3 376 €	33 886 €
90,0%	2 200 €	16 330 €
75,0%	1 446 €	6 142 €
50,0%	1 437 €	3 016 €
25,0%	930 €	1 480 €
10,0%	366 €	754 €
5,0%	174 €	405 €
1,0%	31 €	90 €
0,0%	1 €	1 €

Table 2.2 – Tableau des quantiles des coûts par composante

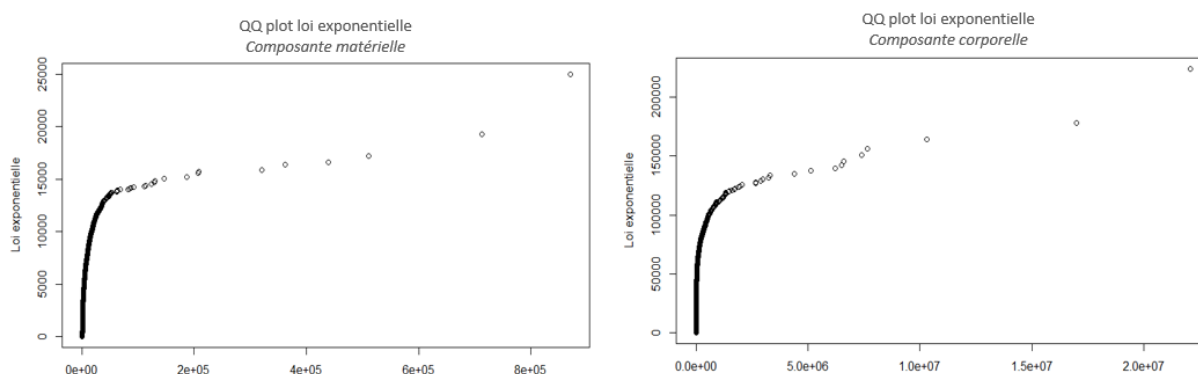


Figure 2.2 – QQ-plot des coûts par composante

D'autre part, la forme concave des QQ-plot de la loi exponentielle sur chaque composante (graphique 2.2) indique également que les données suivent une distribution à queue épaisse et confirme la présence de valeurs extrêmes.

Afin d'éviter que ces sinistres extrêmes ne biaisent les modèles de coût moyen, une solution consiste à écrêter les valeurs extrêmes. Pour cela, la théorie des valeurs extrêmes et en particulier la méthode de dépassement de seuil fournit les outils nécessaires à la détermination de ce seuil d'écrêtement. La charge résiduelle (ou surcrête) sera répartie entre les assurés via un coefficient multiplicatif.

2.2.3.1 Méthodes de détection du seuil d'écrêtement

La théorie des valeurs extrêmes propose différentes méthodes de détection du seuil d'écrêtement telles que la fonction moyennes des excès, l'estimateur de Hill ou encore l'adéquation des excès par une loi de Pareto généralisée.

Ces méthodes sont décrites ci-dessous :

Fonction moyenne des excès

Il s'agit d'une fonction seuil u , permettant de décrire la prédiction du dépassement de seuil lorsqu'un excès se produit. Elle est définie par :

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

L'estimateur de cette fonction moyenne des excès (FME) est le rapport entre la somme des excès et le nombre d'observations dépassant ce seuil :

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

Si l'ajustement des dépassements par une loi de Pareto Généralisée est vérifié avec un seuil u_0 et avec $k < 1$, la fonction moyenne des excès s'écrit pour tout $u > u_0$:

$$E[X - u | X > u] = \frac{\sigma_u}{1 - k} = \frac{\sigma_{u_0} + k(u - u_0)}{1 - k} = \frac{k}{1 - k}u + \frac{\sigma_{u_0} - ku_0}{1 - k}$$

Autrement dit, la fonction moyenne des excès est une fonction linéaire de u .

En pratique, cette approche consiste à analyser le Mean Excess Plot, le nuage de points de la FME en fonction du seuil de déclenchement des valeurs extrêmes et à choisir le seuil u la plus petite valeur à partir de laquelle la FME est linéaire.

Estimateur de Hill

L'estimateur de Hill est défini de la façon suivante :

$$\hat{\gamma}_H(k) = \frac{1}{k} \sum_{i=1}^k \log(X_{(i)}) - \log(X_{(k+1)}), k > 0$$

Avec $X_{(1)}, \dots, X_{(n)}$ les statistiques d'ordre associées à l'échantillon $X_1, \dots, X_n$.

En pratique, il s'agit de déterminer graphiquement, à partir du Hill Plot, le seuil u à partir duquel l'estimateur de Hill devient stable.

Adéquation des excès par une loi de Pareto généralisée (GPD)

Cette méthode consiste à tester l'adéquation des excès au-delà du seuil u avec une loi de Pareto Généralisée (via un *QQ-Plot* par exemple) de fonction de répartition définie par :

$$F^{GPD}_{\sigma, \gamma}(x) = 1 - \left(1 + \gamma \frac{x}{\sigma}\right)_+^{-\frac{1}{\gamma}}, x > u$$

L'adéquation des excès par une GPD utilise la relation suivante :

$$\text{Pour tout } x > u, F(x) = P(X \leq x) = (1 - P(X \leq u))F_u(x - u) + P(X \leq u)$$

où $F_u(x - u)$ est l'estimateur empirique de $F(x)$.

Et d'après la théorie des valeurs extrêmes : $F_u(x - u) \approx F^{GPD}_{\hat{\sigma}, \hat{\gamma}}(x)$ où les paramètres sont estimés par la méthode de maximum de vraisemblance.

2.2.3.2 Détermination du seuil d'écèlement par composante

Une procédure de sélection du seuil d'écèlement combinant les différentes méthodes présentées précédemment est appliquée à la base individuelle des sinistres en distinguant chacune des composantes. Il s'agit alors d'en déterminer un seuil d'écèlement adéquat.

Les éléments détaillés ci-après portent sur la **détermination du seuil d'écèlement de la composante corporelle**. L'analyse de la partie matérielle est disponible en annexe A.

Fonction moyenne des excès

Le graphique de la moyenne des excès des sinistres corporels (*Mean Excess Plot*) permet de déterminer deux seuils candidats, un premier seuil $u_1 = 1\,000\,000\text{€}$ et un second seuil $u_2 = 1\,500\,000\text{€}$.

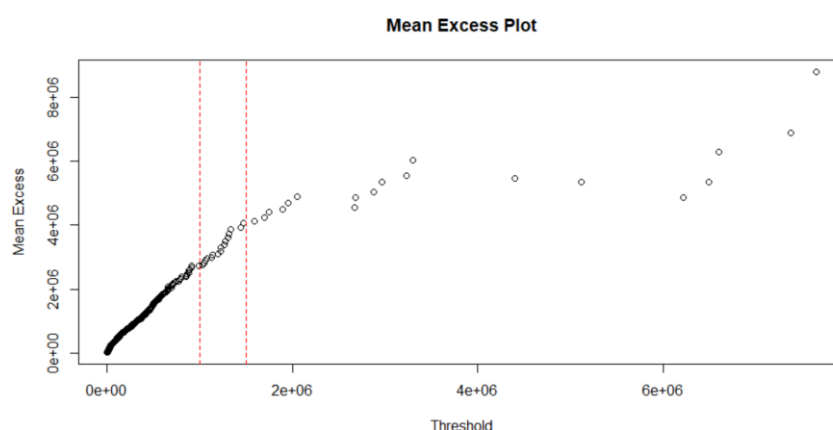


Figure 2.3 – *Mean Excess Plot* de la composante corporelle

Estimateur de Hill

Le seuil d'écèlement correspond au seuil à partir duquel l'estimateur de Hill devient relativement stable. Le graphique de Hill suggère de considérer le seuil $u_2 = 1\,500\,000\text{€}$.

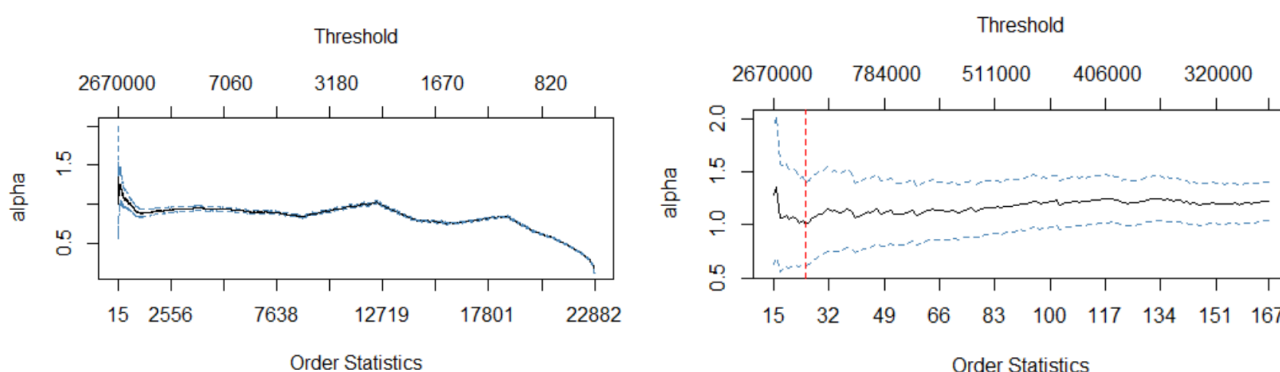


Figure 2.4 – Hill Plot de la composante corporelle

Adéquation des excès par une GPD

Le seuil $u_2 = 1\,500\,000\text{€}$ est utilisé pour ajuster les données à une loi de Pareto Généralisée. Le QQ-Plot des excès permet de confirmer ce seuil d'écèlement puisque les données sont relativement proches de la 1^{ère} bissectrice.

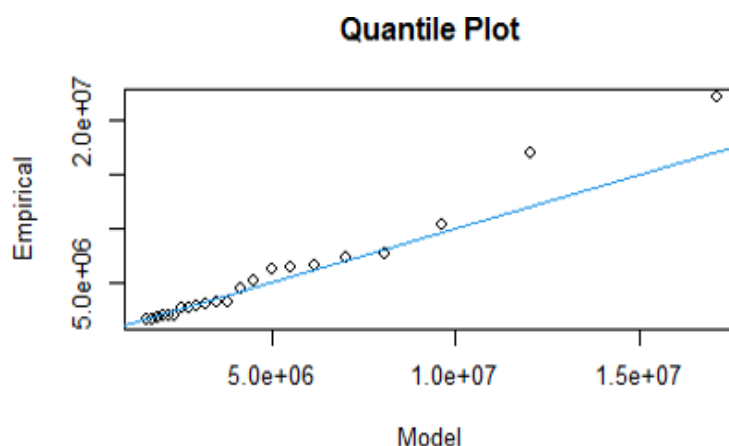


Figure 2.5 – QQ-Plot des excès de la composante corporelle avec une GPD

2.2.3.3 Résultat de l'écrêtement

Les seuils d'écrêtement retenus sont donnés par $u_{RCM} = 50\,000\text{€}$ pour la partie matérielle et $u_{RCC} = 1\,500\,000\text{€}$ pour la partie corporelle.

Les résultats de l'écrêtement par composante sont indiqués ci-dessous :

	Matérielle	Corporelle
Seuil d'écrêtement retenu	50 000 €	1 500 000 €
Nombre de sinistres > seuil	27	22
Poids de la charge en sucrête	1,6%	27,9%

Table 2.3 – Résultat de l'écrêtement par composante

Le seuil d'écrêtement de la composante corporelle apparaît très pertinent puisqu'il permet d'écarter un très faible nombre de sinistres représentant une part relativement importante de la charge corporelle (28% de la charge corporelle pour 22 sinistres). L'effet de l'écrêtement sur la composante matérielle semble beaucoup moins fort puisque la charge écrêtée représente moins de 2% de la charge matérielle.

2.3 Analyse exploratoire des données

2.3.1 Résultats généraux

Le portefeuille d'étude se compose d'un grand volume de données, il comptabilise plus de 5 millions d'années-risques dont les $\frac{3}{4}$ sont issues du canal intermédié. La prime pure définie comme la charge moyenne de sinistres par année-risque s'établit à 115€ sur la période 2013-2018 et se décompose suivant une fréquence de 6,74% et un coût moyen de 1 709€.

Répartition du portefeuille selon le canal de distribution

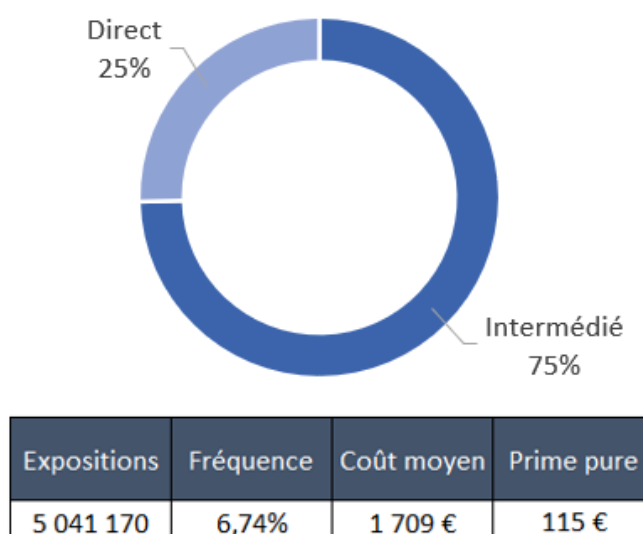


Figure 2.6 – Résultats généraux du portefeuille d'étude

2.3.2 Décomposition de la sinistralité

2.3.2.1 Séparation des composantes matérielles et corporelles

La garantie responsabilité civile automobile couvre les dommages matériels et corporels causés à autrui. Par conséquent, la sinistralité RC se décompose suivant chacune de ces deux composantes qui, par nature, présentent des caractéristiques très distinctes.

Le graphique 2.6 ci-dessous illustre bien cette forte hétérogénéité entre ces deux composantes et en décrit les principales spécificités :

- La composante matérielle représentant 39% de la charge totale des sinistres se distingue par une fréquence de sinistres élevée et un coût moyen faible (sinistres de masse à faible coût)
- La composante corporelle représentant 61% de la charge totale des sinistres se distingue par une fréquence de sinistres très faible et un coût moyen très élevé (sinistres rares à coût très important).

Répartition de la charge de sinistres RC par composante

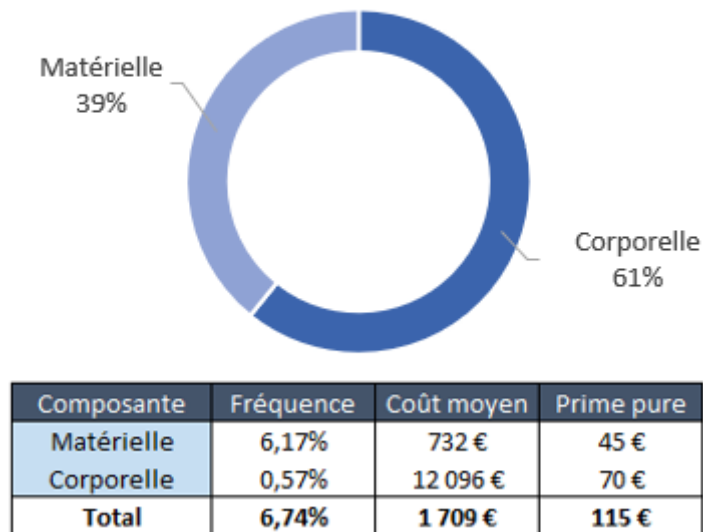


Figure 2.7 – Décomposition de la sinistralité de la garantie RC par composante

D'autre part, ces deux composantes présentent également des distributions bien différentes ce qui justifie de les traiter séparément. Ci-dessous une analyse comparative de leur distribution illustrée par le diagramme en boîtes à moustaches :

Aide à la lecture des boîtes à moustaches :



Avec Q_i les quartiles de la charge de sinistres définis comme suit :

- Le premier quartile Q_1 correspond au montant au-dessous duquel se situe 25% de la charge totale de sinistres.
- Le second quartile Q_2 correspond au montant au-dessous duquel se situe la moitié de la charge totale de sinistres, il s'agit de la médiane.
- Le troisième quartile Q_3 correspond au montant au-dessous duquel se situe 75% de la charge totale de sinistres.

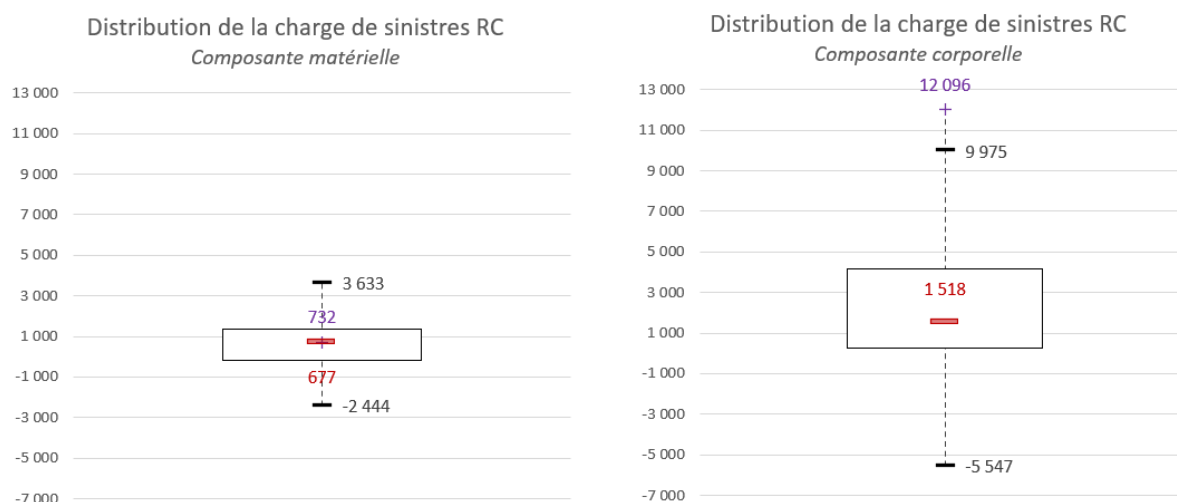


Figure 2.8 – Boîtes à moustaches par composante (yc charge négative)

Le diagramme en boîte à moustaches de la charge de sinistres RC ci-dessus met en évidence une distribution beaucoup plus étalée de la composante corporelle et confirme sa forte volatilité et la présence de valeurs extrêmes tirant la moyenne vers le haut. À noter que les valeurs négatives observées sur chacune des composantes s'expliquent par la présence de sinistres au forfait (IRSA pour la partie matérielle / IRCA pour la partie corporelle).

2.3.2.2 Séparation de la composante matérielle

La convention IRSA (Indemnisation directe de l'assuré et de Recours entre Sociétés d'Assurance) joue un rôle très particulier dans le processus d'indemnisation des sinistres RC matériels. Cette convention, signée par la plupart des compagnies d'assurances (95% des acteurs du marché), est destinée à faciliter le règlement des **dommages matériels** subis par un assuré par **une indemnisation de son propre assureur, même en cas de sinistre non responsable** et dans la mesure de leur droit à réparation, déterminé selon les règles du droit commun. L'assureur effectue ensuite un recours conventionnel auprès de (ou des) assureur(s) adverse(s) pour recevoir une indemnisation forfaitaire (forfait IDA) si le montant des dommages est inférieur au plafond de 6 500€ ou en mode réel sinon.

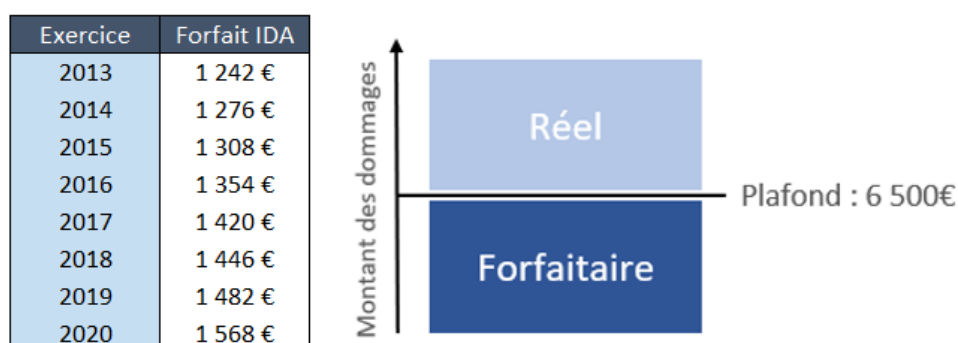


Figure 2.9 – Fonctionnement de la convention IRSA

Mais ce mode d'indemnisation n'est pas sans conséquence pour les assureurs et ceci même en cas de sinistre non responsable des assurés. En effet, dans le cas d'un sinistre non responsable dont le

montant des dommages matériels subis ne dépasse pas le plafond de 6 500€, l'indemnisation selon la convention IRSA a deux conséquences diamétralement opposées pour les assureurs :

$$\text{Coût assureur} = \begin{cases} < 0, & \text{si le montant des dommages subis} < \text{forfait IDA} \\ > 0, & \text{si le montant des dommages subis} > \text{forfait IDA} \end{cases}$$

La ventilation des sinistres matériels suivant cette convention (indemnisation IDA vs non IDA) et le niveau de responsabilité (responsable vs non responsable) est donnée ci-dessous :

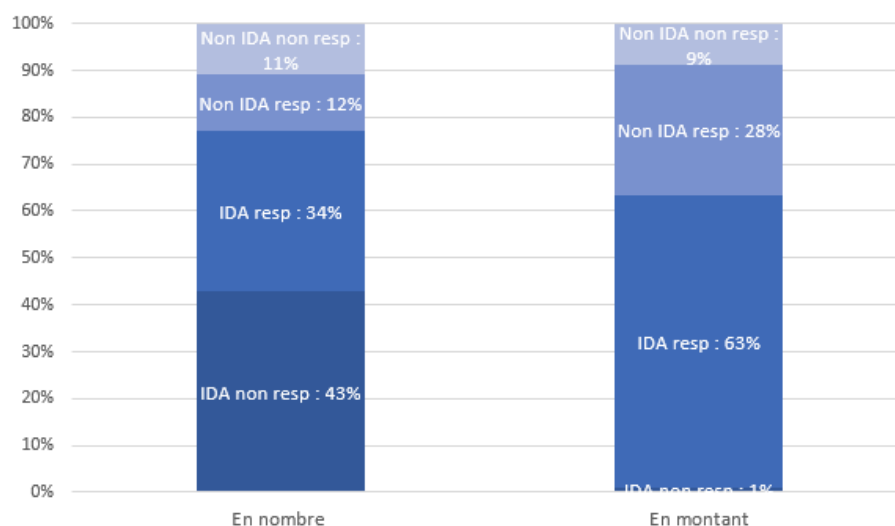


Figure 2.10 – Ventilation des sinistres matériels selon le mode d'indemnisation et le niveau de responsabilité

- L'indemnisation au forfait (IDA resp + IDA non resp) concerne 77% des sinistres matériels en nombre et 64% en montant.
- Les sinistres responsables (IDA resp + non IDA resp) représentent 46% des sinistres matériels en nombre et 90% en montant.

Partant de cette nouvelle décomposition, deux approches sont envisagées pour modéliser le coût moyen de la composante matérielle :

- 1) **Approche exhaustive** : Modélisation du coût moyen en croisant le mode d'indemnisation et le niveau de responsabilité.
Il s'agit de l'approche la plus fine et en principe la plus appropriée puisqu'elle permet un meilleur ajustement aux données des lois théoriques. Cependant cette méthode est difficile à mettre en œuvre, elle nécessite de déployer un nombre considérable de sous-modèles ce qui complexifie grandement leur combinaison.
- 2) **Approche simplifiée** : Modélisation du coût moyen en distinguant uniquement du niveau de responsabilité.

Cette méthode permet de réduire mécaniquement les sous-modèles déployés et de ce fait minimise les erreurs dues à la combinaison des modèles. Néanmoins, le point faible de cette approche réside dans la difficulté à ajuster une loi théorique à la distribution empirique des coûts moyens responsable du fait de la surreprésentation des coûts forfaitaires.

Une comparaison de ces deux approches sera réalisée dans le chapitre suivant qui traitera de la modélisation de la prime pure. Une troisième approche qui consiste à ne pas segmenter le portefeuille suivant le coût moyen (segmentation basée exclusivement sur les modèles de fréquence) ne sera pas abordée dans le cadre de ces travaux puisqu'elle a déjà fait l'objet d'un mémoire intitulé « Modélisation de la prime pure RC Automobile » (Frédéric Bouttier, 2015). Par ailleurs, cette approche semble perfectible puisqu'elle n'exploite pas la connaissance des coûts moyens des sinistres non responsables et/ou hors convention.

2.3.3 Analyse univariée

L'analyse univariée est une étape préalable à la modélisation, elle consiste à étudier l'évolution de la sinistralité (fréquence, coût moyen ou prime pure directement) en fonction des variables disponibles. Elle fournit un premier outil de détection des variables potentiellement influentes. Elle permet également de s'approprier les données et de détecter d'éventuelles sources d'instabilité des résultats (faible exposition, absence de sinistres, ...).

A titre illustratif, les éléments présentés ci-dessous se concentrent sur l'analyse de la **fréquence RC matérielle responsable** et portent sur une sélection de variables parmi les plus pertinentes.

L'âge du conducteur

Le graphique 2.11 met en évidence une variabilité importante de la fréquence de sinistres suivant l'âge du conducteur qui se traduit par une courbe convexe et une décroissance rapide de la fréquence sur le segment des jeunes entre 18-23 ans. Cette amélioration de la sinistralité est très probablement liée à une meilleure expérience de conduite dans les premières années. Par la suite, la fréquence continue de décroître mais dans une moindre mesure jusqu'à 65 ans avant de remonter progressivement. Les conducteurs non désignés (préposés d'une entreprise) sont quant à eux représentés à la droite du graphique (pic d'exposition à ~100 000 années-risques) et se caractérisent par une fréquence relativement importante au même niveau que les jeunes de 21 ans.

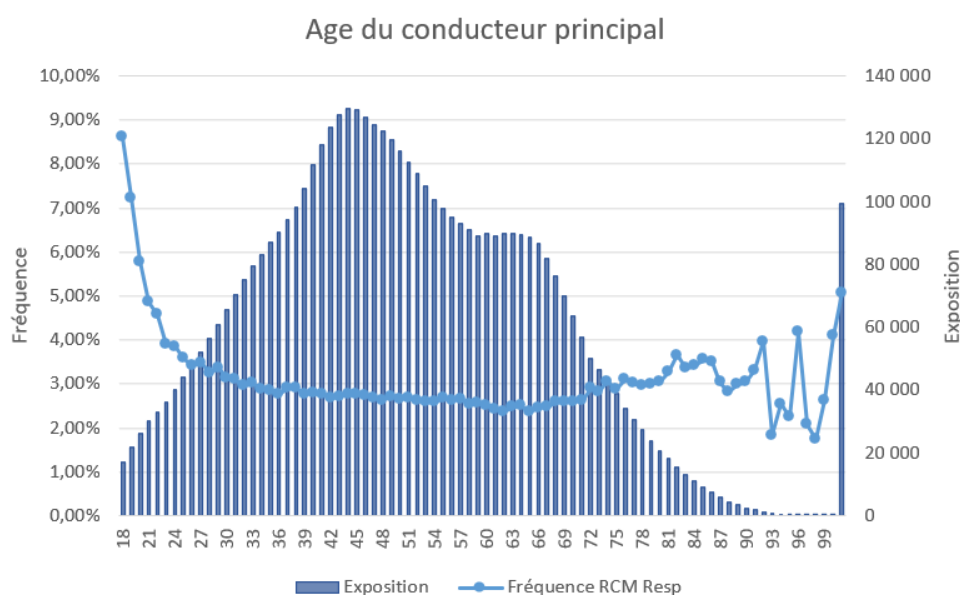


Figure 2.11 – Fréquence RCM responsable en fonction de l'âge du conducteur

Le nombre d'années à bonus 50

Le nombre d'années à bonus 50 complète la connaissance de la variable du CRM (indicateur synthétique de l'expérience de sinistralité) et permet de segmenter davantage la grande population des CRM 50. Le graphique ci-dessous montre une liaison de type linéaire entre la fréquence de sinistres et le nombre d'années à bonus 50. Elle est naturellement plus faible pour les CRM 50 avec une ancienneté supérieure ou égale à 6 ans et plus élevée pour les CRM > 50.

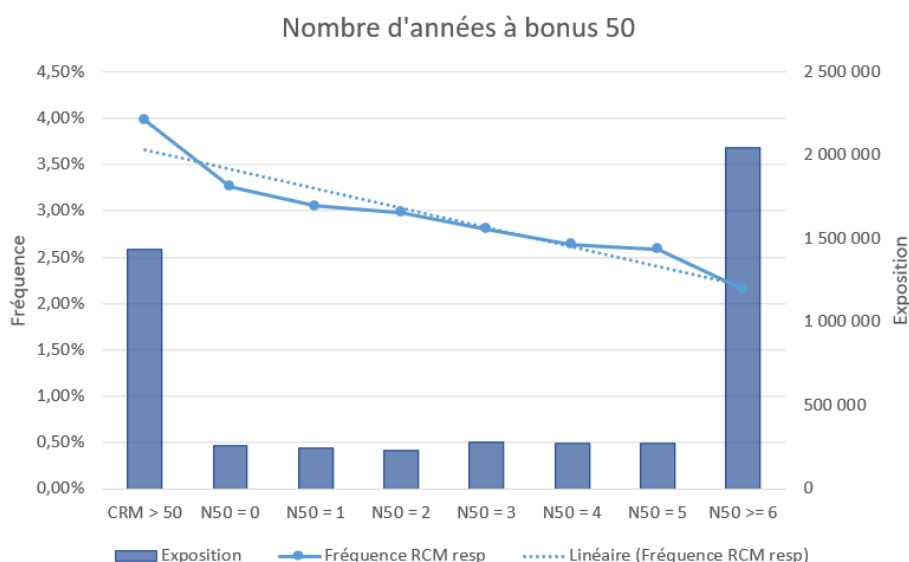


Figure 2.12 – Fréquence RCM responsable en fonction du nombre d'années à bonus 50

L'ancienneté du véhicule

Le graphique ci-dessous montre une évolution décroissante de la fréquence de sinistres selon l'ancienneté du véhicule, elle est très élevée sur les véhicules récents et plus basse sur les véhicules

anciens. A noter le faible volume d'exposition des véhicules de 30 ans et plus qui justifierait un regroupement de ces modalités pour gagner en robustesse des estimateurs.

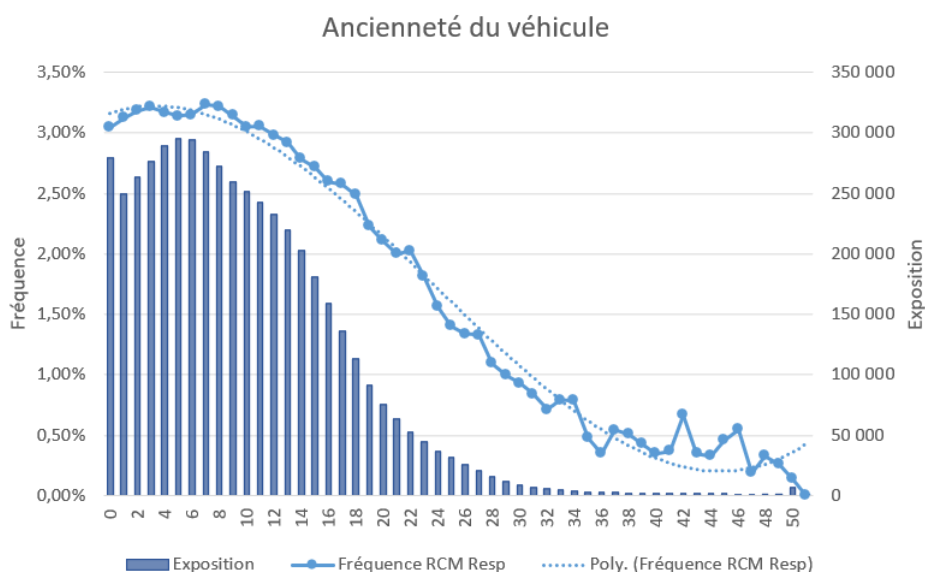


Figure 2.13 – Fréquence RCM responsable en fonction de l'ancienneté du véhicule

Les forfaits kilométriques

L'étude de la fréquence de sinistres selon le forfait kilométrique fait apparaître une opposition entre le segment des « petits » rouleurs (forfait 5 000km) qui se caractérisent par une fréquence très basse à près de 1,50% et le segment des « gros » rouleurs (sans forfait) qui se distingue par une fréquence de 3,50%, soit un rapport de la fréquence de 2,3 entre ces populations.

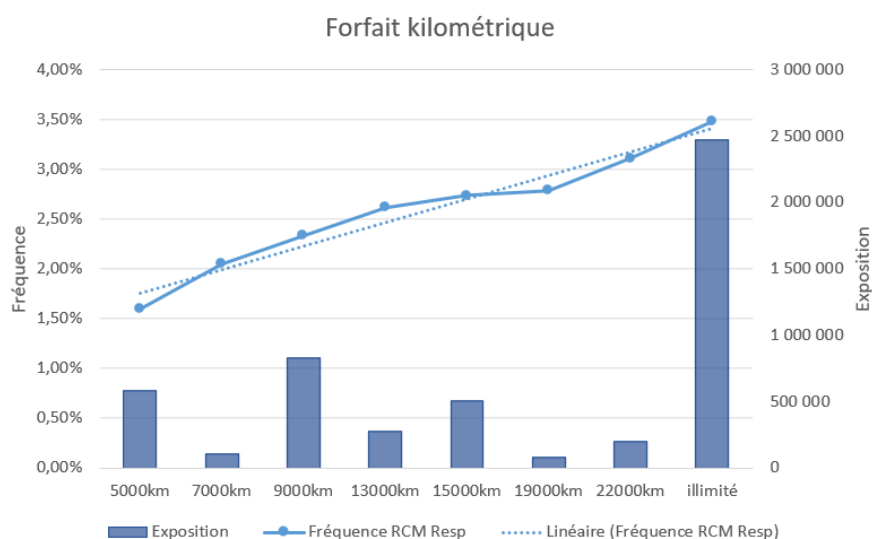


Figure 2.14 – Fréquence RCM responsable en fonction du forfait kilométrique

2.4 Etude des corrélations entre variables explicatives

Dans le cadre d'un modèle de régression, une forte corrélation entre variables explicatives est problématique car elle entraîne une augmentation de la variance des estimateurs les rendant instables et difficilement interprétables. C'est pourquoi il est important d'étudier les corrélations en amont de la modélisation afin de détecter et quantifier ces liaisons.

Pour cela, deux analyses complémentaires ont été menées :

- Une étude des corrélations par la méthode du V de Cramer
- Une classification automatique de variables

2.4.1 Méthode du V de Cramer

L'approche classique dans la détection des corrélations consiste à utiliser la statistique du V de Cramer qui permet d'obtenir directement l'intensité de la liaison entre deux variables qualitatives ou quantitatives préalablement discrétisées. Le V de Cramer est un indicateur synthétique et très lisible, sa valeur varie entre 0 (liaison nulle) et 1 (liaison maximale).

Sa statistique est donnée par la formule suivante :

$$V \text{ de Cramer} = \sqrt{\frac{\chi^2}{\chi_{max}^2}}$$
$$\text{avec } \chi^2 = \sum_{ij} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}$$

Le seuil empirique de 30% est généralement retenu indiquant une forte corrélation si celui-ci est dépassé. L'analyse du corrélogramme ci-dessous permet d'identifier visuellement les variables les plus corrélées, elle est complétée par une lecture de la matrice des corrélations indiquant l'intensité exacte de la liaison. Il apparaît par exemple une forte corrélation entre l'ancienneté de permis et l'âge du conducteur (V de Cramer de 62%), l'ancienneté du véhicule et la formule souscrite (V de Cramer de 47%) ou encore la classe de prix SRA et le groupe SRA (V de Cramer de 44%).

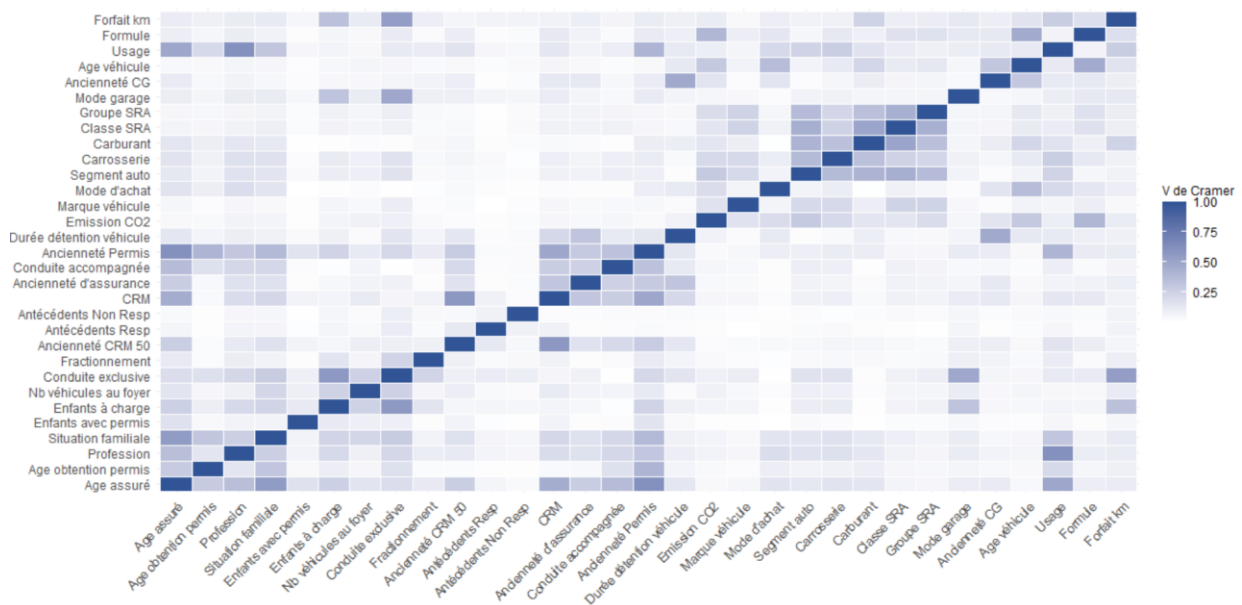


Figure 2.15 – Corrélogramme des variables explicatives

2.4.2 Classification automatique de variables

Le but de la classification de variables est de regrouper les variables les plus liées afin de construire des classes de variables homogènes participant à une meilleure lisibilité des corrélations entre les variables. Elle est très utile dans un objectif de réduction de la dimension lorsque les variables à analyser sont nombreuses avec de fortes corrélations mais également dans un objectif de sélection de variables car elle permet de s'assurer d'avoir capté toute l'information disponible en ayant sélectionné au minimum un représentant dans chaque classe de variables.

Cette opération a été menée à l'aide du package *ClustOfVar* du logiciel *R* qui permet de classer simultanément des variables quantitatives et qualitatives. L'approche par la classification ascendante hiérarchique (CAH) a été privilégiée car elle ne nécessite pas d'a priori sur le nombre de classes à retenir contrairement à une approche *k-means*.

2.4.2.1 Présentation de la méthode

Cette méthode met en œuvre une CAH sur les variables. Elle commence par une partition en singleton en considérant chaque variable comme une classe et procède à des agrégations successives de deux classes jusqu'à l'obtention d'une seule classe contenant la totalité des variables. La suite des partitions est représentée par un arbre appelé *dendrogramme*.

L'agrégation entre deux classes s'appuie sur une notion de distance ou mesure de dissimilarité qui mesure la perte d'homogénéité lorsque deux classes sont agrégées.

Par ailleurs, l'algorithme fournit une évaluation de la stabilité des classes par une approche de type *bootstrap* constituant une aide précieuse dans le choix du nombre de classes à retenir.

Mesure d'agrégation entre deux classes

La distance entre deux classes de variables A et B est donnée par :

$$d(A, B) = H_A + H_B - H_{A \cup B}$$

Où H_k le critère d'homogénéité de la classe C_k

Homogénéité d'une classe de variable

L'homogénéité H_k de la classe C_k est une mesure d'adéquation entre les variables de la classe C_k et sa variable synthétique Z_k définie comme la combinaison linéaire des variables de la classe C_k ou encore la première composante principale de l'analyse factorielle de données mixtes (AFDM).

L'homogénéité H_k est donnée par :

$$H_k = \sum_{\substack{j \in C_k \\ j \text{ quantitative}}} r^2(x_j, Z_k) + \sum_{\substack{j \in C_k \\ j \text{ qualitative}}} \eta^2(x_j, Z_k) = \lambda_1^k$$

Où

- $r^2(x_j, Z_k)$ représente le carré du coefficient linéaire de Pearson entre la variable x_j et la variable synthétique Z_k
- $\eta^2(x_j, Z_k)$ représente le carré du rapport de corrélation de x_j avec Z_k
- λ_1^k est la première valeur propre de l'AFDM appliquée aux variables de la classe C_k

L'homogénéité H de la partition $\{C_1, \dots, C_K\}$ est définie comme la somme des homogénéités de ses K classes et peut être vue comme une mesure globale de qualité de la classification qui doit être maximale :

$$H = \sum_{k=1}^K H_k = \sum_{k=1}^K \lambda_1^k$$

Stabilité d'une partition

Une évaluation de la stabilité d'une partition est réalisée en tirant B échantillons *bootstrap* et en comparant les classifications obtenues avec la classification initiale selon l'indice de Rand qui mesure la proportion des concordances entre les deux classifications parmi l'ensemble de toutes les paires de variables. Une valeur de l'indice égale à 1 indique une classification parfaitement stable.

Choix du nombre de classes

Le choix du nombre de classes à retenir s'appuie sur le panel de critères suivant :

- L'examen attentif du *dendrogramme* afin de déterminer des classes dites « naturelles ».
- Le graphique de la perte d'inertie interclasse (perte d'homogénéité lorsque deux classes sont agrégées) suivant la règle du coude.
- La stabilité de la classification en fonction du nombre de classes.

2.4.2.2 Application au portefeuille

L'analyse du graphique 2.16 de la perte d'inertie interclasse en fonction du nombre de classes indique la présence de 3 coudes orientant le choix vers une classification à 2, 7 ou 9 classes, toutes les 3 stables au sens de l'indice de Rand avec un indice de stabilité de 1 (graphique 2.15). Cependant une classification à 2 classes semble beaucoup trop restrictive (seulement 20,8% de l'inertie expliquée) et peu pertinente d'un point de vue métier alors qu'une classification à 7 ou 9 classes permettrait de réduire le nombre de variables tout en conservant une part significative de l'inertie expliquée (52,5% pour une classification à 7 classes et 59,5% pour 9 classes).

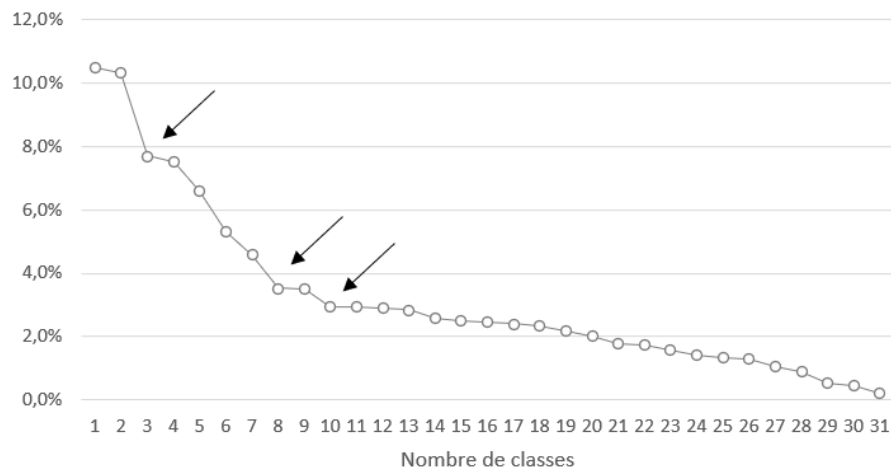


Figure 2.16 – Perte de l'inertie interclasse en fonction du nombre de classes

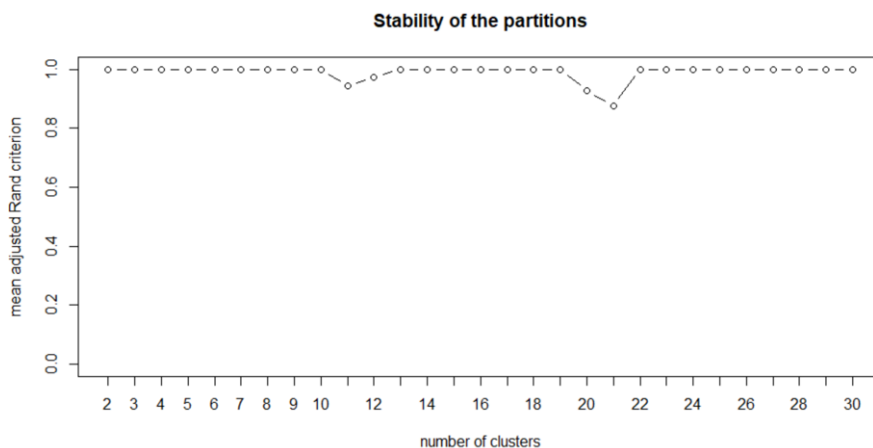


Figure 2.17 – Stabilité des partitions de la classification de variables

L'examen du *dendrogramme* ci-dessous (graphique 2.18) permet finalement de fixer le nombre de classes à retenir à 9 qui est préférable à une classification à 7 groupes puisqu'elle apparaît plus pertinente d'un point de vue métier. En effet, une classification à 9 groupes permet d'isoler le groupe des antécédents à la souscription (classe n°4).

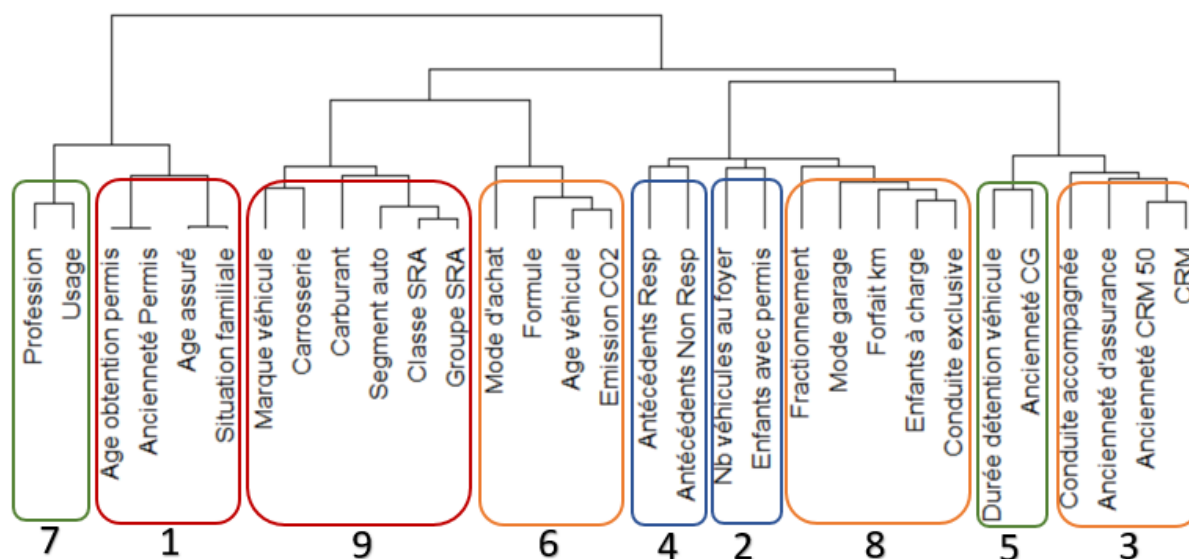


Figure 2.18 – Dendrogramme de la classification de variables

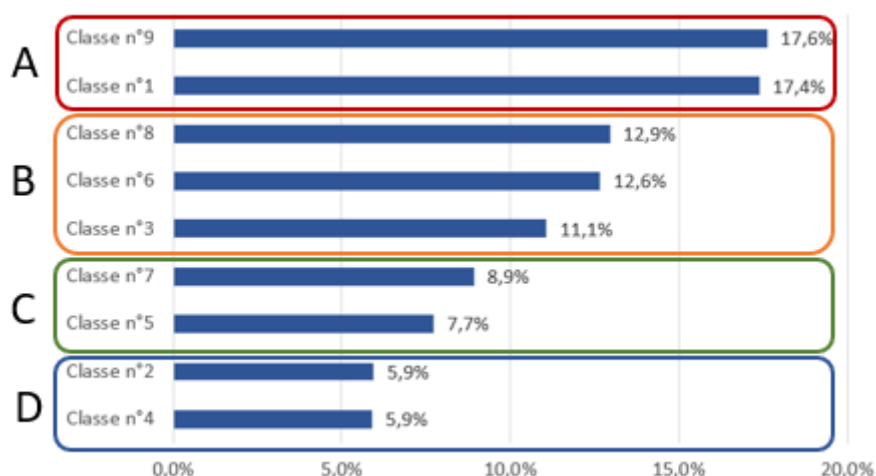


Figure 2.19 – Contribution des classes à la variance expliquée

Interprétation des classes retenues

Une hiérarchie des classes de variables suivant leur contribution à la variance expliquée est donnée par le graphique 2.19 ci-dessus et fait apparaître 4 niveaux d'explicabilité des classes de variables :

- Le groupe A (35% de la variance expliquée) est constitué des 2 classes les plus contributrices (classes n°9 et n°1), à savoir les caractéristiques du véhicule (groupe et classe de prix SRA, segment et type de carrosserie) et les informations liées au conducteur (âge du conducteur, situation familiale, ...).
- Le groupe B (36% de la variance expliquée) se compose de 3 classes de variables, la classe n°8 qui reflète le niveau d'utilisation du véhicule (forfait kilométrique, conduite exclusive, ...), la

classe n°6 renseignant sur le niveau de couverture du véhicule et la classe n°3 relative à l'expérience de sinistralité des assurés (CRM, ancienneté d'assurance, ...).

- Le groupe C (17% de la variance expliquée) regroupe la classe n°7 définie par l'usage du véhicule et la profession de l'assuré et la classe n°5 expliquée par la durée de détention du véhicule (précédent ou actuel).
- Le groupe D (12% de la variance expliquée) est celui qui apporte le moins d'information, il se compose des classes n°2 et n°4 faisant référence respectivement aux informations de type foyer et aux antécédents de sinistres à la souscription.

Matrice de corrélations entre classes de variables

Une matrice de corrélation réduite constituée des classes de variables issues de la classification à 9 groupes est donnée ci-dessous. Elle permet de vérifier que les classes de variables sont peu corrélées entre elles, ce qui est globalement le cas excepté peut être des classes n°3 et n°5 qui affichent un coefficient de corrélation de 31% qui s'explique par une liaison relativement importante entre la variable de l'ancienneté d'assurance appartenant à la classe n°3 et la durée de détention appartenant à la classe n°5 (graphique 1.21).

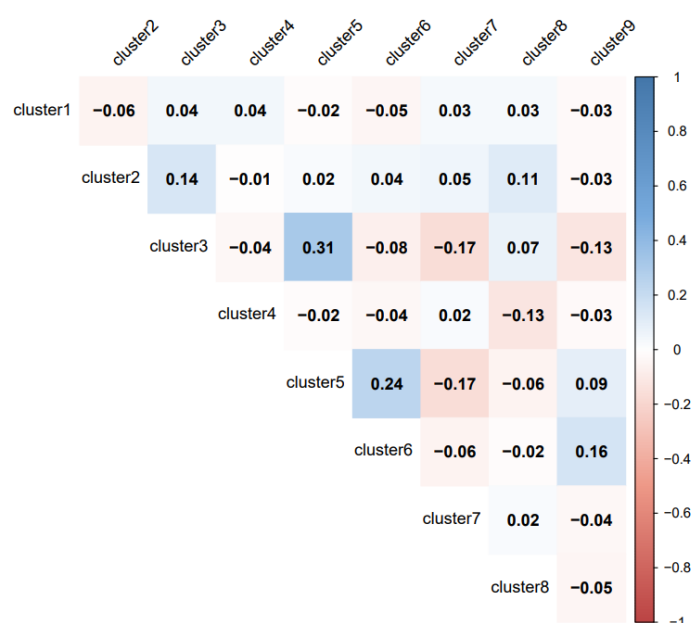


Figure 2.20 – Matrice des corrélations entre les classes de variables

	Durée détention véhicule	Ancienneté carte grise	Conduite accompagnée	Ancienneté d'assurance	Ancienneté CRM 50	CRM
Durée détention véhicule	1,00					
Ancienneté carte grise	0,48	1,00				
Conduite accompagnée	0,12	0,05	1,00			
Ancienneté d'assurance	0,32	0,13	0,25	1,00		
Ancienneté CRM 50	0,14	0,10	0,21	0,16	1,00	
CRM	0,21	0,13	0,28	0,32	0,58	1,00

Figure 2.21 – Matrice de corrélations entre les variables initiales des classes n°3 et n°5

Chapitre 3

Modélisation de la prime pure

Ce troisième chapitre est dédié à la modélisation de la prime pure de la garantie RC par une méthode largement éprouvée, à savoir les modèles linéaires généralisés (GLM). Cette technique est très utilisée dans le secteur de l'assurance car elle offre de nombreux avantages tels qu'une bonne lisibilité du tarif, une structure tarifaire multiplicative qui facilite sa mise en œuvre ou encore la possibilité d'ajuster manuellement le tarif sur des segments stratégiques.

Comme évoqué précédemment, les composantes matérielles et corporelles seront traitées séparément selon une approche classique de fréquence-coût. Le modèle de prime pure sera alors obtenu en combinant chacun de ces sous-modèles. L'élaboration des modèles sera réalisée à partir d'un échantillon d'apprentissage représentant 80% de la base d'étude et leur performance sera évaluée sur les 20% restant constituant la base de test. Cet échantillonnage vise à mesurer la capacité de généralisation des modèles et à se prémunir d'un éventuel sur-apprentissage.

Ce chapitre sera également l'occasion d'aborder un sujet incontournable dans la tarification d'un contrat d'assurance automobile, à savoir la création d'un zonier qui permet de capter l'information géographique du risque.

3.1 Motivation de l'approche

En assurance automobile, une approche fréquence-coût moyen est généralement retenue pour modéliser la prime pure car elle permet de d'isoler les effets propres à chaque facteur. En effet, alors que la fréquence de sinistres est principalement liée au profil des conducteurs, à leur exposition au risque ou à leur environnement, le coût moyen s'exprime davantage par les caractéristiques du véhicule.

3.1.1 Le cadre usuel de la tarification IARD

Soit S le montant total de la charge de sinistres, alors dans le modèle collectif, S peut s'écrire sous la forme :

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

Avec :

- N , la variable aléatoire désignant le nombre de sinistres survenus pendant la période considérée
- X_i , la variable aléatoire désignant le coût du sinistre i

Où :

- Les variables aléatoires N et X_i sont indépendantes et admettent des moments d'ordre 2

- X_i est une suite de variables aléatoires positives indépendantes et identiquement distribuées (*iid*)

Sous ces différentes hypothèses, la prime pure définie comme l'espérance de la charge des sinistres, s'écrit :

$$\begin{aligned}
 E(S) &= E\left(\sum_{i=1}^N X_i\right) \\
 &= E\left(E\left[\sum_{i=1}^N X_i | N\right]\right) \\
 &= E\left(\sum_{i=1}^N E(X_i)\right) \\
 &= E(N \times E(X)) \\
 &= E(N) \times E(X)
 \end{aligned}$$

Cette dernière égalité traduit que, dans le cadre du modèle collectif, modéliser la prime pure d'une garantie donnée revient à modéliser séparément la fréquence et le coût moyen.

La modélisation de la prime pure peut donc être réalisée suivant deux approches, la modélisation séparée de la fréquence et du coût moyen d'une part et de l'autre, la modélisation directe de la prime pure. Chacune de ces approches possède ses avantages et ses inconvénients :

L'approche fréquence-coût

Avantages :

- **Précision de la tarification** : Elle permet de prendre en compte dans la tarification à la fois les facteurs qui influencent la fréquence et le coût moyen.
- **Interprétabilité des modèles** : Les modèles de fréquence et de coût moyen peuvent être plus faciles à interpréter car ils se concentrent sur des aspects spécifiques du risque.
- **Flexibilité dans la modélisation** : Il est possible d'utiliser différents types de loi pour modéliser la fréquence et le coût moyen.

Inconvénients :

- **Indépendance des composantes** : Elle suppose une indépendance entre la fréquence et le coût moyen, ce qui peut ne pas être toujours réaliste.
- **Complexité de la tarification** : Elle nécessite de combiner les estimations issues des modèles de fréquence et de coût moyen.

L'approche directe

Avantages :

- **Simplicité dans la modélisation** : Cette approche est plus simple car elle nécessite un seul modèle pour estimer la prime pure faisant intervenir une loi de *Tweedie*. Elle permet également de s'affranchir de l'hypothèse d'indépendance entre la fréquence et le coût moyen.
- **Parcimonie du modèle** : Une modélisation directe en prime pure fait généralement intervenir un nombre plus limité de facteurs.

Inconvénients :

- **Précision de la tarification** : Elle peut ne pas capter aussi précisément les facteurs spécifiques qui influencent la fréquence et le coût moyen que la modélisation séparée.
- **Interprétabilité des résultats** : L'interprétabilité des résultats peut être plus complexe car il peut être difficile de démêler l'impact des facteurs sur la prime pure.

3.1.2 Tests d'indépendance

La modélisation de la prime pure selon une approche fréquence-coût suppose une hypothèse forte d'indépendance entre la fréquence et le coût moyen. Celle-ci, bien que communément admise, peut être vérifiée par un test statistique tel que le test d'indépendance du khi-deux encore le test d'indépendance basé sur le coefficient de corrélation des rangs de Spearman qui ont tous deux l'avantage d'être des tests non paramétriques (ils ne reposent pas sur des hypothèses spécifiques concernant la distribution sous-jacente des données).

Test du χ^2

Le test d'indépendance du khi-deux est utilisé pour évaluer s'il existe une association significative entre deux variables catégorielles.

Les hypothèses de ce test sont les suivantes :

$$\begin{cases} H_0 : \text{les deux variables sont indépendantes} \\ H_1 : \text{les deux variables sont liées} \end{cases}$$

Sous H_0 (indépendance des variables), la statistique du test est définie par :

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - T_{ij})^2}{T_{ij}}$$

Où O_{ij} et T_{ij} représentent respectivement les effectifs observés et théoriques.

Le nombre de degrés de liberté du χ^2 est égal à :

$$p = (\text{nombre de lignes} - 1) \times (\text{nombre de colonnes} - 1)$$

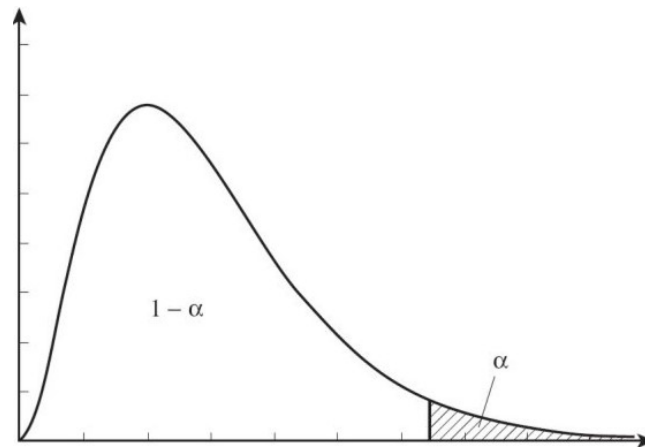


Figure 3.1 – Densité de la loi du khi-deux

L'hypothèse d'indépendance est acceptée au niveau α si la p – valeur $= P(\chi^2 > \chi^2_{1-\alpha}) > \alpha$.

Le coefficient de corrélation des rangs de Spearman

Le test d'indépendance basé sur le coefficient de corrélation des rangs de Spearman est utilisé pour évaluer s'il existe une association monotone (même non linéaire) significative entre deux variables qui peuvent être continues, discrètes ou ordinales.

Soit n le nombre d'observations, le coefficient de Spearman est donné par la formule :

$$\rho = 1 - \frac{6 \sum_{i=1}^n (r_{xi} - r_{yi})^2}{n(n^2 - 1)}$$

Où r_{xi} et r_{yi} représentent respectivement le rang de chaque observation i selon la variable X puis selon la variable Y .

Sous H_0 (indépendance des variables), la statistique du test est définie par :

$$Z = \sqrt{\rho^2 \frac{n-2}{1-\rho^2}} \sim T_{n-2}$$

Où T_{n-2} est la loi de Student à $n-2$ degrés de liberté.

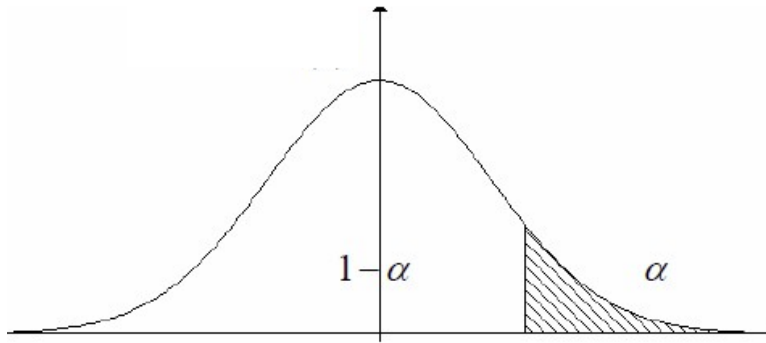


Figure 3.2 – Densité de la loi de Student

L'hypothèse d'indépendance est acceptée au niveau α si la p – valeur = $P(Z > T_{1-\alpha}) > \alpha$.

Application aux données :

La fréquence et le coût moyen étant deux quantités continues, le test d'indépendance basé sur le coefficient de corrélation des rangs de Spearman (test en t) apparaît le plus approprié. D'autre part, il est à noter que le test du khi-deux peut être moins robuste en grande dimension.

Ci-dessous le résultat du test d'indépendance pour les 3 composantes du modèle :

Composante	Coef de corrélation	p-valeur
RCM resp	-0,00285	0,6631
RCM non resp	0,01596	0,2642
RCC	-0,01170	0,3936

Table 3.1 – Résultat du test de Spearman

La condition d'indépendance entre la fréquence et le coût moyen est vérifiée sur l'ensemble des composantes ce qui permet d'envisager une modélisation séparée fréquence-coût.

3.2 Rappels théoriques

3.2.1 Les modèles linéaires généralisés

Les modèles linéaires généralisés (GLM) sont une généralisation des modèles linéaires classiques ayant pour objectif de mettre en relation une variable aléatoire à expliquer Y avec un ensemble de variables explicatives X . Mais contrairement aux modèles linéaires classiques, les GLM n'imposent pas une distribution normale de la variable à expliquer et la relation entre l'espérance conditionnelle de $Y|X$ et les variables explicatives n'est pas nécessairement linéaire.

En revanche, dans le cadre des GLM, la variable à expliquer $Y = (Y_1, \dots, Y_n)$ composée de n variables aléatoires *iid* doit suivre une loi de probabilité appartenant à la famille exponentielle. Cette dernière condition implique que Y admet une densité de la forme :

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

Avec :

- θ appelé paramètre naturel
- ϕ appelé paramètre de dispersion
- $a(\cdot)$, $b(\cdot)$ et $c(\cdot)$ des fonctions réelles

Exemple de lois exponentielles : Normale, Poisson, Gamma, Binomiale, Binomiale négative, ...

Ces différents paramètres sont spécifiés en fonction de la loi choisie. Par ailleurs, l'espérance et la variance d'une loi Y suivant une distribution de la forme exponentielle sont données par :

$$E(Y) = b'(\theta), \text{Var}(Y) = b'' \cdot a(\phi)$$

Par exemple pour une loi de Poisson $P(\lambda)$:

$$f(y|\theta, \phi) = \exp(y\ln(\lambda) - \lambda - \ln(y!))$$

Avec $\theta = \ln(\lambda)$, $a(\phi) = 1$, $b(\theta) = \exp(\theta)$ et $c(y, \phi) = -\ln(y!)$

En appliquant les formules précédentes :

$$E(P(\lambda)) = b'(\theta) = \lambda \text{ et } \text{Var}(P(\lambda)) = b'' \cdot a(\lambda) = \lambda$$

La formule du modèle linéaire généralisée est de la forme :

$$g(E(Y|X = x)) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

où g est une fonction strictement monotone et dérivable appelée fonction de lien. Elle détermine la relation entre le prédicteur linéaire et l'espérance de la variable expliquée. Les fonctions de lien les plus fréquentes sont rappelées ci-dessous.

Fonction de lien	$g(u)$	Loi de $Y X$
Identité	u	Normale
Log	$\ln(u)$	Poisson
Inverse	u^{-1}	Exponentielle
Puissance	u^p	Gamma
Logit	$\ln\left(\frac{u}{1-u}\right)$	Binomiale

Table 3.2 – Fonctions de lien usuelles

3.2.2 Estimation des paramètres du modèle

L'estimation des paramètres β du modèle est obtenue par maximum de vraisemblance. Cette méthode très utilisée et donnant de bons estimateurs, consiste à maximiser la probabilité d'obtenir les réalisations observées $y_1, \dots, y_n$ sur un n -échantillon *iid*.

L'objectif est de déterminer les paramètres qui maximisent la fonction de vraisemblance ou de manière équivalente, la fonction log-vraisemblance donnée par :

$$l(\theta, \phi, y) = \ln(L(\theta, \phi, y)) = \ln\left(\prod_{i=1}^n f(y_i | \theta, \phi)\right) = \sum_{i=1}^n \left(\frac{y_i \theta - b(\theta)}{a(\phi)} + c(y, \phi) \right)$$

Il s'agit de chercher $\hat{\beta}$ tel que $\hat{\beta} = \operatorname{argmax}_{\beta} l(\theta, \phi, y)$.

L'estimation par maximum de vraisemblance conduit à résoudre « l'équation de la vraisemblance » :

$$\frac{\partial \ln(L(\theta, \phi, y))}{\partial \beta} = 0$$

Puis de vérifier que la dérivée seconde est négative (présence d'un maximum) :

$$\frac{\partial^2 \ln(L(\theta, \phi, y))}{\partial \beta^2} < 0$$

De plus, l'estimation par maximum de vraisemblance permet d'obtenir un intervalle de confiance des paramètres estimés.

3.2.3 Significativité des paramètres du modèle

Une fois les paramètres estimés, il est important de s'assurer de leur significativité. Cette vérification est effectuée par un test statistique appelé test de Wald.

Test de Wald

Il s'agit d'un test d'hypothèse destiné à tester la significativité de chaque coefficient du modèle.

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$$

Sous H_0 (nullité de la $j^{\text{ième}}$ modalité de la variable i), la statistique du test est définie par :

$$W = \left(\frac{\hat{\beta}_j}{\hat{\sigma}_j} \right)^2 \sim \chi_1^2$$

En conséquence, l'hypothèse nulle sera rejetée au niveau α si la p -valeur = $P(W > \chi_{1;1-\alpha}^2) < \alpha$. Dans le cas contraire, le coefficient sera considérée comme nul et donc non significatif.

3.2.4 Qualité d'ajustement et validation du modèle

La qualité d'ajustement du modèle est mesurée par une comparaison entre les observations et les estimations. Pour cela, une mesure de la déviance complétée d'une analyse des résidus est généralement réalisée.

La déviance

Cette quantité est une mesure d'écart du modèle par rapport au modèle saturé, c'est-à-dire le modèle parfait qui admet autant de paramètres que d'observations.

Soit $L(\beta_{max})$ la vraisemblance du modèle saturé et $L(\hat{\beta})$ celle du modèle considéré, alors la déviance est donnée par :

$$D = -2 \times \log \left(\frac{L(\hat{\beta})}{L(\beta_{max})} \right) = 2 \times \log (L(\beta_{max}) - L(\hat{\beta})) \geq 0$$

Un bon ajustement du modèle se traduira par une déviance relativement faible.

Cette statistique suit approximativement une loi du Khi-deux à $n-p-1$ degrés de liberté, où n est égal au nombre d'observations et p le nombre de paramètres estimés dans le modèle considéré. Ce dernier résultat permet de construire un test de significativité du modèle. Une règle empirique consiste à comparer la valeur $D/(n-p-1)$ à 1 de telle sorte qu'une valeur très supérieure à 1 indique un modèle mal ajusté.

Les résidus

Les résidus mesurent l'écart entre l'observation réelle et la valeur prédite par le modèle. L'analyse des résidus permet une identification des observations qui pourraient expliquer une éventuelle inadéquation du modèle aux données.

Il existe différents types de résidus mais les plus utilisés sont les résidus de Pearson et de déviance :

- Résidus de Pearson : $r_i^P = \frac{y_i - \mu_i}{\sqrt{V(\mu_i)}}$

- Résidus de déviance : $r_i^D = \text{signe}(y_i - \mu_i)\sqrt{d_i}$, $D = \sum_{i=1}^n d_i$

L'analyse des résidus est généralement graphique et consiste à vérifier le caractère normal de leur distribution, l'égalité de leur variance (homoscédasticité) et leur indépendance.

3.2.5 Critères de sélection de modèles

Le processus de sélection du meilleur modèle doit nécessairement s'appuyer sur l'analyse de plusieurs indicateurs testés en parallèle permettant de donner un avis objectif complet sur les forces et faiblesses de chaque modèle considéré. Ces critères de sélection doivent notamment tenir compte de la complexité des modèles, de leur pouvoir de discrimination ou encore de leur robustesse.

Les critères de pénalisation AIC et BIC

Afin d'assurer une bonne généralisation du modèle, il est nécessaire qu'il soit le plus parcimonieux possible. Or le seul critère de la déviance peut conduire à un phénomène de sur-apprentissage du fait de l'amélioration mécanique de cet indicateur en fonction du nombre de variables du modèle. Les critères AIC et BIC apportent une réponse à cette problématique par une pénalisation de la log-

vraisemblance suivant le nombre de paramètres p du modèle. Le meilleur modèle sera celui qui minimise ces critères.

$$AIC = -2 \ln(L(\hat{\beta})) + 2p$$

$$BIC = -2 \ln(L(\hat{\beta})) + p \ln(n)$$

L'indice de Gini

L'indice de Gini est un indicateur qui évalue la capacité d'un modèle à segmenter le risque. Il varie entre 0 et 1, une valeur proche de 1 illustre une bonne discrimination du modèle. Il peut être représenté graphiquement par l'aire entre la 1^{ère} bissectrice et la courbe de Lorenz $L(\alpha)$ qui permet d'étudier les inégalités de répartition d'une grandeur positive cumulée.

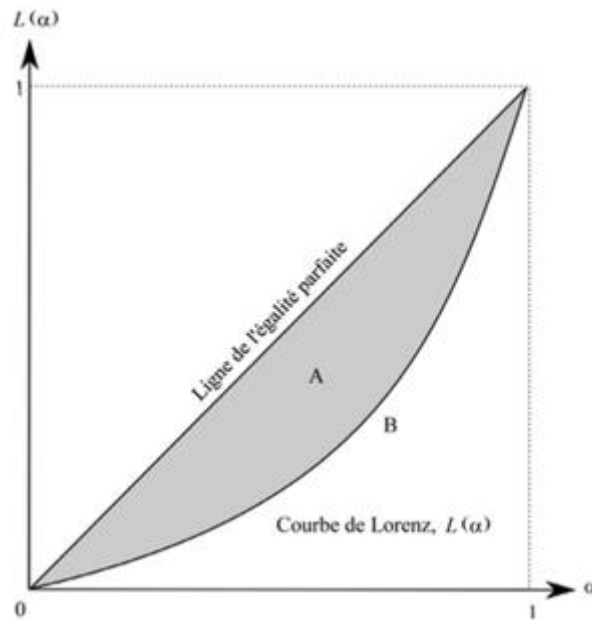


Figure 3.3 – Courbe de Lorenz

$$\text{Indice de Gini} = \frac{A}{A + B} = 2A$$

La robustesse

Ce critère exprime la capacité des modèles à ne pas être perturbés par une modification dans les données. Il s'agit de mesurer le pouvoir de généralisation des modèles sur une autre base que celle qui a servi à construire les modèles.

3.2.6 Stratégie de sélection de variables

3.2.6.1 La méthode Stepwise

La méthode Stepwise est une méthode ascendante de sélection de variables qui, à chaque étape, effectue un test de significativité des variables déjà présentes pour éventuellement éliminer la moins significative d'entre elles. L'algorithme s'arrête lorsque plus aucune variable ne permet d'améliorer le modèle au sens du critère choisi.

3.2.6.2 La méthode LASSO

La méthode LASSO (*Least Absolute Shrinkage and Selection Operation*) consiste à ajouter une contrainte sur la norme L_1 des coefficients par le biais d'une fonction de pénalisation dans le problème de minimisation de l'erreur quadratique et ceci afin de diminuer la variance des estimateurs. L'objectif de cette méthode est la recherche d'un modèle parcimonieux. La fonction à minimiser est la suivante :

$$\sum_{i=1}^n \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j|, \lambda \geq 0$$

La valeur du coefficient λ influence l'intensité de la pénalisation :

- Lorsque $\lambda = 0$ (pas de pénalisation), l'estimateur obtenu est celui des moindres carrés ordinaires.
- La pénalisation augmente en fonction de la valeur de λ et les modèles sont de plus en plus parcimonieux (une valeur de λ trop grande renvoie un modèle nul).

Le calibrage du paramètre λ est donc d'une importance majeure, celui-ci est réalisé selon une méthode de minimisation d'un critère de choix (MSE, déviance, MAE, AUC, ...) en procédant par validation croisée. L'algorithme utilisé fournit deux valeurs pour le paramètre de pénalisation λ :

- $\lambda_{\min}$ correspond à la valeur de λ qui minimise l'erreur de validation croisée. Cette approche est généralement la plus utilisée dans un problème de prédiction.
- λ_{1se} correspond à la plus grande valeur de λ telle que son erreur soit inférieure à une erreur-type du minimum. Etant donné que $\lambda_{1se} \geq \lambda_{\min}$, le nombre de variables retenues est logiquement plus réduit qu'avec la valeur $\lambda_{\min}$.

3.3 Modélisation de la fréquence

La modélisation de la fréquence RC fait intervenir 3 sous-modèles constitués d'un modèle de fréquence matérielle responsable, d'un modèle de fréquence matérielle non responsable et d'un modèle de fréquence corporelle. La séparation suivant le mode d'indemnisation (IDA/non IDA) n'a pas été retenue puisqu'elle n'a pas d'effet sur la fréquence de sinistres. La distinction suivant le niveau de responsabilité fait naturellement sens puisqu'un sinistre responsable implique un comportement fautif du conducteur et donc un certain profil qu'il convient de déterminer.

L'élaboration du zonier sera également traitée dans cette section et une mesure de son apport dans les différents modèles de fréquence sera réalisée.

Afin de ne pas dupliquer inutilement les analyses, **l'ensemble des résultats présentés tout au long de cette section concerne le modèle de fréquence RC matérielle responsable**. Les principaux résultats relatifs au modèle de fréquence RC matérielle non responsable est disponible en annexe B.

3.3.1 Détermination de la loi

En règle générale, la modélisation de la fréquence de sinistres automobile fait intervenir la loi de Poisson. Cependant, afin de déterminer la loi la plus appropriée aux données, deux tests d'adéquations de lois ont été effectués, la première avec une loi de Poisson et la seconde avec une loi Binomiale négative.

Ci-dessous, la comparaison de l'ajustement de la loi empirique avec ces deux lois théoriques. Une première analyse graphique montre clairement un meilleur ajustement avec la loi de Poisson. D'autre part, le test de Kolmogorov-Smirnov qui consiste à mesurer l'écart maximum D (en valeur absolue) entre la fonction de répartition empirique et la fonction de répartition théorique confirme ce dernier résultat.

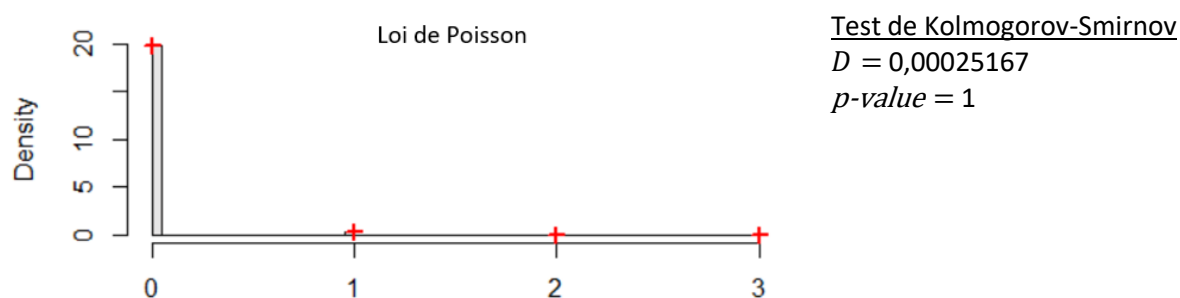


Figure 3.4 – Ajustement de la fréquence RCM responsable à une loi de Poisson

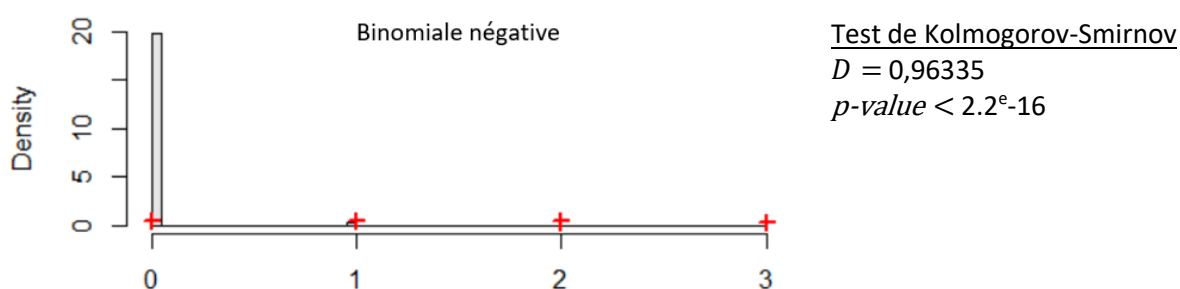


Figure 3.5 – Ajustement de la fréquence RCM responsable à une loi binomiale négative

3.3.2 Prise en compte des contraintes

L'exposition au risque impacte la fréquence de sinistres par définition, il est donc nécessaire d'en tenir compte dans les modèles. Pour cela, une variable *offset* de l'exposition est introduite sous la forme $\log(\text{exposition})$.

Une autre contrainte propre à l'assurance automobile doit également être prise en compte, il s'agit de la contrainte réglementaire du CRM. Celle-ci consiste à niveler le tarif en fonction du niveau de CRM de l'assuré. De la même manière que pour l'exposition, une seconde variable *offset* égale à $\log(\text{CRM})$ est donc introduite dans les modèles.

3.3.3 Sélection de variables

La sélection de variables se déroule en plusieurs étapes en appliquant différents filtres successifs qui visent à obtenir le meilleur sous-ensemble de variables possible.

Tout d'abord, un premier filtre « naturel » est appliqué sur les variables candidates en écartant celles qui n'ont pas de lien véritable avec la garantie RC, il s'agit par exemple des variables relatives aux franchises dommage. Il s'agit aussi d'éliminer les variables qui pourraient biaiser les résultats et qui expliqueraient un effet canal de distribution. C'est notamment le cas de la variable conduite exclusive qui identifie de manière détournée le canal de distribution et doit donc être mise de côté pour l'instant.

Ensuite, il convient de prendre en compte l'analyse des corrélations menée lors de l'analyse exploratoire des données en écartant les variables les plus corrélées entre elles et qui expliquent le même phénomène. Par exemple, la variable de l'âge obtention de permis est écartée car elle peut être déduite à partir de l'âge du conducteur et de l'ancienneté de permis. De même, le segment du véhicule (variable SRA) peut être approché par la combinaison d'autres variables telles que la classe de prix, le groupe du véhicule et le type de carrosserie qui apportent davantage d'information.

Enfin, sur la base des variables restantes, une hiérarchie des variables les plus significatives est établie en s'appuyant sur les méthodes décrites dans la section 3.2.6. Concrètement, il s'agit de tester en priorité les variables qui ont recueilli le plus de « votes » lors du processus de sélection de variables par les méthodes Stepwise et Lasso, puis de compléter la sélection par les autres variables si celles-ci permettent d'améliorer la performance du modèle. Cette stratégie vise à optimiser la sélection de variables en combinant plusieurs méthodes.

Les principaux résultats sont décrits ci-dessous :

Le tableau 3.3 dénombre les variables retenues par chaque méthode, le modèle le plus parcimonieux est obtenu par la méthode LASSO avec le paramètre λ_{1se} (9 variables) et le plus complexe est obtenu par la méthode Stepwise (17 variables).

Méthode	Critère de sélection	Nb variables retenues
Stepwise	AIC	17
LASSO $\lambda_{\min}$	Déviance	15
LASSO λ_{1se}	Déviance	9

Table 3.3 – Complexité du modèle de fréquence RCM responsable en fonction de la méthode de sélection

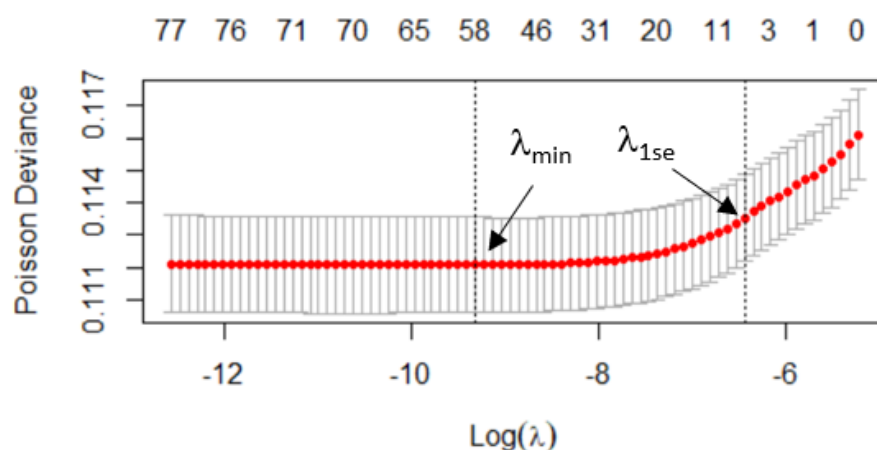


Figure 3.6 – Evolution de la déviance du modèle de fréquence RCM responsable selon la valeur du paramètre $\log(\lambda)$ et du nombre de régresseurs

Une synthèse de la sélection de variables du modèle de fréquence RC matérielle responsable est donnée par le tableau 3.4 qui recense les variables retenues par chaque méthode, le nombre de « votes » recueillis par chaque variable (1,2 ou 3 fois) et enfin la sélection finale qui se compose de 17 variables dont la plupart sélectionnées unanimement par les deux méthodes.

Variable	Stepwise	Lasso $\lambda_{\min}$	Lasso λ_{1se}	Nb sélections	Retenu
Mode d'achat		✓		1	✓
Age conducteur	✓	✓	✓	3	✓
Durée d'assurance	✓	✓	✓	3	✓
Ancienneté carte grise	✓	✓		2	✓
Ancienneté véhicule	✓	✓	✓	3	✓
Carburant	✓	✓		2	✓
Carrosserie	✓	✓	✓	3	✓
Profession		✓		1	✓
CRM	✓	✓	✓	3	offset
Mode garage	✓			1	✓
Marque véhicule	✓			1	
Ancienneté CRM 50	✓	✓	✓	3	✓
Enfants avec permis	✓	✓		2	✓
Antécédents responsables	✓	✓	✓	3	✓
Antécédents non responsables	✓			1	✓
Nb voitures au foyer	✓	✓		2	✓
Classe de prix SRA	✓			1	
Forfait km	✓	✓	✓	3	✓
Usage	✓	✓	✓	3	✓
Nb variables sélectionnées	17	15	9		

Table 3.4 – Synthèse de la sélection de variables du modèle de fréquence RCM responsable

Une détection des interactions via la méthode LASSO a également été réalisée, elle est représentée sous la forme d'une matrice et les interactions sont hiérarchisées en moyennant les coefficients sur chaque croisement. Le graphique 3.7 indique notamment une interaction entre le forfait kilométrique et l'ancienneté d'assurance ou encore entre la profession et le forfait kilométrique.

Cependant, ces différentes interactions n'ont pas été retenues dans le modèle final car leur apport n'était pas suffisamment significatif au sens du critère de l'AIC.

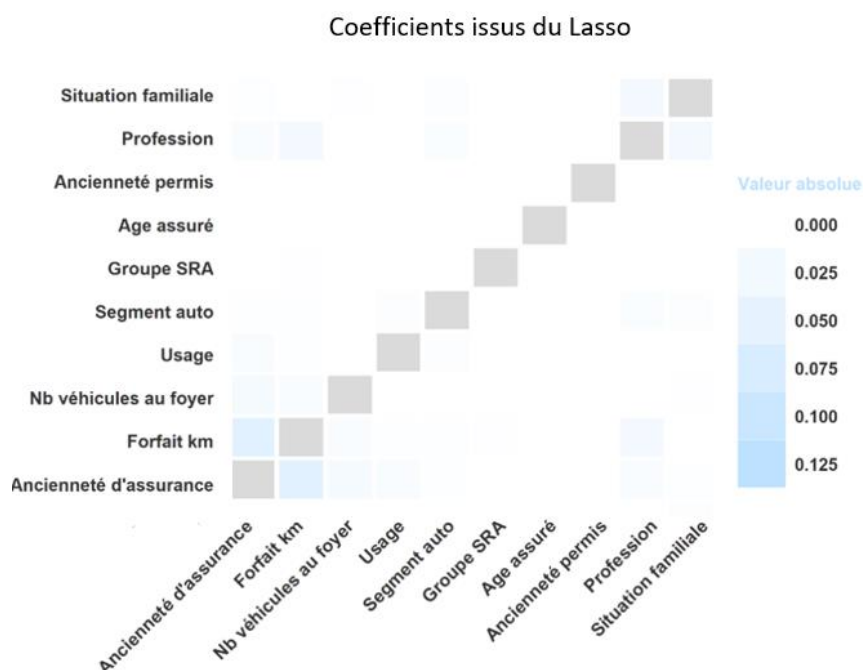


Figure 3.7 – Détection des interactions du modèle de fréquence RCM responsable

3.3.4 Retraitements des variables dans les modèles

Cette étape est essentielle pour réduire la complexité des modèles, elle se compose principalement de deux types de retraitements :

- **Le regroupement de modalités** présentant un faible niveau d'expositions, d'une logique métier ou d'un manque de significativité de l'estimateur d'une modalité.
- **L'utilisation de splines** pour les variables de type ordinal qui présentent une tendance avérée. Cette simplification permet notamment de gommer les fluctuations trop brutales entre deux valeurs consécutives (lissage des estimateurs).

Un exemple concret est présenté ci-dessous, il concerne la variable de l'ancienneté du véhicule et illustre les deux types de retraitements décrits ci-dessus.

Le premier retraitement appliqué est le regroupement de modalités, ici le choix s'est naturellement porté sur un regroupement des véhicules de 30 ans et plus car il correspond à un seuil de passage vers une dénomination en véhicule de collection.

Le second retraitement est celui du lissage permettant d'obtenir une courbe de coefficients lissée respectant la tendance naturelle des estimateurs.

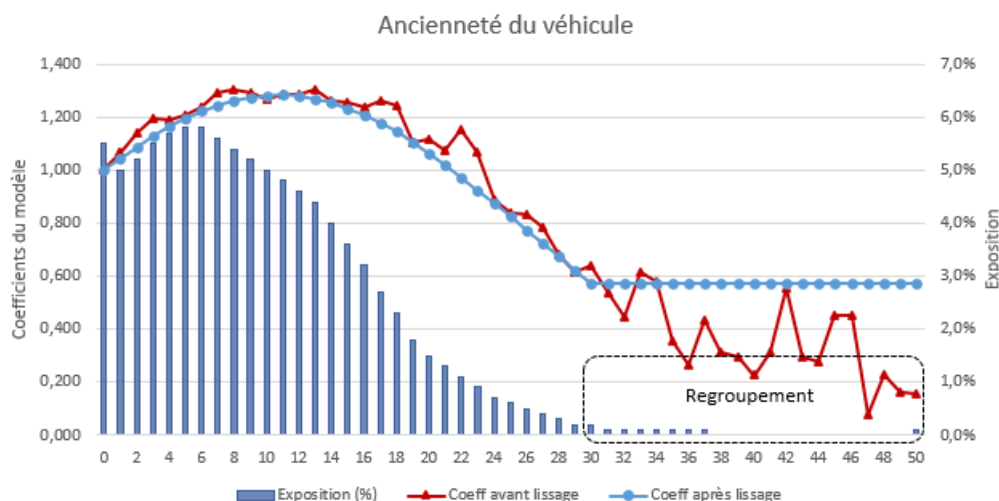


Figure 3.8 – Retraitements de l'ancienneté du véhicule

3.3.5 Stabilité des variables dans le temps

Afin de garantir une bonne généralisation des modèles, il est nécessaire de s'assurer de la stabilité dans le temps des coefficients associés à chacune des modalités de chaque variable retenue. Ce contrôle est réalisé en croisant chacune des variables avec l'exercice de survénance, ce qui permet d'observer la tendance prédite sur chaque exercice. Ainsi, si la variable présente une tendance similaire sur chaque exercice, elle sera conservée et dans le cas contraire elle fera l'objet de regroupements supplémentaires pour la stabiliser dans le temps ou sera tout simplement supprimée.

Un exemple de variable stable dans le temps est donné par le graphique 3.9 montrant une même tendance polynomiale de la courbe des estimateurs de l'ancienneté du véhicule qui se vérifie sur les différents exercices de survénance (2013-2018).

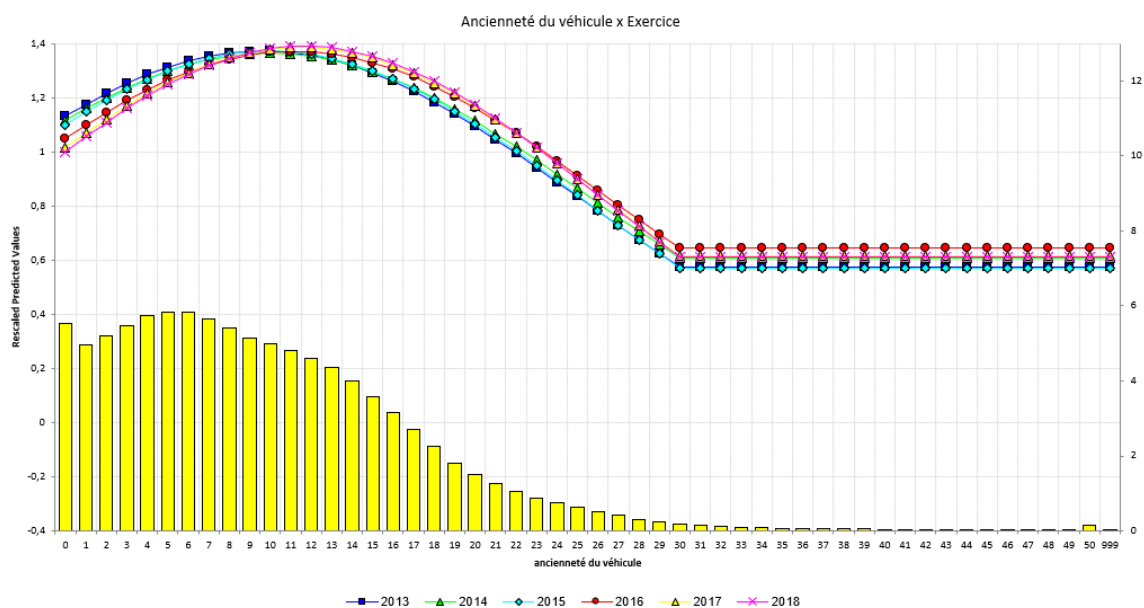


Figure 3.9 – Stabilité dans le temps de l'ancienneté du véhicule

3.3.6 Elaboration du zonier

En assurance automobile, la zone géographique fait généralement partie des facteurs qui explique le mieux la sinistralité et plus particulièrement la fréquence de sinistres. Cette forte influence s'explique par exemple par la densité du trafic routier, les conditions météorologiques ou encore l'état ou la configuration des routes (routes plus ou moins accidentogènes). C'est pourquoi il est important d'intégrer cette composante géographique dans la tarification en supplément des autres variables déjà présentes.

3.3.6.1 Approche retenue

La méthode de zonage retenue et qui est aussi la plus fréquemment utilisée, est celle du lissage spatial. Elle consiste à modéliser les résidus d'une modélisation hors effet géographique à l'aide de variables externes, l'utilisation des variables géographiques permettant de prédire la partie géographique des résidus. Elle est suivie d'une étape de lissage qui utilise l'information des zones voisines, le paramètre de lissage étant déterminé selon différents critères (fiabilité, robustesse des zones et stabilité temporelle, avis d'expert). L'avantage du lissage est qu'il permet de s'affranchir des problèmes d'exposition et de réduire les variations entre zones voisines. Les zones géographiques sont, quant à elles, déterminées suivant différentes méthodes de classification, certaines différencient les groupes par leur volume (méthode quantile ou poids égal) et d'autres par l'écart entre les coefficients (méthode Ward, Average Linkage, Centroïde).

Les différentes étapes de l'élaboration d'un zonier sont schématisées ci-dessous :

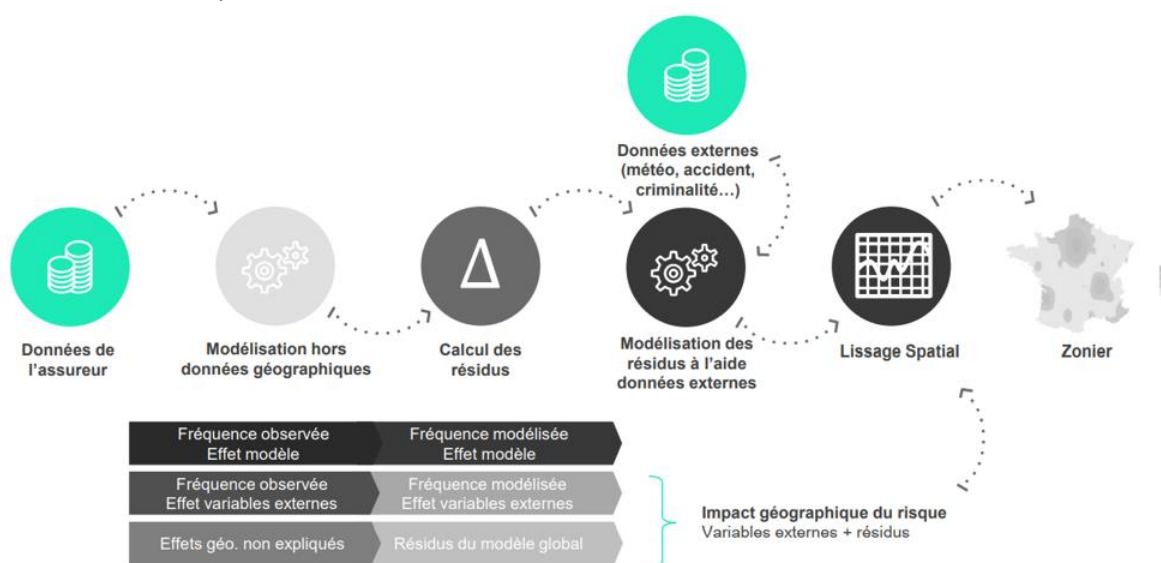


Figure 3.10 – Schéma de création d'un zonier par la méthode du lissage spatial

Le lecteur souhaitant se documenter davantage sur les différentes techniques de zonage pourra par exemple se référer au mémoire « Techniques de zonage en tarification IARD » (Brice Carles, 2017).

3.3.6.2 Application aux données

Le zonier est établi à la maille commune *INSEE* à partir des résidus du modèle de fréquence RC matérielle obtenu après combinaison des modèles de fréquence matérielle responsable et non

responsable. La méthode de classification considérée est celle d'une répartition équilibrée en 12 zones de même exposition afin de satisfaire à une contrainte liée à une distribution du produit sur le canal intermédiaire.

Le graphique 3.11 met en évidence une croissance stricte de la fréquence RC matérielle observée par zone avec une fréquence à 3,85% pour la zone 1 (la moins risquée) et une fréquence à 8,87% pour la zone 12 (la plus risquée) ce qui laisse présager une bonne discrimination par le zonier.

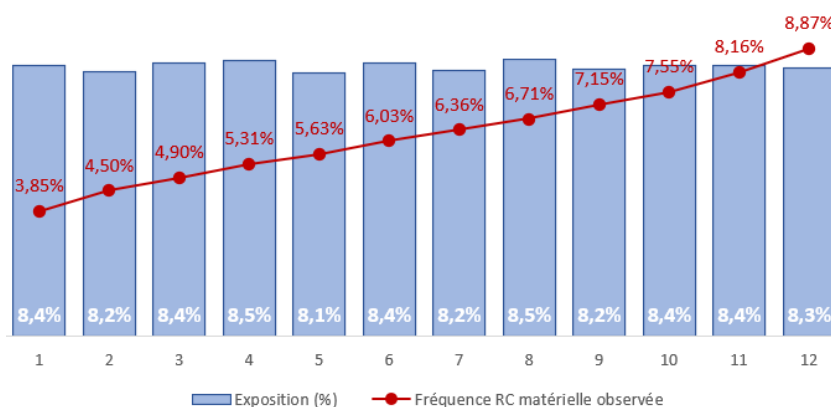


Figure 3.11 – Fréquence RCM observée par zone géographique

Ci-dessous la carte résultante du découpage en 12 zones :

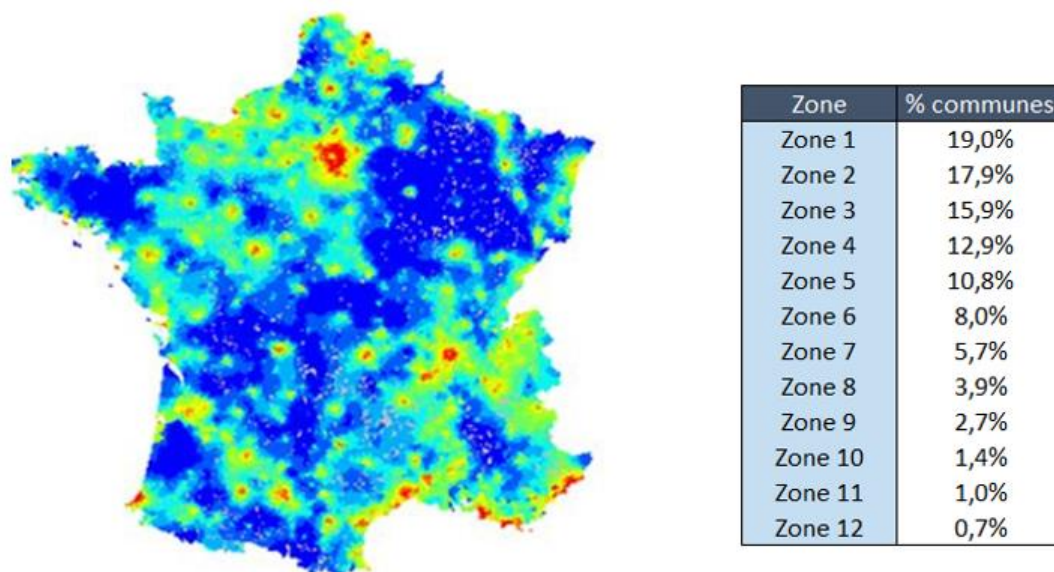


Figure 3.12 – Zonier RC omnicanal et répartition des communes par zone géographique

Le dégradé de couleur va du bleu pour les zones les moins risquées au rouge pour les plus risquées. Les zones les plus touchées (zones 11 et 12 - en rouge sur la carte) ne concernent que peu de communes et se concentrent sur les villes de moyennes ou de grandes tailles et en particulier la région parisienne, la région lyonnaise et le sud-est de la France (la région PACA). Ces différentes régions se caractérisent par une forte densité de population et un trafic routier très saturé et donc potentiellement accidentogène, il n'est donc pas très étonnant de les retrouver parmi les zones les plus risquées.

3.3.7 Résultats du modèle

3.3.7.1 Présentation du modèle

Une manière relativement simple de résumer un modèle est par exemple de hiérarchiser les variables retenues en fonction de leur influence dans le modèle. La méthode présentée ci-dessous permet d'établir un classement des variables en fonction de leur importance au sens du critère de la déviance.

Le gain pour chaque variable au modèle est calculé comme suit :

$$\text{Gain (\%)} = \frac{\text{Déviance sans la variable} - \text{Déviance avec toutes les variables}}{\text{Déviance sans aucune variable} - \text{Déviance avec toutes les variables}}$$

La contribution de chaque variable au modèle est alors obtenue en divisant son gain par la somme des gains de toutes les variables.

Le graphique 3.13 permet d'identifier les variables les plus influentes dans le modèle de fréquence RC matérielle responsable, à savoir le zonier qui se détache très nettement des autres variables avec une contribution de 24,5%, suivi de l'âge du conducteur à 17,1% et enfin du forfait kilométrique avec une contribution de 14,5%, soit une contribution cumulée de ces 3 variables de 56,1%. Inversement, les antécédents responsables et l'ancienneté d'assurance sont les variables qui contribuent le moins et semblent n'avoir qu'un apport très modeste dans le modèle, ce qui pourrait s'expliquer par une information captée en grande partie par la variable du CRM.

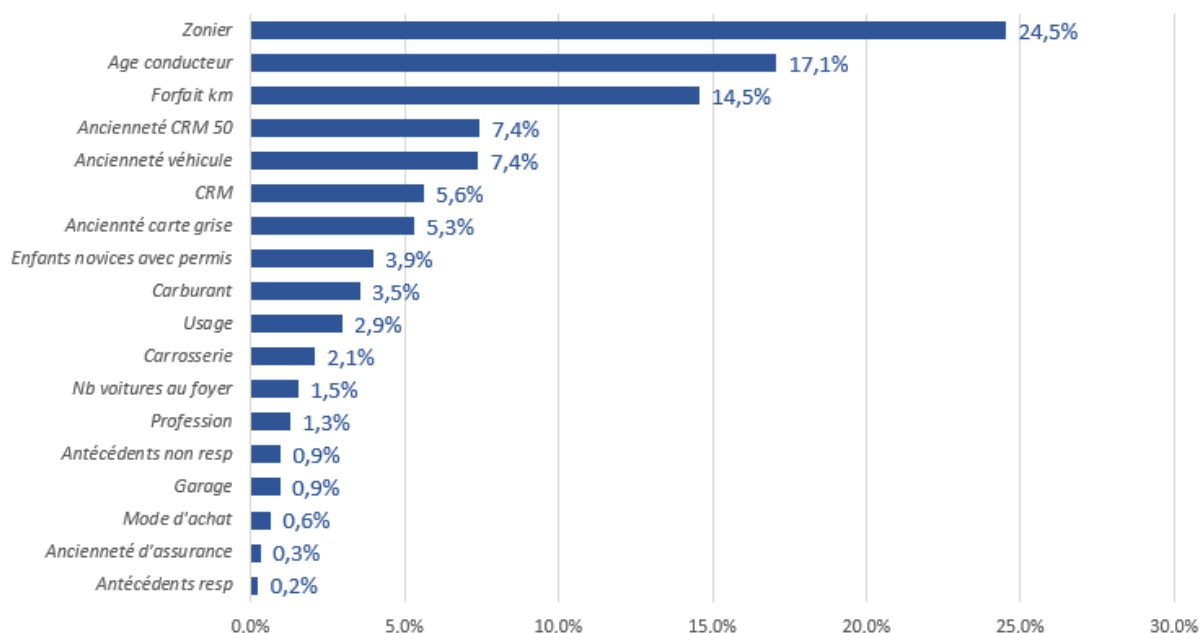


Figure 3.13 – Contributions des variables au modèle de fréquence RCM responsable

3.3.7.2 Analyse et interprétation des résultats

Une analyse des résultats est réalisée en parcourant chacune des variables qui composent le modèle. Cette étude porte notamment sur la comparaison des fréquences prédites et observées. Un extrait de cette étude est présenté ci-dessous et concernent les variables de la zone géographique (graphique 3.14), de l'âge du conducteur (graphique 3.15) et de l'ancienneté du véhicule (graphique 3.16). Sur chacun de ces graphiques, il est possible de visualiser simultanément la fréquence observée (en rose), la fréquence prédite (en vert foncé) et les coefficients du modèle (en vert clair).

La zone géographique

Comme attendu, la fréquence augmente en fonction du niveau de risque de la zone géographique et les coefficients du modèle sont aussi fidèles à ce constat. Concernant l'ajustement du modèle, il est satisfaisant malgré quelques écarts ponctuels et notamment une surestimation sur les zones 1 et 10. L'écart constaté pour la modalité « 999 », qui correspond aux risques mal rattachés à une commune, est non significatif compte-tenu de son faible niveau d'exposition.

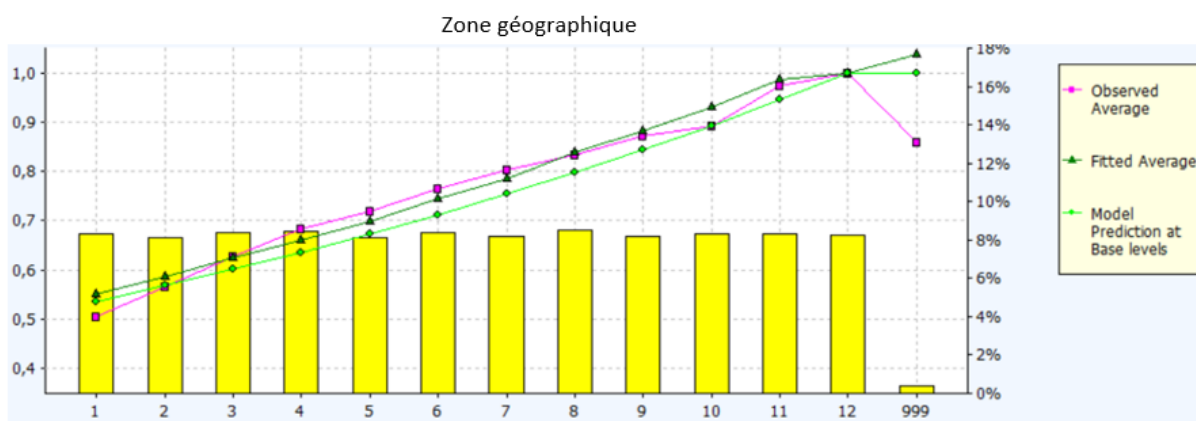


Figure 3.14 – Comparaison des fréquences RCM responsable prédites et observées selon la zone géographique

L'âge du conducteur

Le modèle semble particulièrement bien ajusté sur cette variable puisque la courbe de la fréquence prédite est très proche de l'observée. Concernant la courbe des coefficients, elle apparaît désaxée par rapport à la fréquence observée, ce phénomène est généralement rencontré en présence d'une corrélation avec une autre variable. Ici, il s'agit d'une corrélation avec le CRM qui a été introduit en offset pour satisfaire à la contrainte réglementaire.

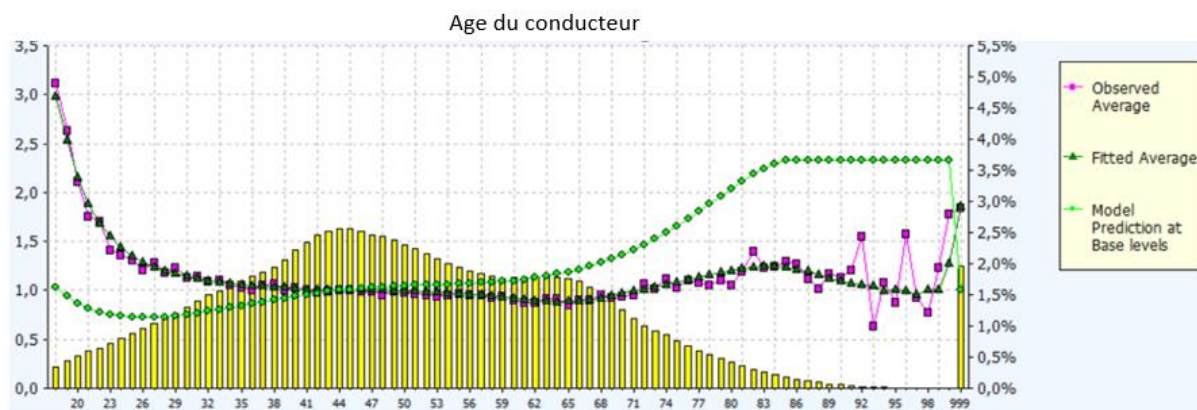


Figure 3.15 – Comparaison des fréquences RCM responsable prédites et observées selon l'âge du conducteur

L'ancienneté du véhicule

L'adéquation aux données apparaît de bonne qualité malgré une légère sous-estimation de la fréquence sur les véhicules récents (entre 1-3 ans). L'allure de la courbe des coefficients est quant à elle en phase avec l'évolution de la fréquence observée qui décrit une tendance polynomiale ce qui est fidèle aux premiers constats établis lors de l'analyse univariée.

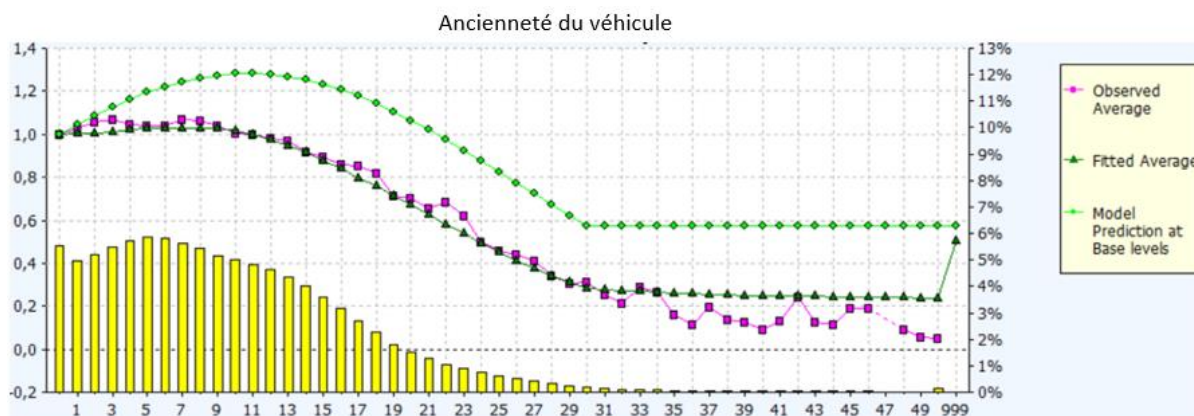


Figure 3.16 – Comparaison des fréquences RCM responsable prédites et observées selon l'ancienneté du véhicule

3.3.7.3 Validation du modèle

Test de significativité du modèle

Le rapport de la déviance sur le nombre de degrés de liberté est de 10%, bien inférieur à 1, ce qui permet de valider la pertinence du modèle au sens de ce critère.

Déviance	ddl	Déviance / ddl
961 330	9 640 059	10%

Table 3.5 – Rapport de la déviance du modèle de fréquence RCM responsable

Qualité d'ajustement global

Ce critère de qualité consiste en une comparaison en moyenne des valeurs prédites et observées. Le tableau ci-dessous montre une bonne qualité d'ajustement du modèle à la fois sur l'échantillon d'apprentissage et sur la base test puisque la fréquence moyenne prédite est très proche de la fréquence moyenne observée.

exercice	Apprentissage			Test		
	Fréquence estimée	Fréquence observée	écart (%)	Fréquence estimée	Fréquence observée	écart (%)
2013	2,99%	3,00%	-0,3%	3,01%	3,07%	-2,0%
2014	2,94%	2,90%	1,4%	2,95%	2,88%	2,4%
2015	2,90%	2,94%	-1,4%	2,91%	2,87%	1,4%
2016	2,88%	2,89%	-0,3%	2,87%	2,93%	-2,0%
2017	2,85%	2,86%	-0,3%	2,83%	2,84%	-0,4%
2018	2,82%	2,81%	0,4%	2,80%	2,79%	0,4%
Total	2,90%	2,90%	0,0%	2,89%	2,89%	0,0%

Table 3.6 – Erreur moyenne d'ajustement du modèle de fréquence RCM responsable

Indice de Gini

L'indice de Gini mesure le pouvoir de segmentation du modèle, des valeurs trop éloignées sur les échantillons d'apprentissage et de test peuvent traduire un sur-apprentissage du modèle. Ici les résultats sont rassurants car les valeurs de l'indice de Gini sont très proches sur les deux jeux de données.

Apprentissage	Test
0,2694	0,2718

Table 3.7 – Indice de Gini du modèle de fréquence RCM responsable

Courbe de lift

La courbe de lift est utilisée pour mesurer la qualité d'ajustement du modèle sur la base de test selon une distribution en quantile. Le graphique ci-dessous représente la fréquence prédite (en bleu) et la fréquence observée (en rouge) suivant une distribution en vingtile. Une bonne adéquation du modèle aux données est constatée lorsque les deux courbes sont proches voire superposées, ce qui est manifestement le cas ici. A noter toutefois une légère surestimation sur le dernier vingtile qui correspond aux 5% des risques avec la fréquence la plus élevée.

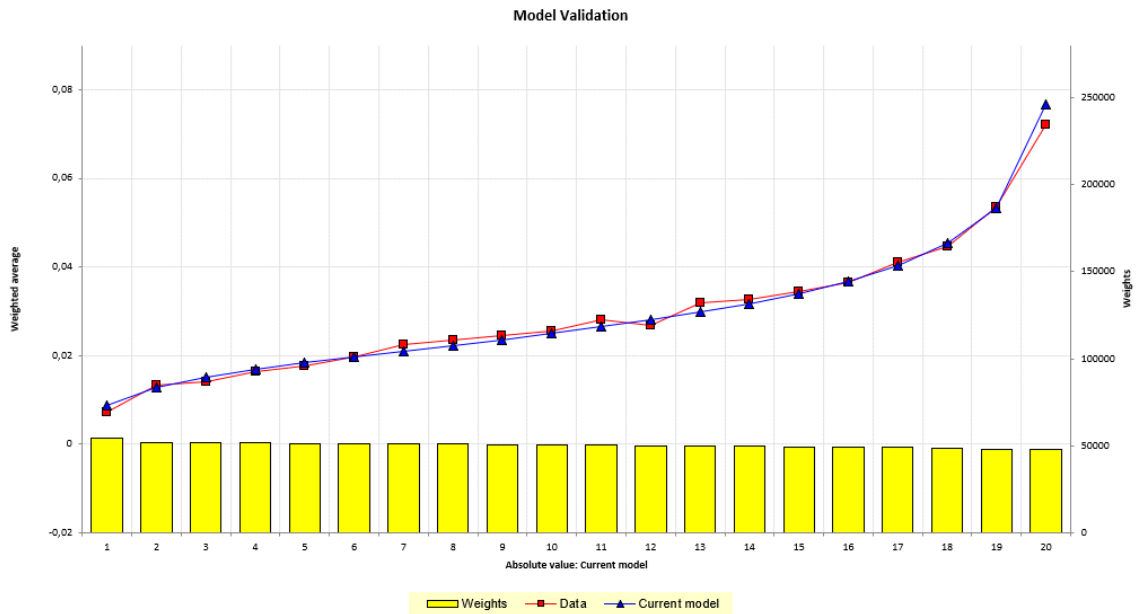


Figure 3.17 – Courbe de lift du modèle de fréquence RCM responsable

Robustesse du modèle

La robustesse du modèle est mesurée en comparant les coefficients estimés sur la base d'apprentissage de chaque variable à ceux obtenus sur la base de test. Des coefficients proches sur les deux jeux de données permettent de conclure à une stabilité des coefficients du modèle. Par exemple, le graphique 3.18 montre que les coefficients de l'ancienneté du véhicule estimés en apprentissage (en bleu) sont très proches de ceux obtenus en test (en vert clair).

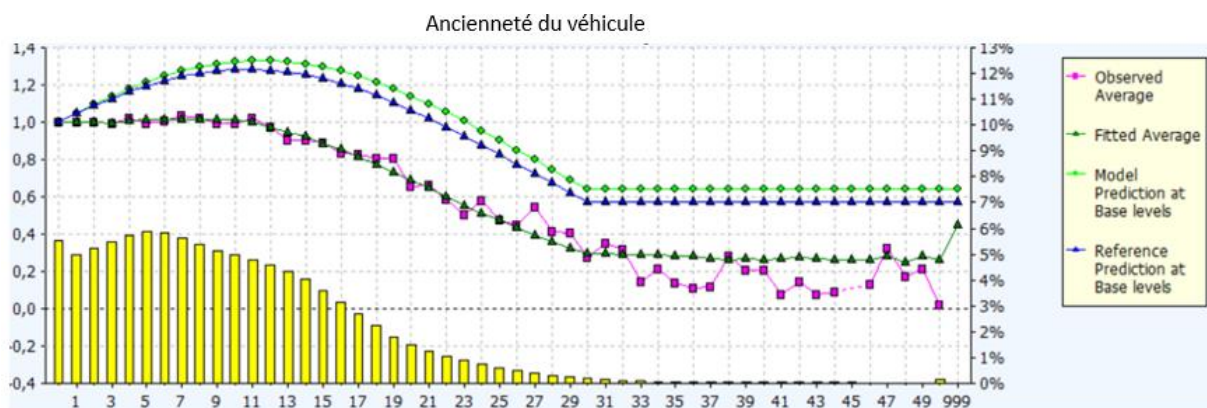


Figure 3.18 – Robustesse des estimateurs de l'ancienneté du véhicule du modèle de fréquence RCM responsable

Etude des résidus

L'étude des résidus consiste en un examen graphique des résidus standardisés de la déviance. Le graphique 3.19 représentant les résidus du modèle de fréquence RC matérielle responsable montre

un nuage de points homogène, de forme elliptique et centré autour de 0, ce qui permet de valider les hypothèses du modèle.

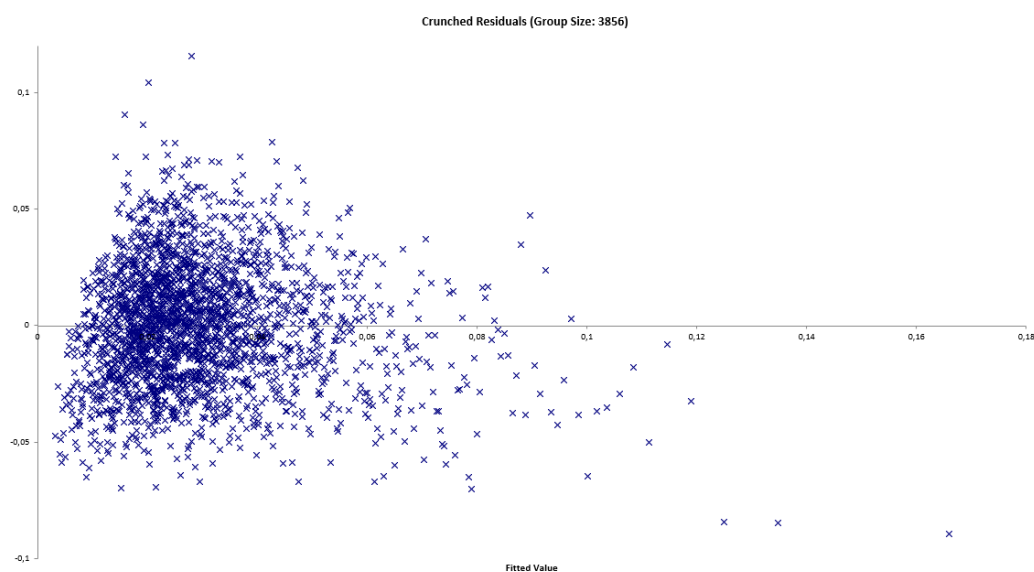


Figure 3.19 – Résidus standardisés de la déviance du modèle de fréquence RCM responsable

3.4 Modélisation du coût moyen

La modélisation du coût moyen RC porte sur la charge attritionnelle positive nette de recours en distinguant d'une part la composante matérielle et de l'autre la composante corporelle.

Comme précisé en section 2.3.2.2, le mode d'indemnisation impacte les modèles de coût moyen et en particulier le modèle de coût moyen matériel responsable du fait de la surreprésentation des sinistres à coût forfaitaire (forfaits IDA). Deux approches de la modélisation du coût moyen matériel responsable ont donc été testées, la première en séparant les sinistres IDA des non IDA (approche exhaustive) et la seconde en les regroupant (approche simplifiée). Dans l'approche exhaustive, seule la composante non IDA responsable fait l'objet d'une segmentation par des variables explicatives, le modèle IDA responsable se composant uniquement du montant du forfait IDA et de l'exercice de survenance.

Dans la suite de cette section, l'accent est mis sur la modélisation du coût moyen RC matériel responsable, les principaux résultats relatifs au modèle de coût moyen RC matériel non responsable est disponible en annexe C.

3.4.1 Détermination de la loi

De manière générale, l'ajustement aux données par une loi théorique est assez délicat à obtenir pour les modèles de coûts moyens en raison des coûts sur-représentés et seule la validation graphique est considérée car les tests d'adéquation sont quasi-systématiquement rejetés. Les deux lois les plus adaptées pour modéliser le coût moyen en assurance automobile sont la loi Gamma et la Log-normale.

Les analyses graphiques présentées ci-dessous concernent l'ajustement des données à une loi Gamma (plus appropriée aux données que la Log-normale) selon les deux approches testées. Et comme attendu, l'approche simplifiée fait apparaître une plus grande erreur d'ajustement consécutive à un pic de sinistres qui correspond aux coûts forfaitaires (forfaits IDA responsables).

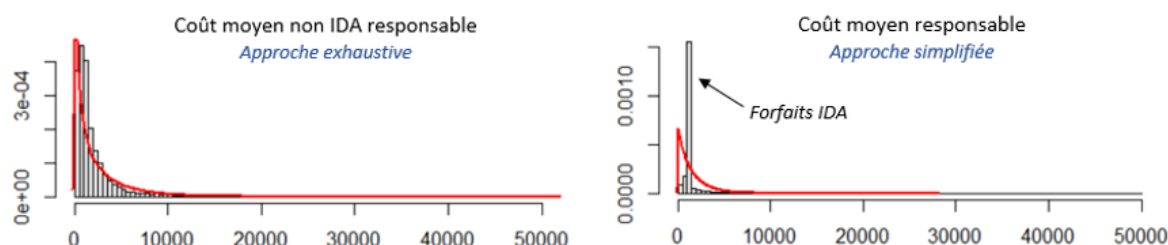


Figure 3.20 – Ajustement du coût moyen RCM responsable à une loi Gamma

3.4.2 Présentation du modèle

Les deux approches ont ensuite fait l'objet d'une analyse comparative de leur structure tarifaire et des coefficients associés à chacune des modalités. Concernant leur structure, elles sont identiques aussi bien sur le critère de la complexité (nombre de variables du modèle) que de la segmentation elle-même. En effet, les deux modèles se composent de 3 variables qui sont l'âge du conducteur, la classe de prix SRA et le type de formule souscrite.

De manière générale, les courbes de coefficients suivent la même tendance mais se différencient assez nettement au niveau de leur intensité, elle est plus modérée avec l'approche simplifiée. Ce dernier résultat pourrait s'expliquer par la répartition des coûts forfaitaires sur les différents profils de risque qui atténueraient l'amplitude des estimateurs.

Ci-dessous l'analyse comparative des approches détaillée par critère tarifaire :

L'âge du conducteur

Comparativement à l'approche exhaustive, le segment des jeunes conducteurs est davantage préservé avec l'approche simplifiée avec un impact compris entre -17% et -10% sur les 18-27 ans. Inversement, le segment des conducteurs seniors est plus touché avec un impact supérieur à +5% sur les 56 ans et plus.

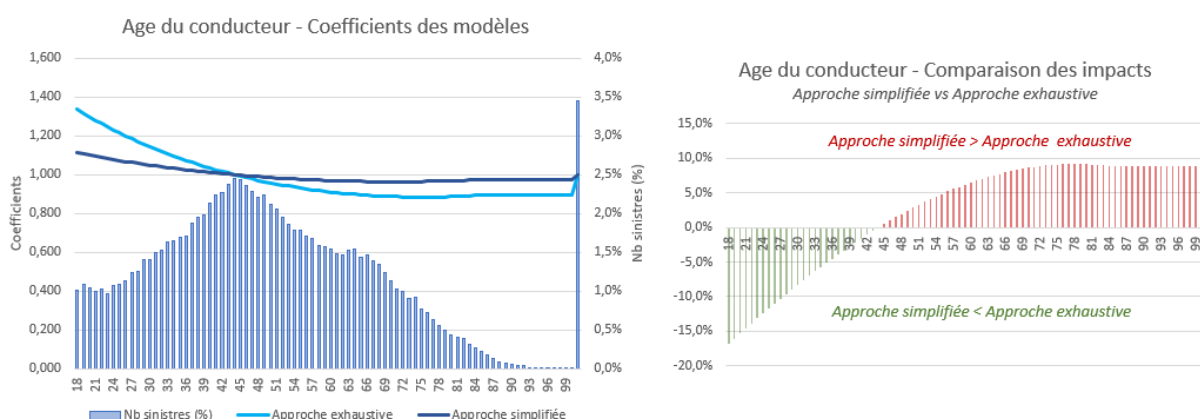


Figure 3.21 – Comparaison des approches suivant l'âge du conducteur

La classe de prix SRA

De la même manière que pour l'âge du conducteur, les différences sont visibles aux deux extrémités de la courbes des coefficients. L'approche simplifiée se caractérise par un impact plus graduel suivant la montée en gamme du véhicule.

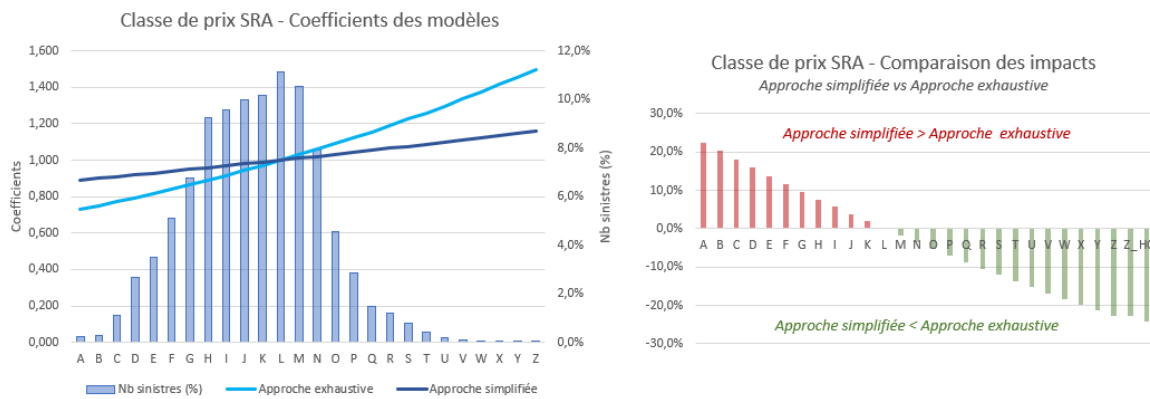


Figure 3.22 – Comparaison des approches suivant la classe de prix SRA

Le type de formule

L'influence de la formule dans le modèle de coût moyen est beaucoup moins prononcée avec l'approche simplifiée avec une amplitude de seulement +11% entre la formule Tiers mini et la Tous risques contre +50% avec l'approche exhaustive.

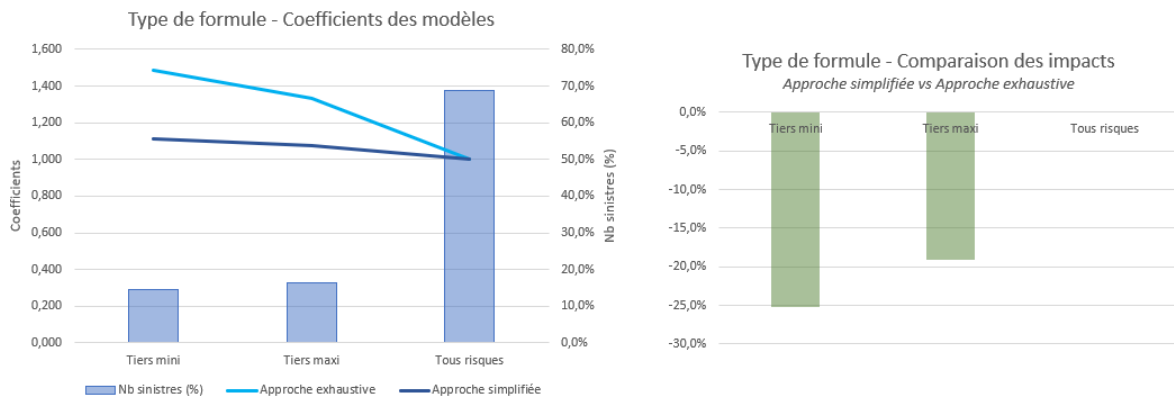


Figure 3.23 – Comparaison des approches suivant le type de formule

3.4.3 Performance du modèle

Les deux approches ont enfin fait l'objet d'une comparaison de leur performance par le biais de l'examen des courbes de lift établies sur la base de test. Et sur ce critère, c'est l'approche simplifiée qui apparaît la plus ajustée aux données au regard de la plus grande proximité entre les prédictions (courbe bleue) et les observations (courbe rouge) comme le montrent les graphiques 3.24 et 3.25 ci-dessous.

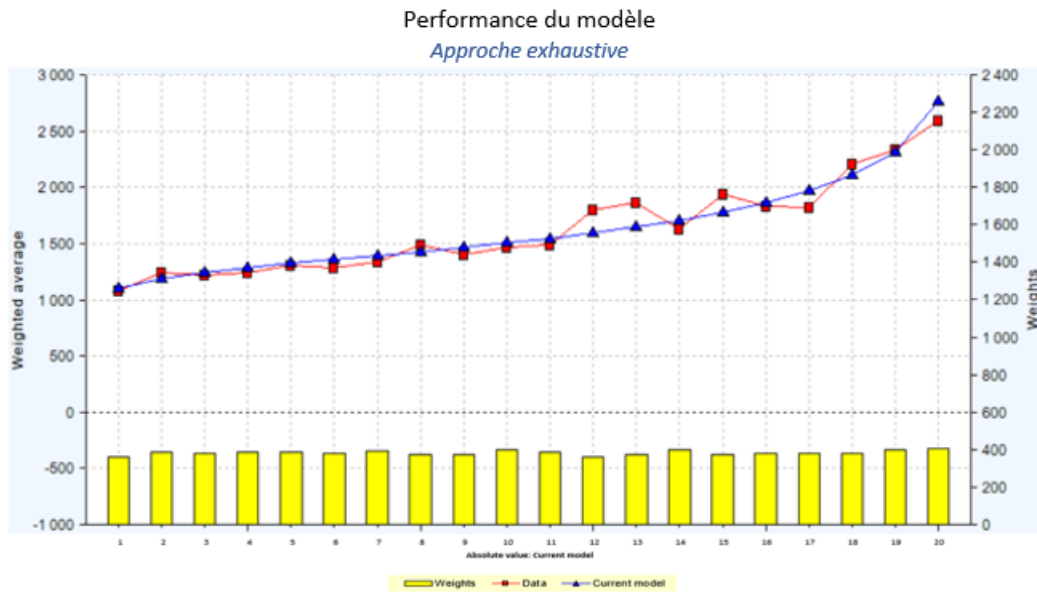


Figure 3.24 – Approche exhaustive – Courbe de lift du modèle de coût moyen RCM responsable

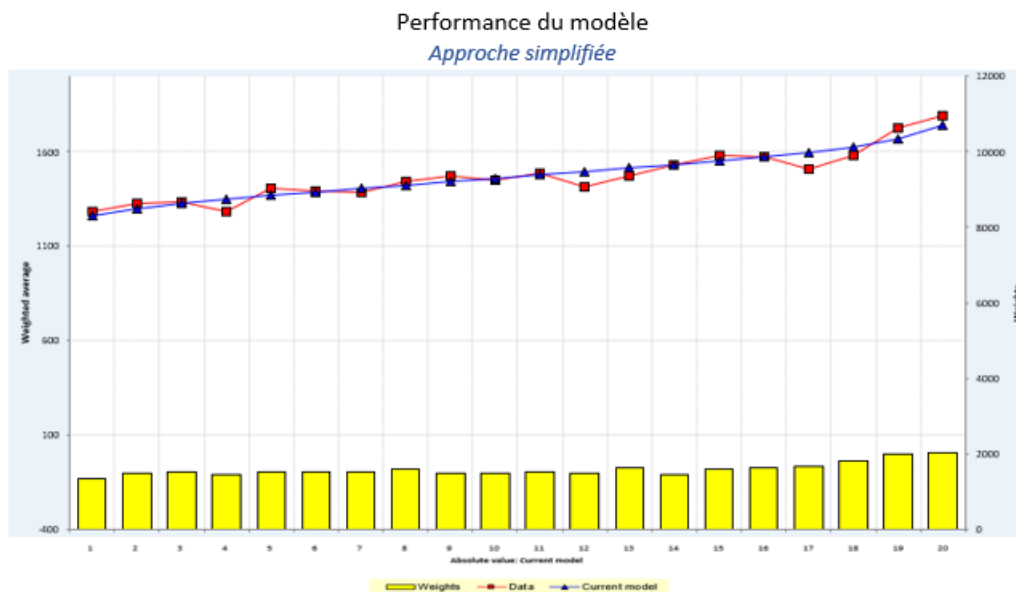


Figure 3.25 – Approche simplifiée – Courbe de lift du modèle de coût moyen RCM responsable

En résumé, bien que l'approche exhaustive soit la plus adaptée aux hypothèses du modèle (meilleur ajustement de la distribution des sinistres à une loi théorique), les deux approches ont finalement abouti à la même segmentation tarifaire avec un avantage sur le critère de la performance pour l'approche simplifiée qui pourrait s'expliquer par une plus grande robustesse des estimateurs par un effet volume de données. Compte-tenu de ces éléments, **le choix s'est porté sur l'approche simplifiée pour modéliser le coût moyen de la composante matérielle.**

Critère de décision	Approche	
	Exhaustive	Simplifiée
Adéquation des hypothèses		
Pertinence de la segmentation		
Performance du modèle		
Combinaison des sous-modèles		

Table 3.8 – Comparatif des modèles de coût moyen RCM responsable selon l’approche

3.4.4 Validation du modèle

Qualité d’ajustement global

Le tableau 3.9 montre une bonne qualité d’ajustement du modèle de coût moyen matériel responsable avec un coût moyen estimé très proche de l’observé aussi bien en apprentissage qu’en test, ce qui vient confirmer les précédents résultats.

exercice	Apprentissage			Test		
	Coût moyen estimé	Coût moyen observé	écart (%)	Coût moyen estimé	Coût moyen observé	écart (%)
2013	1 259 €	1 277 €	-1,4%	1 265 €	1 274 €	-0,7%
2014	1 309 €	1 304 €	0,4%	1 314 €	1 327 €	-0,9%
2015	1 363 €	1 349 €	1,1%	1 369 €	1 334 €	2,6%
2016	1 420 €	1 393 €	1,9%	1 430 €	1 428 €	0,1%
2017	1 478 €	1 493 €	-1,0%	1 486 €	1 485 €	0,1%
2018	1 537 €	1 551 €	-0,9%	1 546 €	1 563 €	-1,1%
Total	1 396 €	1 396 €	0,0%	1 403 €	1 403 €	0,0%

Table 3.9 – Erreur moyenne d’ajustement du modèle de coût moyen RCM responsable

Robustesse du modèle

La robustesse du modèle de coût moyen est également vérifiée sur chaque critère tarifaire du modèle. Ci-dessous le résultat de l’analyse concernant la classe de prix du véhicule démontrant la robustesse des estimateurs estimés sur la base d’apprentissage.

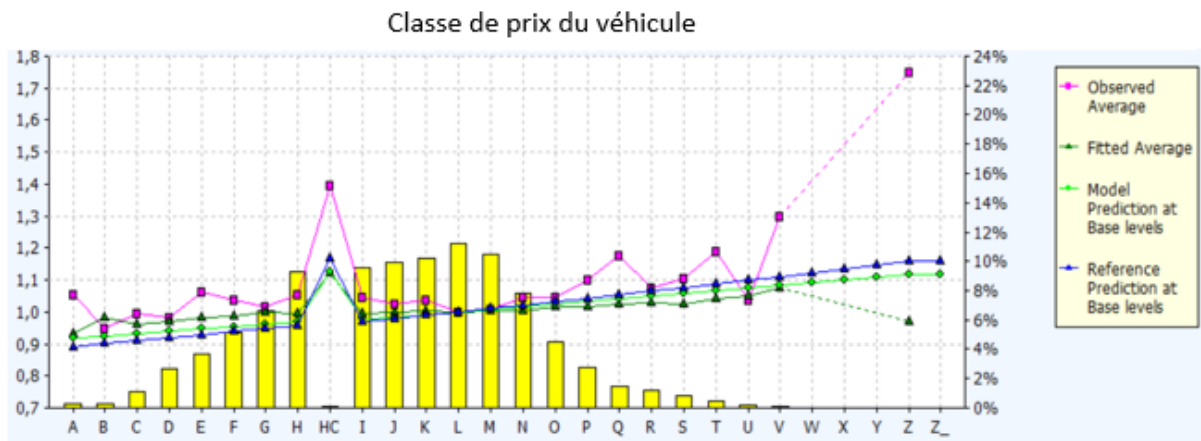


Figure 3.26 – Robustesse des estimateurs de la classe de prix SRA du modèle de coût moyen RCM responsable

Etude des résidus

Sur le graphique 3.27, le nuage de points des résidus est bien concentré sur l'intervalle $[-2, 2]$ et sa structure ne présente pas de forme particulière. De plus, l'histogramme des résidus donné par le graphique 3.28 montre clairement une concentration autour de 0 et une distribution proche de la loi normale centrée réduite.

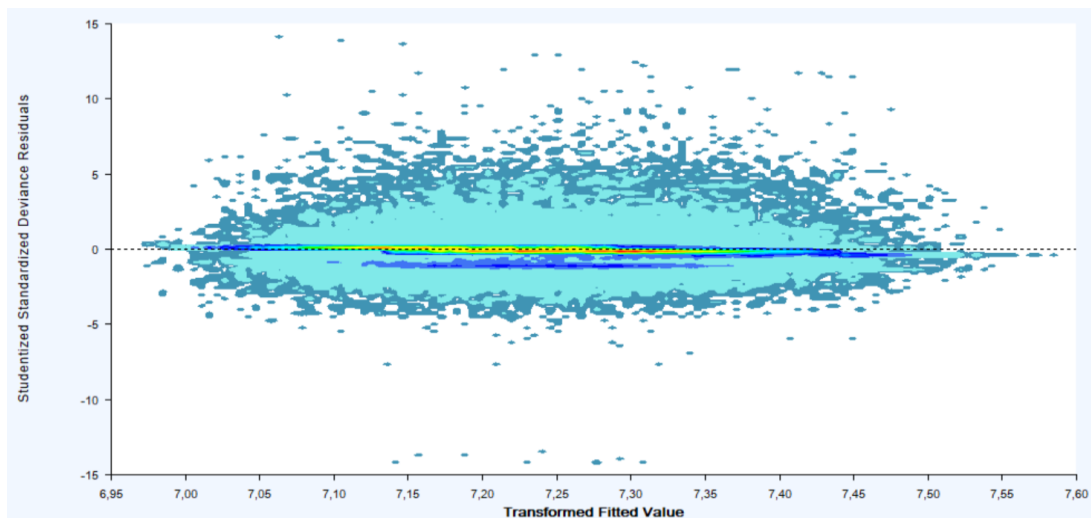


Figure 3.27 – Résidus de la déviance du modèle de coût moyen RCM responsable

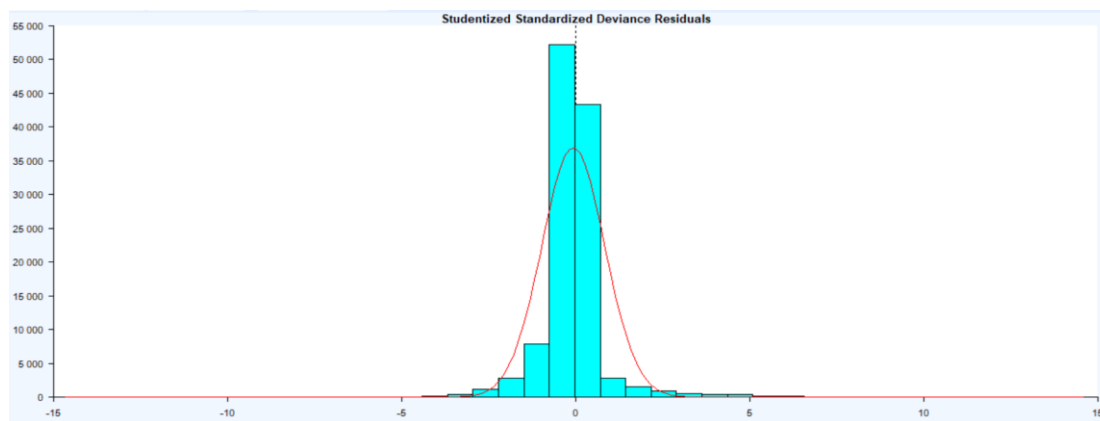


Figure 3.28 – Histogramme des résidus du modèle de coût moyen RCM responsable

3.5 Modèle de prime pure

Par décomposition de la charge globale RC selon ses composantes naturelles distinctes (partie matérielle et corporelle) et par linéarité de l'espérance, la prime pure globale de la garantie RC est le résultat de la combinaison de ses sous-modèles (graphique 3.29) et s'écrit sous la forme :

$$PP_{globale\ RC} = E(S_{mat+corp}) = E(S_{mat}) + E(S_{corp}) = PP_{mat} + PP_{corp}$$

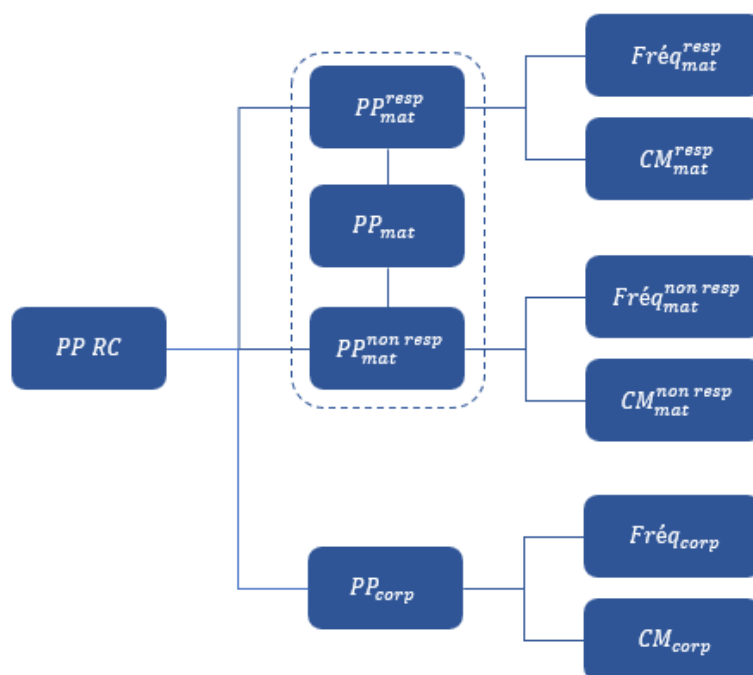


Figure 3.29 – Elaboration du modèle de prime pure de la garantie RC

Plus précisément,

$$PP_{globale RC} = \left(PP_{mat}^{resp} \times (1 + \alpha) + PP_{mat}^{non resp} \times (1 + \beta) + PP_{corp} \times (1 + \gamma) \right) \times (1 + \delta)$$

Avec :

- $PP = Fréq \times CM$ pour chaque composante (modélisation séparée de la fréquence et du coût moyen)
- α, β, γ le coefficient de redressement des charges modélisées de chaque composante égal au rapport de la charge observée sur la charge attritionnelle modélisée. La valeur de ces différents coefficients est donnée ci-dessous (tableau 3.10). Il est à noter un coefficient de redressement inférieur à 1 pour la composante matérielle non responsable du fait de la part importante de sinistres à coût négatif dans la charge observée correspondant aux recours IDA perçus.

α	1,0085
β	0,2949
γ	1,2790

Table 3.10 – Coefficient de redressement de la prime pure par composante

- δ le coefficient d'ajustement global égal au rapport de la charge totale observée sur la charge totale prédite. Cet ajustement est calculé sur l'ensemble des données et a pour but de corriger les écarts générés par la combinaison des sous-modèles. Sa valeur est inférieure à 1% (tableau 3.11), ce qui indique une bonne calibration du modèle de prime pure.

Charge observée	580 783 831 €
Charge totale modélisée	586 324 599 €
Coefficient d'ajustement	0,9905

Table 3.11 – Coefficient d'ajustement global de la charge modélisée

Enfin, la dernière étape consiste à ajuster un GLM au modèle de prime pure combiné afin d'obtenir un ensemble unique de coefficients pour chaque facteur et une structure multiplicative du tarif. Ces différentes étapes ont été réalisées à l'aide du *Model Combiner* du logiciel Emblem.

Le modèle final de prime pure ainsi obtenu prend la forme :

$$Prime Pure = \alpha \sum_i \beta_i X_i$$

Où :

- α représente la prime pure de base.
- X_i sont les facteurs de risque qui influent sur le modèle.
- β_i sont les coefficients associés à chaque facteur de risque.

Chapitre 4

Analyse des résultats et comparaison avec l'existant

Ce dernier chapitre s'intéresse à l'étude approfondie du modèle de prime pure et à sa pertinence selon le canal de distribution. A cet effet, la première partie de ce chapitre sera dédiée à l'ajustement, la compréhension et l'étude de la performance du modèle de prime pure. Ces différentes analyses seront menées au global mais également à un niveau plus fin en parcourant une variété de critères tarifaires.

Dans un second temps, une comparaison avec le modèle existant sera réalisée afin de mieux comprendre ce qui les différencie. Cette étude comparative portera 3 principaux thèmes que sont la structure des modèles, l'influence des variables tarifaires et le zonier.

Enfin, une dernière partie sera consacrée à l'évaluation de la pertinence du modèle selon le canal de distribution, l'objectif étant de juger de la capacité du nouveau modèle de prime pure à tenir la promesse de l'omnicanalité et d'un tarif unique quel que soit le canal sollicité.

4.1 Ajustement du modèle de prime pure

Le modèle de prime pure a fait l'objet d'ajustements a posteriori gérés directement dans la modélisation afin de prendre en compte les corrélations entre les variables tarifaires et de conserver l'équilibre technique du modèle.

Ces ajustements concernent deux points en particulier :

- La cohérence des impacts tarifaires liés aux antécédents de sinistres à la souscription
- La valorisation de la clause de conduite exclusive qui avait été laissée volontairement de côté lors de la phase de construction des modèles.

4.1.1 Les antécédents de sinistres à la souscription

Les antécédents de sinistres font parties des éléments obligatoires figurant sur le relevé d'informations et permettent de retracer l'historique de sinistralité du conducteur sur une période donnée (généralement sur les 24-36 derniers mois). Lors de la souscription d'un contrat d'assurance automobile, ces informations sont utilisées par l'assureur pour accepter ou refuser le risque mais également pour adapter la prime d'assurance en appliquant une majoration plus ou moins importante en présence d'antécédents de sinistres responsables ou non. Mais cet impact tarifaire est généralement plus important en cas de présence d'antécédents responsables.

Ci-dessous le résultat de l'ajustement des antécédents à la souscription qui permet de corriger une incohérence tarifaire entre les antécédents responsables et non responsables, l'effet de la variable

des antécédents responsables dans le modèle initial étant en grande partie capté par le CRM. Cet ajustement consiste d'une part à réduire l'impact des antécédents non responsables et d'autre part à accentuer celui des responsables en s'assurant en même temps d'une cohérence tarifaire.

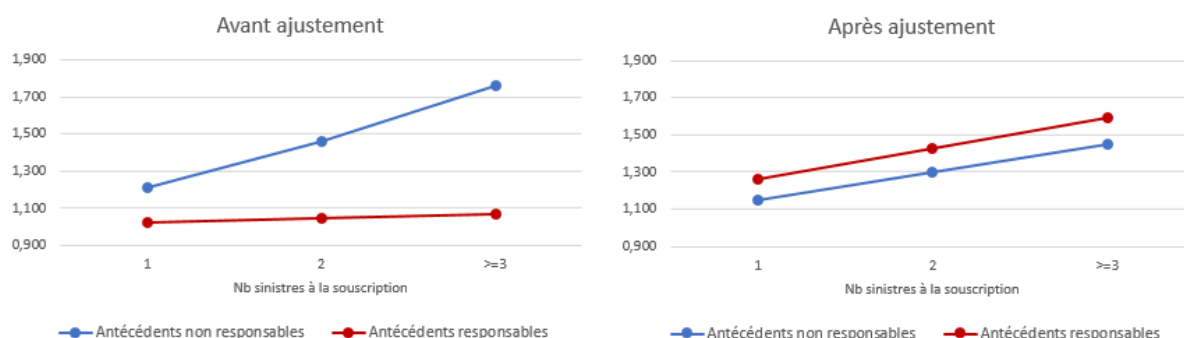


Figure 4.1 – Ajustement des antécédents de sinistres à la souscription

4.1.2 La clause de conduite exclusive

Le second ajustement est de type commercial et consiste à valoriser la réduction tarifaire accordée en cas de souscription à l'option de la conduite exclusive. Pour rappel, cette clause limite l'usage du véhicule aux seuls conducteurs désignés aux contrats (conducteur principal et secondaire) et de ce fait restreint le risque de survenance d'un sinistre pour l'assureur.

L'impact de la conduite exclusive est déterminé en comparant la prime pure moyenne des assurés en conduite libre avec celle des assurés en conduite exclusive. Le gain de sinistralité associé permet de valoriser la clause de conduite exclusive représentant ici une réduction de 10%.

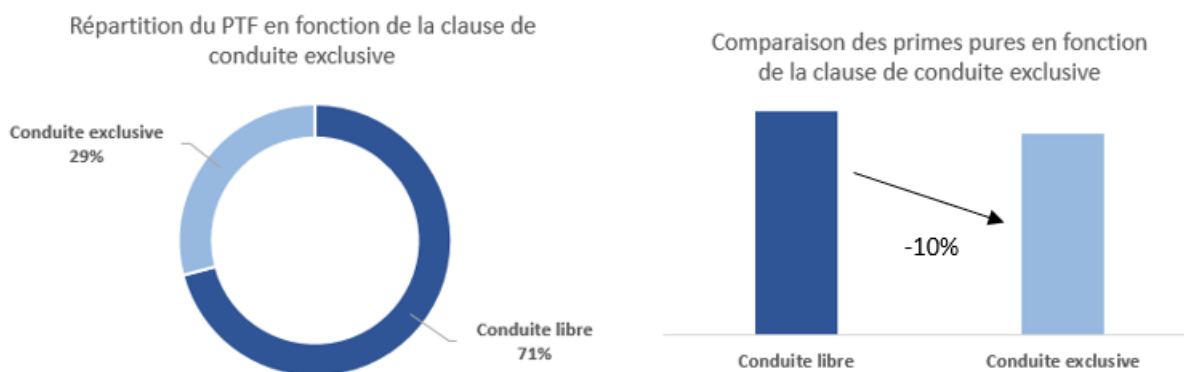


Figure 4.2 – Valorisation de l'impact de la conduite exclusive

4.2 Analyse du modèle de prime pure

4.2.1 Contributions des variables au modèle

Comme pour les modèles de fréquence et de coût moyen, une restitution des résultats par l'analyse des contributions des variables est présentée ci-dessous. Ces contributions sont ici calculées par la méthode des relativités qui estime le gain de chaque variable en considérant la relativité et la

volumétrie de chacune de ses modalités. La contribution d'une variable au modèle est donc fonction de l'amplitude de ses coefficients tarifaires et de l'exposition associée.

Le graphique 4.3 précise l'importance relative de chaque variable au modèle de prime pure et montre une certaine cohérence dans la hiérarchie des variables les plus influentes. En effet, en tête du classement, le forfait kilométrique qui exprime une exposition au risque en quantifiant le nombre de kilomètres annuels parcourus, le zonier qui permet de capter l'environnement géographique et enfin le CRM qui reflète l'expérience de sinistralité du conducteur. En queue de classement, deux variables à moins de 1%, il s'agit du mode d'achat du véhicule et des antécédents de sinistres responsables à la souscription.

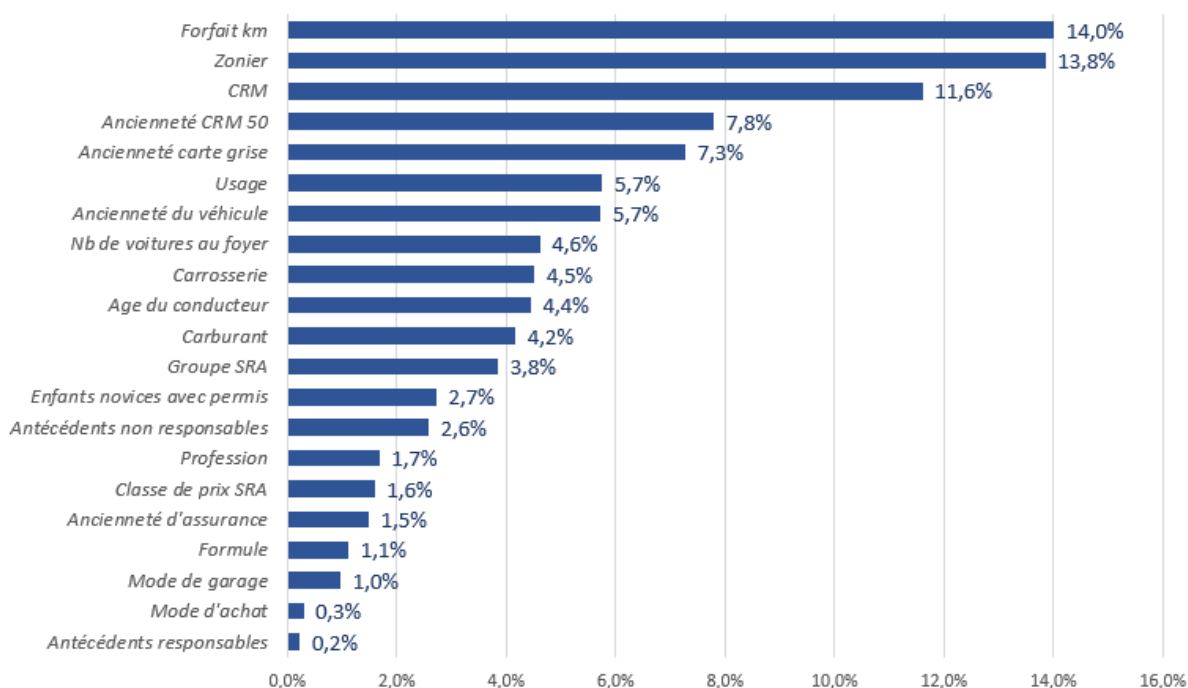


Figure 4.3 – Contributions des variables au modèle de prime pure

Le mode d'achat

Les véhicules en LOA/LLD représentent seulement 3,6% du portefeuille d'étude (graphique 4.4), d'où la faible contribution apparente de cette variable (0,3%). Cependant, la décision de conserver cette variable est stratégique, elle est motivée par les récentes évolutions du marché automobile (section 1.1.3) et la forte croissance de ce mode de financement.

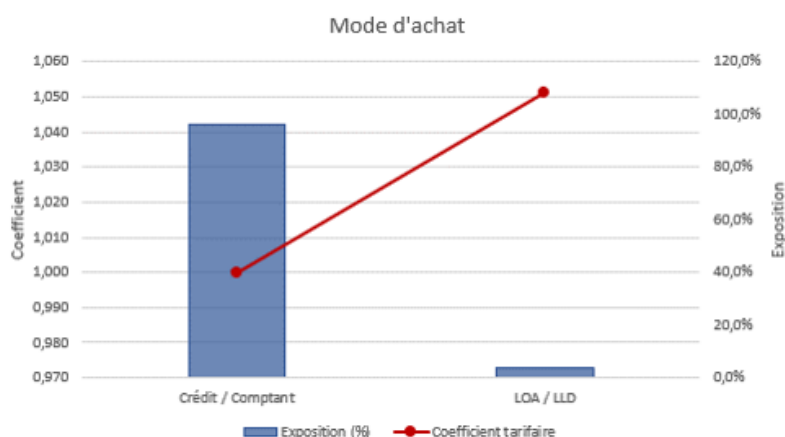


Figure 4.4 – Impact du mode d'achat dans le modèle de prime pure

Les antécédents responsables

Malgré une contribution négligeable (0,2%), cette variable est également conservée dans le modèle car elle permet de majorer davantage les assurés avec des antécédents responsables (cohérence tarifaire avec les non responsables) et de lutter contre l'antisélection.

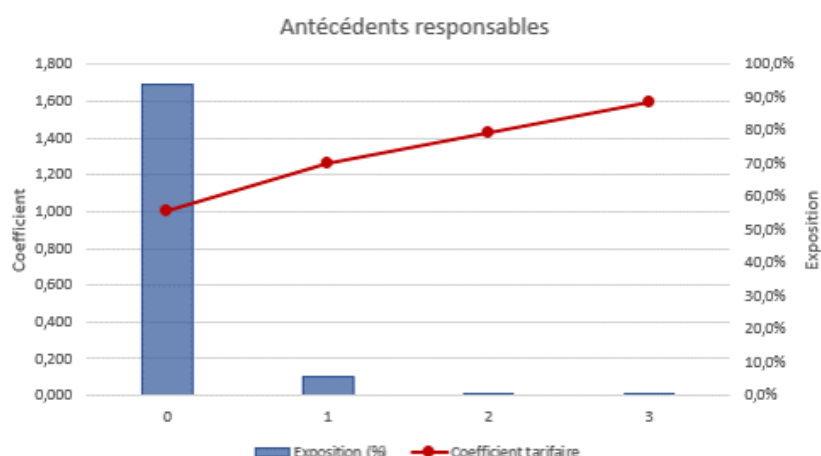


Figure 4.5 – Impact des antécédents responsables dans le modèle de prime pure

4.2.2 Pouvoir de segmentation des variables tarifaires

L'analyse des contributions est complétée par une mesure du pouvoir de segmentation de chaque variable. Elle se base sur l'amplitude des coefficients et le *spread* associé (écart entre le coefficient le plus élevé et le plus petit). Ces deux indicateurs sont représentés dans le graphique 4.6 ci-dessous de la manière suivante :

- L'amplitude des coefficients s'apprécie graphiquement en considérant la longueur du segment dont les extrémités correspondent aux coefficients min et max de la variable.
- Le niveau du spread est indiqué directement sur le graphique (au-dessus de la borne max de la variable considérée).

Par exemple, pour la variable de l'âge du conducteur, l'amplitude des coefficients varie entre 0,949 pour la borne min et 1,488 pour la borne max, soit un spread de 57%. Ce graphique permet d'identifier rapidement les variables les plus segmentantes (spreads les plus élevés), à savoir l'ancienneté du véhicule (spread de 123%), le zonier (spread de 94%) et le forfait kilométrique (spread de 73%). A l'opposé, les variables qui segmentent le moins la prime pure sont le mode de garage et la formule souscrite (spread de 4%).

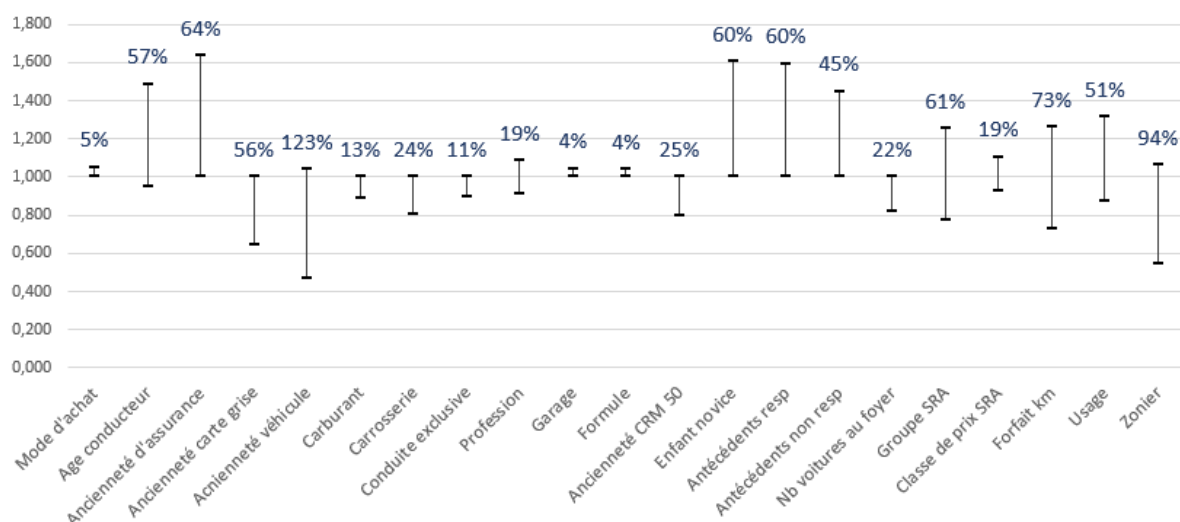
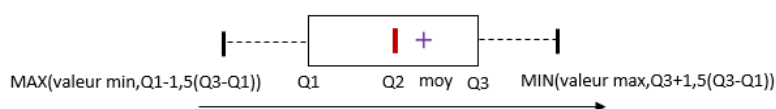


Figure 4.6 – Amplitude et spread des variables du modèle de prime pure

4.2.3 Distribution de la prime pure

Une représentation synthétique des principales caractéristiques de la prime pure modélisée est donnée par le diagramme en boîte à moustaches. Cet outil permet d'en étudier la distribution et d'identifier d'éventuelles points atypiques qui nécessiteraient un ajustement du tarif. Cette étude est réalisée au global pour une analyse macro et par critère tarifaire pour une analyse comparative par segment.

Aide à la lecture des boîtes à moustaches (rappel) :



4.2.3.1 Analyse de la distribution au global

Le graphique 4.7 montre une distribution légèrement asymétrique à droite de la prime pure modélisée puisque la médiane (~100€) est plus proche du 1^{er} quartile (70€) que du 3^{ème} quartile (140€). Ce phénomène s'explique par la présence de profils à haut risque cumulant les critères les plus défavorables (sans forfait kilométrique, usage tournée, zone géographique la plus risquée, véhicule récent et haut de gamme, ...) et tirant la moyenne vers le haut.

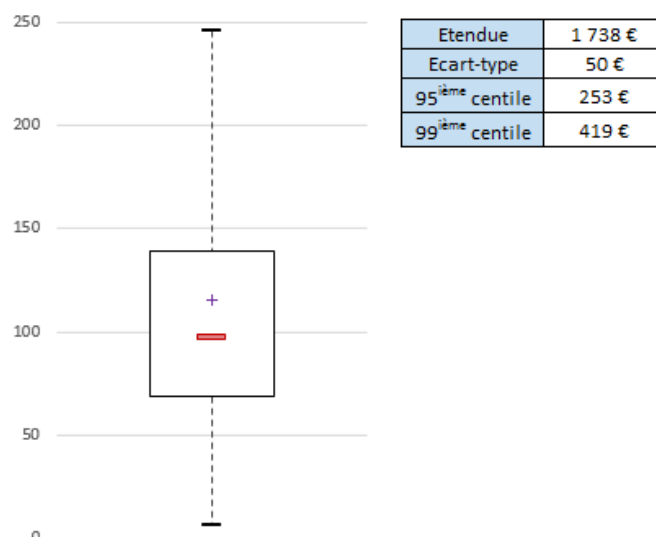


Figure 4.7 – Distribution de la prime pure modélisée

4.2.3.2 Analyse de la distribution par variable

Un autre grand intérêt de ces diagrammes en boîte à moustaches est de pouvoir estimer et comparer facilement les distributions entre chaque population et de mesurer la variabilité du modèle à l'intérieur d'un segment donné.

Une analyse de la distribution sur quelques critères est donnée ci-dessous :

L'âge du conducteur

Le graphique 4.8 met clairement en évidence une forte variabilité de la prime pure modélisée chez les jeunes conducteurs de moins de 21 ans (intervalle interquartile très étalé) et à tendance à diminuer en fonction de l'âge du conducteur. Pour chaque segment, la distribution semble relativement symétrique, les moyennes étant très proches des médianes et de manière cohérente, la prime pure moyenne est décroissante en fonction de l'âge du conducteur.

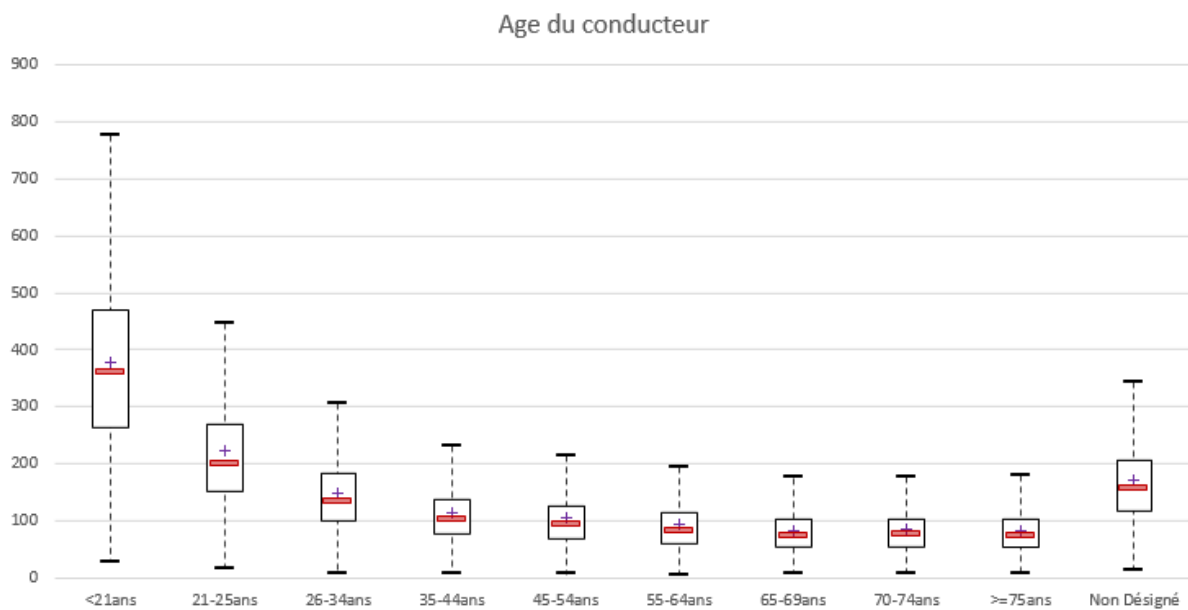


Figure 4.8 – Distribution de la prime pure modélisée en fonction de l'âge du conducteur

La zone géographique

Le graphique 4.9 montre une variabilité croissante de la prime pure selon la zone géographique avec une dispersion plus importante pour les zones les plus risquées. Les primes pures moyennes vont également dans ce sens, elles augmentent avec le niveau de risque de la zone géographique.

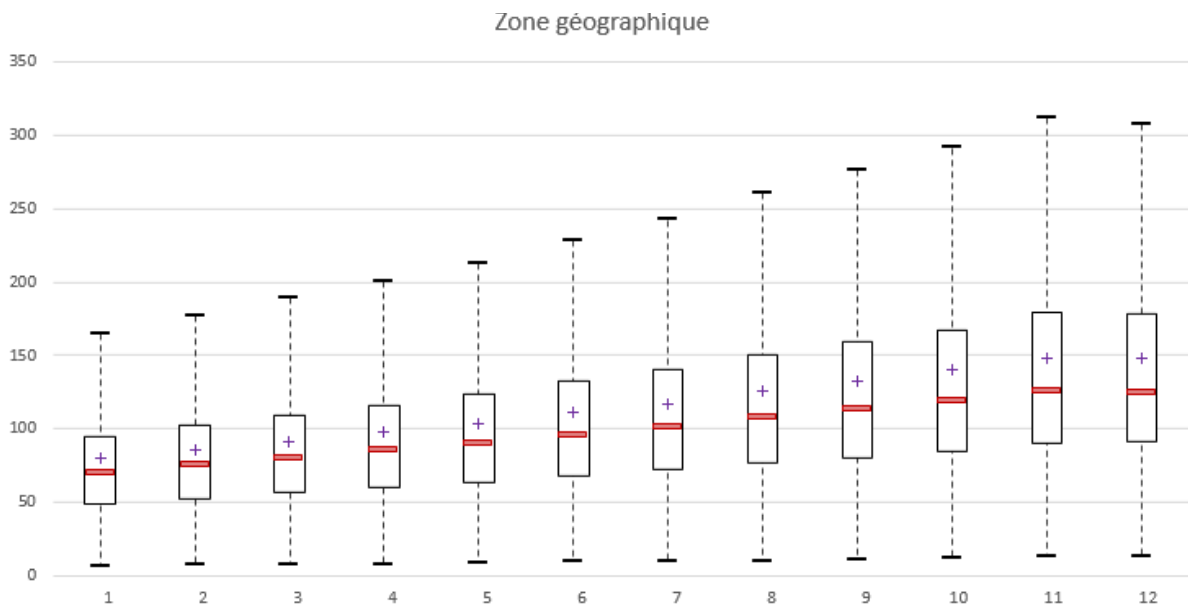


Figure 4.9 – Distribution de la prime pure modélisée en fonction de la zone géographique

L'usage du véhicule

La distribution de la prime pure selon l'usage du véhicule semble suivre une certaine logique, elle apparaît de plus en plus étalée selon le degré d'utilisation du véhicule avec une distribution moins étendue pour l'usage « Vie privée » et plus dispersée pour l'usage « Tournées ». Cette cohérence se retrouve dans l'évolution de la prime pure moyenne qui suit également la même tendance.

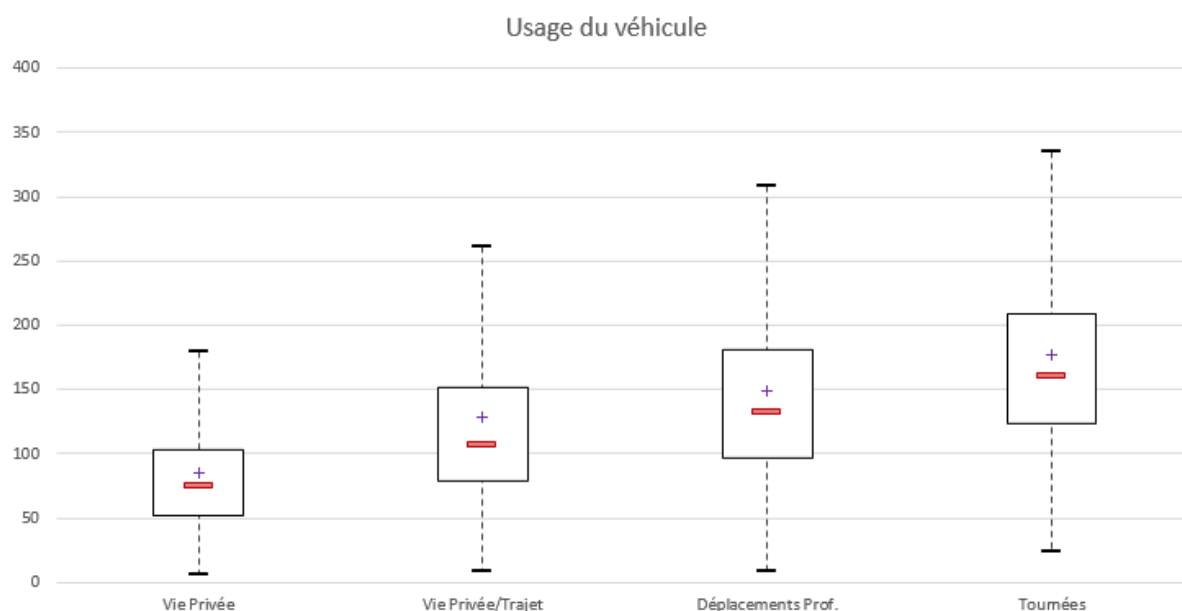


Figure 4.10 – Distribution de la prime pure modélisée en fonction de l'usage du véhicule

4.3 Performance du modèle de prime pure

4.3.1 Qualité prédictive du modèle

La qualité prédictive du modèle est évaluée sur l'échantillon test représentant 20% du portefeuille d'étude.

Les indicateurs utilisés pour mesurer la performance sont :

- Le S/PPM qui correspond au ratio de la charge de sinistres observée sur la prime pure modélisée. Une sous-estimation du modèle est constatée si le S/PPM est supérieur à 1 et à l'opposé, une surestimation est observée si le S/PPM est inférieur à 1.
- L'indice de Gini qui renseigne sur le pouvoir de discrimination du modèle.

Une comparaison de ces indicateurs entre les échantillons d'apprentissage et de test permet de juger de la fiabilité et de la robustesse du modèle.

Indicateurs	Apprentissage	Test
Poids	80%	20%
S/PPM	101,30%	94,78%
Indice Gini	36,7%	35,6%

Table 4.1 – Indicateurs de performance du modèle de prime pure

Alors que le S/PPM est quasiment à l'équilibre sur l'échantillon d'apprentissage (101,30%), un écart significatif est constaté sur l'échantillon test avec un S/PPM de 94,78%, soit une surestimation de la charge de sinistres observée de 5,50%. L'analyse des résultats par composante du risque (graphique 4.11) montre que cet écart ne concerne que la composante corporelle avec un S/PPM de 91,3%. Cette surestimation de la charge corporelle s'explique par une répartition inégale de la charge de graves entre les deux échantillons (22,7% en apprentissage vs 12,5% en test). Pour rappel, la charge corporelle de graves est mutualisée entre les différents risques. Quant à l'ajustement sur la partie matérielle, il est excellent aussi bien sur l'échantillon d'apprentissage qu'en test avec un écart négligeable. Ce dernier résultat vient appuyer le choix de l'approche retenue pour modéliser la composante matérielle.

Enfin, concernant l'indice de Gini, il est comparable sur les deux échantillons, ce qui traduit un bon compromis entre complexité et pouvoir de généralisation du modèle.

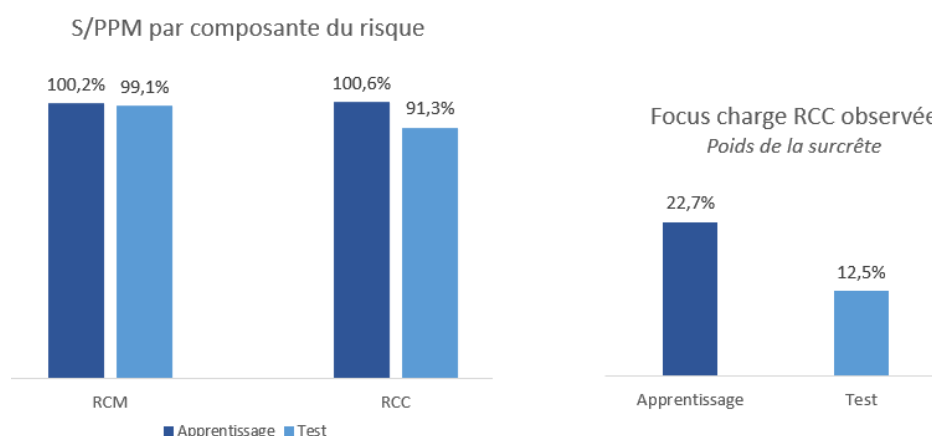


Figure 4.11 – Performance du modèle de prime pure par composante

4.3.2 Analyse des résultats par critère tarifaire

L'étude de la performance du modèle se termine par une analyse détaillée des résultats par variable tarifaire suivant les deux échantillons. Comme précédemment, la métrique utilisée est celle du S/PPM en distinguant chaque composante (RCM pour la partie matérielle et RCC pour la partie corporelle). D'autre part, en raison de la répartition inégale de la charge de graves entre les deux échantillons, il convient d'être prudent dans l'analyse de l'ajustement sur la composante corporelle, les enseignements tirés de l'échantillon d'apprentissage ne se vérifiant pas toujours sur l'échantillon test.

L'âge du conducteur

Le modèle est très bien ajusté sur la partie matérielle puisque le S/PPM RCM est à l'équilibre sur l'ensemble des classes d'âge et sur les deux échantillons. L'analyse sur la partie corporelle est plus délicate de par sa nature très volatile avec des résultats qui diffèrent selon l'échantillon considéré. Par exemple, la sinistralité des jeunes de 21-25 ans, fortement dégradée par une charge corporelle grave importante (poids des graves de 54%), semble sous-estimée en apprentissage mais surestimée sur l'échantillon test.



Figure 4.12 – Ajustement du modèle de prime pure selon l'âge du conducteur

L'ancienneté de la carte grise

Les résultats sont très satisfaisants sur le segment des cartes grises de moins de 9 ans qui représentent 84% des expositions. Entre 9-14 ans, le modèle a tendance à surestimer légèrement la prime pure observée. Au-delà de 14 ans, les résultats ne sont plus significatifs du fait d'une volumétrie trop faible.

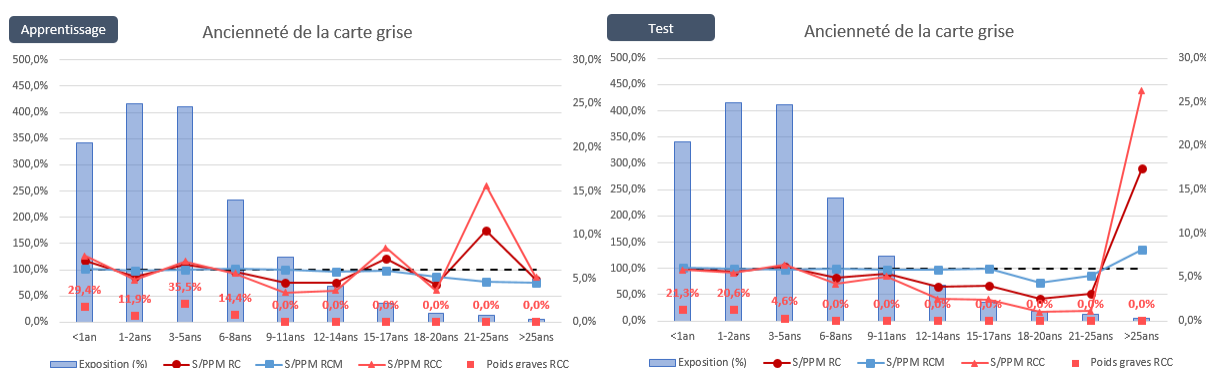


Figure 4.13 – Ajustement du modèle de prime pure selon l'ancienneté de la carte grise

Le mode de carburant

Le modèle est bien ajusté sur les principaux types de carburant (« Essence », « Turbo diesel » et « Diesel » dans une moindre mesure). La faible volumétrie des autres modalités (« Hybride », « Electrique » et « Autre ») ne permet pas de conclure sur les résultats de la modélisation. Faute d'une volumétrie suffisante, ces types de carburant ont été regroupés avec la modalité « Essence ».

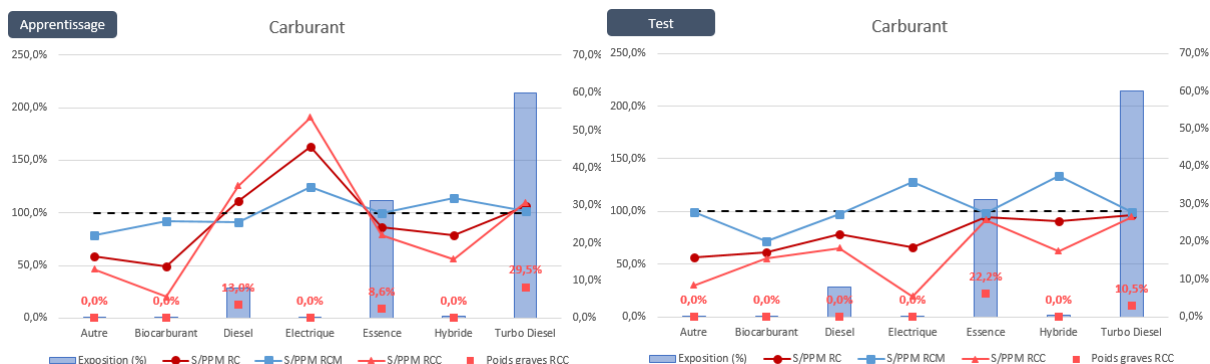


Figure 4.14 – Ajustement du modèle de prime pure selon le carburant

Les forfaits kilométriques

De manière générale, l'ajustement du modèle est bon sur les principaux forfaits kilométriques et plus particulièrement sur le forfait illimité (49% des expositions) avec un alignement parfait de la prime pure modélisée avec l'observée. Une surestimation du forfait 22 000 km (4% des expositions) est à noter, elle s'explique par une simplification de la courbe des coefficients par une droite linéaire.

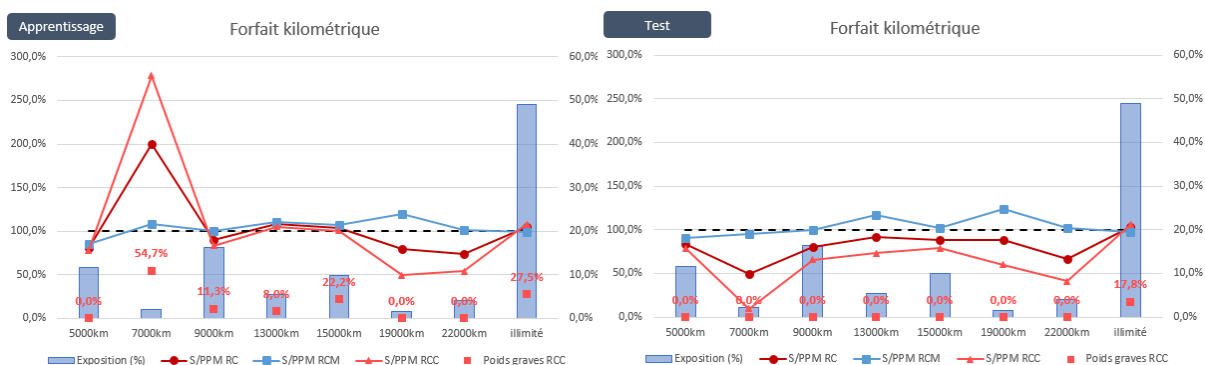


Figure 4.15 – Ajustement du modèle de prime pure selon le forfait kilométrique

La zone géographique

Le modèle ne semble pas très bien ajusté sur cette variable et en particulier en ce qui concerne le risque corporel. Ceci étant, ce résultat n'est pas très surprenant car le zonier a été élaboré à partir des résidus du modèle de fréquence matérielle, il serait donc plus judicieux de se concentrer sur la partie matérielle. Et à ce niveau, les résultats sont beaucoup plus rassurants puisque l'ajustement sur le risque matériel est bon malgré des écarts sur les deux zones extrêmes (surestimation de la zone 1 et sous-estimation de la zone 12).

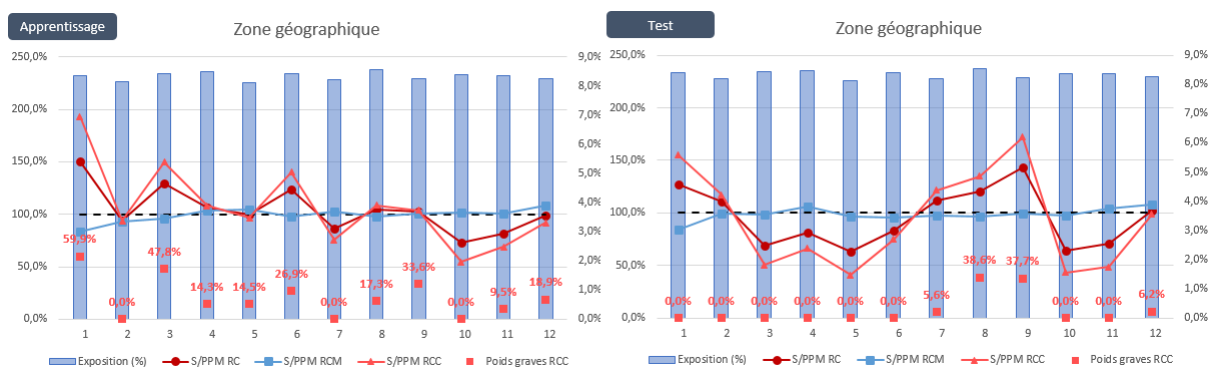


Figure 4.16 – Ajustement du modèle de prime pure selon la zone géographique

Les antécédents responsables

Les profils de risque avec antécédents responsables sont fortement surestimés par le modèle aussi bien en apprentissage qu'en test. Cet effet résulte de l'ajustement a posteriori du modèle de prime pure visant une cohérence tarifaire entre les antécédents de sinistre à la souscription.

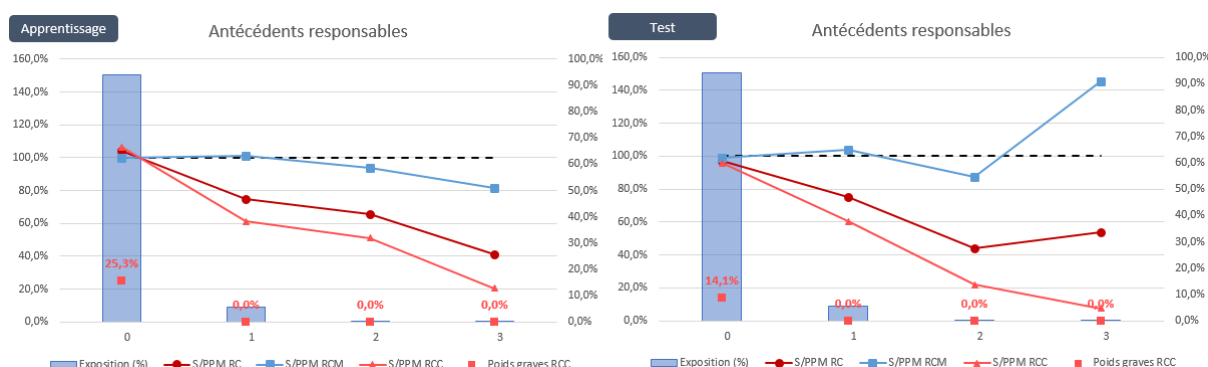


Figure 4.17 – Ajustement du modèle de prime pure selon les antécédents responsables

4.4 Analyse comparative des modèles

Afin de mettre en lumière ce qui différencie le nouveau modèle de prime pure de l'existant, une comparaison des modèles est réalisée. Pour rappel, le modèle existant a été élaboré exclusivement à partir de l'expérience du portefeuille des agents généraux (portefeuille *Vectura*) et adapté au canal du direct par des ajustements ciblés.

Les modèles sont comparés selon différents critères en commençant par une étude comparative de leur structure tarifaire, une attention particulière est apportée à la zone géographique compte-tenu de son influence dans la tarification. Puis les distributions des primes pures modélisées sont comparées en distinguant le canal de distribution.

Dans la suite de ce mémoire, le terme « **modèle Omnicanal** » fera référence au nouveau modèle de prime pure constitué à partir de l'expérience des deux portefeuilles tandis que le terme « **modèle CU** » renverra au modèle existant.

4.4.1 Comparaison des structures tarifaires

Le tableau 4.2 ci-dessous met en évidence une structure tarifaire très proche entre les deux modèles avec une base commune de 20 variables, soit une correspondance structurelle de 91%. Cette forte similarité s'explique logiquement par le poids important du portefeuille des agents généraux dans la base d'étude (75% des risques étudiés sont issus du portefeuille *Vectura*).

Critère tarifaire	Modèle CU	Modèle omnicanal
Mode d'achat		✓
Age conducteur	✓	✓
Age obtention permis	✓	
Ancienneté d'assurance	✓	✓
Ancienneté carte grise	✓	✓
Ancienneté véhicule	✓	✓
Carburant	✓	✓
Carrosserie	✓	✓
Emission CO2	✓	
Conduite exclusive - Commercial	✓	✓
Profession	✓	✓
CRM	✓	✓
Mode de garage	✓	✓
Formule souscrite	✓	✓
Groupe marque véhicule	✓	
Ancienneté CRM 50	✓	✓
Enfants novices non assurés	✓	✓
Antécédents responsables	✓	✓
Antécédents non responsables	✓	✓
Nb voitures au foyer		✓
Situation familiale	✓	
Groupe SRA	✓	✓
Classe de prix SRA	✓	✓
Forfait kilométrique	✓	✓
Usage du véhicule	✓	✓
Zonier	✓	✓
Nb variables	24	22

Table 4.2 – Structure tarifaire des modèles

Le modèle Omnicanal se distingue néanmoins par une structure tarifaire légèrement moins complexe, il comptabilise 22 critères tarifaires contre 24 pour le modèle CU. Dans le détail, les différences avec le modèle CU sont les suivantes :

- **Suppression de 4 variables** : âge obtention du permis, groupe marque du véhicule, situation familiale et émission de CO₂ du véhicule.
- **Ajout de 2 nouvelles variables** : mode d'achat du véhicule et nombre de voitures au foyer.

Si les variables ajoutées ne semblent représenter une amélioration majeure par rapport au modèle CU au regard de leur modeste contribution (contribution de 0,3% pour le mode d'achat et de 4,6% pour le nombre de véhicules au foyer), les variables supprimées pourraient quant à elles

participer à un meilleur ajustement du modèle Omnicanal sur la canal direct. En effet, ces quelques variables ne semblent pas totalement adaptées à la sinistralité du direct avec des tendances parfois contraires avec le canal intermédié, ce qui pourrait expliquer partiellement l'inadéquation du modèle CU au canal direct.

Par exemple, le graphique 4.18 comparant les primes pures empiriques selon la situation familiale montre clairement un effet quasi-neutre de cette variable sur le canal direct contrairement au canal intermédié où elle semble avoir un certain impact sur la sinistralité.

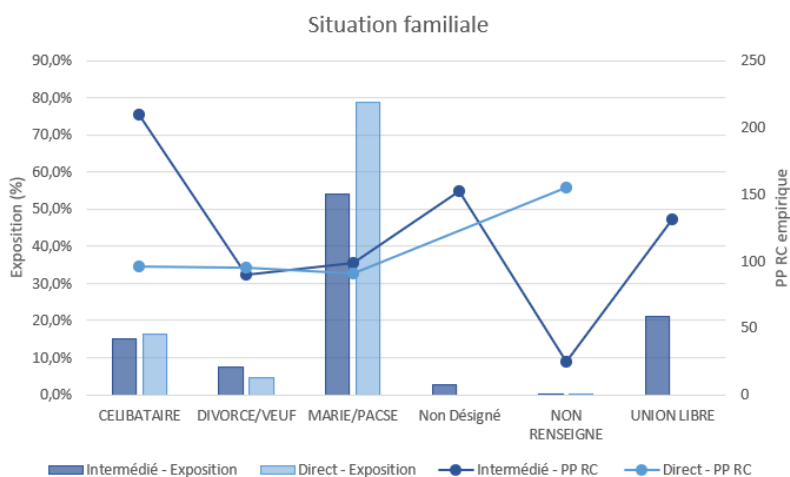


Figure 4.18 – Comparaison des primes pures empiriques selon la situation familiale

4.4.2 Comparaison par critères tarifaires

Une analyse détaillée des modèles est ensuite réalisée, elle se base sur une comparaison par critères tarifaires et consiste à confronter les modèles suivant l'amplitude, le spread et les coefficients de chacune de leurs variables communes.

Le graphique 4.19 ci-dessous comparant les amplitudes et les spreads des variables tarifaires des modèles montre des différences notables dans leur manière de segmenter le risque. En effet, il apparaît que le modèle Omnicanal se distingue par une segmentation plus forte suivant l'ancienneté du véhicule (spread de 123% contre seulement 13% pour le modèle CU) alors que le modèle CU se distingue davantage sur l'âge du conducteur (spread de 104% contre 57% pour le modèle Omnicanal). Des différences significatives apparaissent également sur d'autres variable tarifaires tels que le forfait kilométrique, l'ancienneté d'assurance ou encore le groupe SRA du véhicule. Les deux modèles se rejoignent toutefois sur quelques critères comme le zonier ou l'ancienneté du CRM 50. La variable de la conduite exclusive est, quant à elle, à mettre à part car son impact est purement commercial et de même intensité sur les deux modèles (réduction tarifaire de 10% en cas de conduite exclusive).

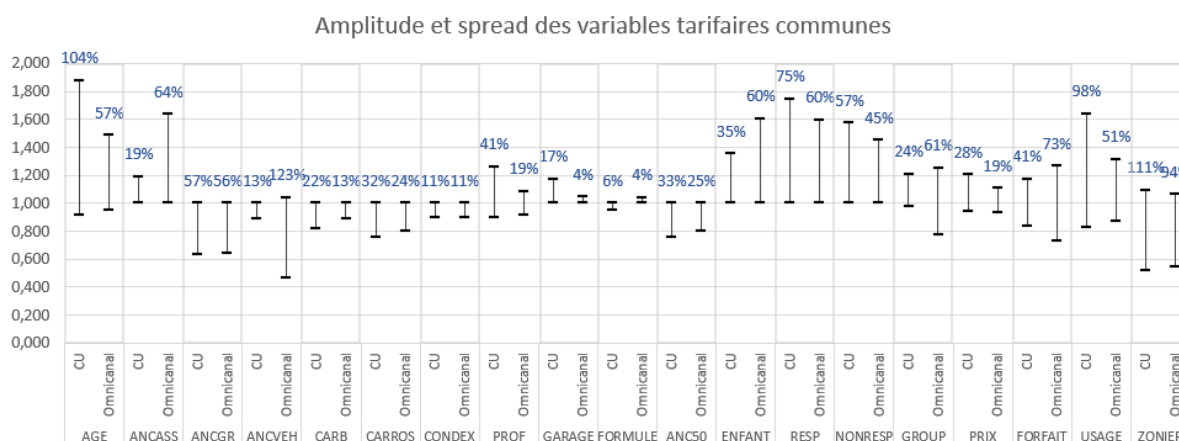


Figure 4.19 – Comparaison des amplitudes et spreads des variables tarifaires des modèles

Une mesure plus précise des impacts tarifaires entre les deux modèles est obtenue en comparant directement les coefficients sur chacun des critères. Un exemple de cette étude est donné ci-dessous et concerne l'âge du conducteur. Le graphique 4.20 montre une tendance similaire des courbes des coefficients tarifaires (graphique de gauche) mais avec des différences notables sur certaines tranches d'âge (graphique de droite). Comparativement au modèle CU, le modèle Omnicanal affiche une baisse entre 5%-10% sur le segment des jeunes conducteurs de moins de 25 ans (5% des expositions) et une hausse du même ordre pour les seniors de 65-80 ans (17% des expositions).

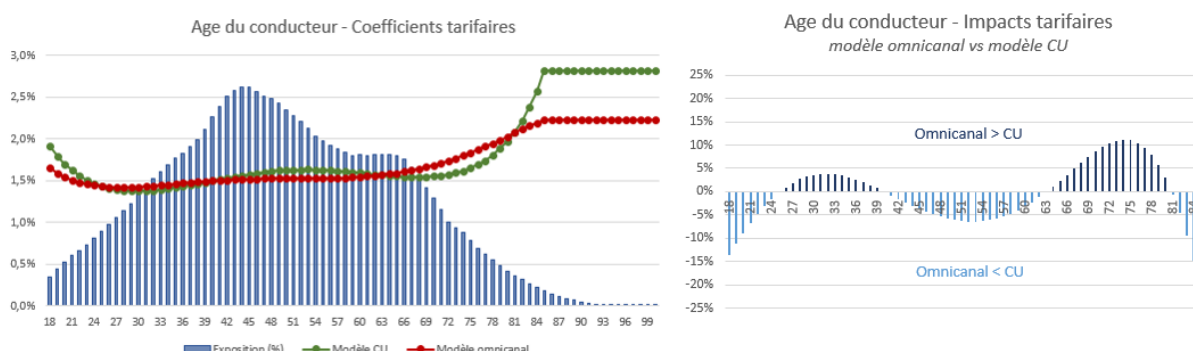


Figure 4.20 – Comparaison des variables tarifaires des modèles

4.4.3 Comparaison des zoniers

Avec une contribution de 13,8%, le zonier est l'une des variables les plus discriminantes dans le modèle de prime pure omnicanal. Et il en va de même pour le modèle CU, d'où l'attention particulière accordée à l'analyse cette variable.

Concernant la méthodologie utilisée dans la création des zoniers, elle est identique pour les deux modèles et s'appuie sur la méthode du lissage spatial des résidus de fréquence RC matérielle avec un découpage en 12 zones de même exposition. Mais contrairement au zonier omnicanal, le zonier CU n'a pas été construit à partir d'une base commune mais uniquement à partir du portefeuille des agents généraux. Il a ensuite été adapté au canal direct par un dézonage de certaines communes afin de prendre en considération l'expérience du direct.

Ci-dessous une comparaison du zonier Client Unique après dézonage (graphique 4.21) et du zonier omnicanal (graphique 4.22). A première vue, ces zoniers semblent être assez équivalents puisque les zones les plus touchées concernent globalement les mêmes régions et notamment l’Île-de-France, le Nord-Pas-de-Calais, le Rhône-Alpes ou encore le Languedoc-Roussillon. Mais à un niveau plus fin, des différences significatives apparaissent et certaines communes sont reclassées par le zonier omnicanal suite à une meilleure connaissance du risque apportée par l’expérience du direct.

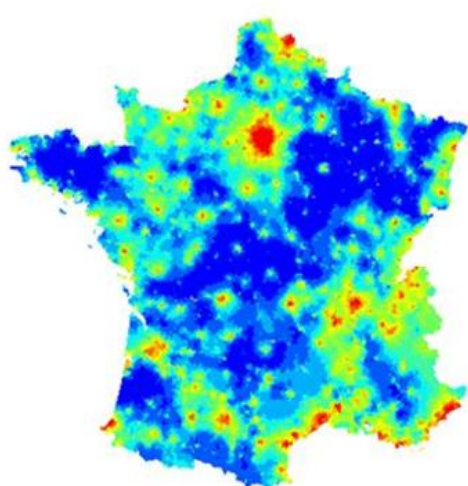


Figure 4.21 – Zonier Client Unique

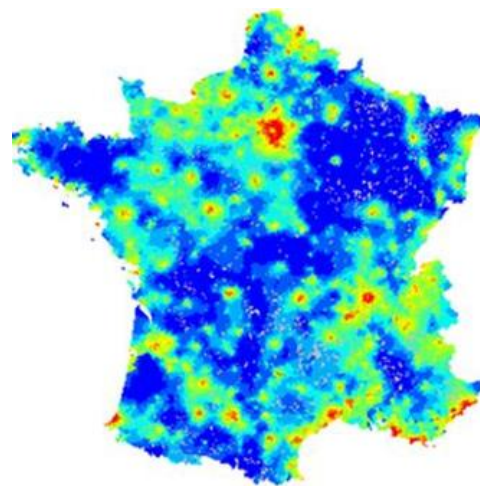


Figure 4.22 – Zonier Omnicanal

Par ailleurs, les deux portefeuilles se caractérisent par une répartition inégale selon la zone géographique, le portefeuille du direct étant plus largement représenté dans les zones élevées reflétant la différence d’exposition géographique entre les deux portefeuilles. Cependant, cette forte concentration du direct dans les zones élevées est plus limitée avec le zonier omnicanal (-2,1 points pour la zone 11 / -2,8 points pour la zone 12), ce qui traduit un certain remaniement des communes.

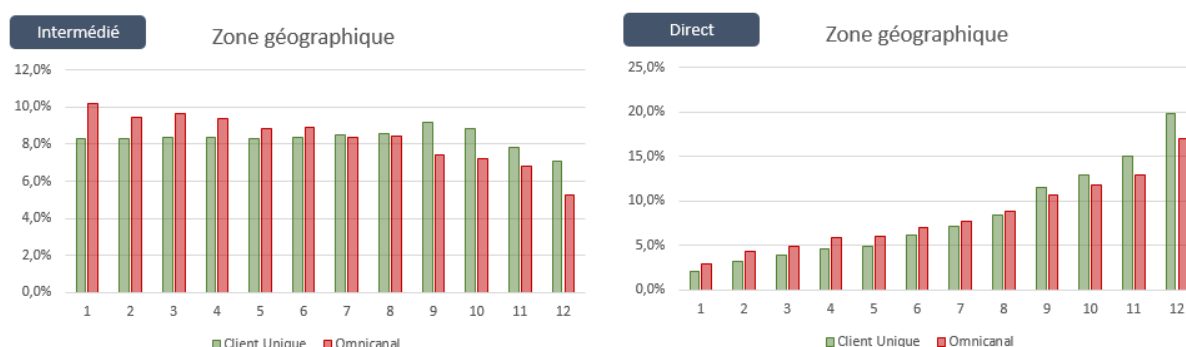


Figure 4.23 – Répartition des portefeuilles par zone géographique

Le graphique 4.24 représentant les écarts de zones entre le zonier omnicanal et CU permet de mieux se rendre compte des divergences de classement de communes entre les deux zoniers. Il montre qu’en effet, le zonier omnicanal déclassé un nombre important de communes par rapport au zonier CU puisque 43% des communes (~15 000 communes) connaissent une baisse d’au moins une zone et que seulement 16% des communes (~5 600 communes) connaissent une hausse.

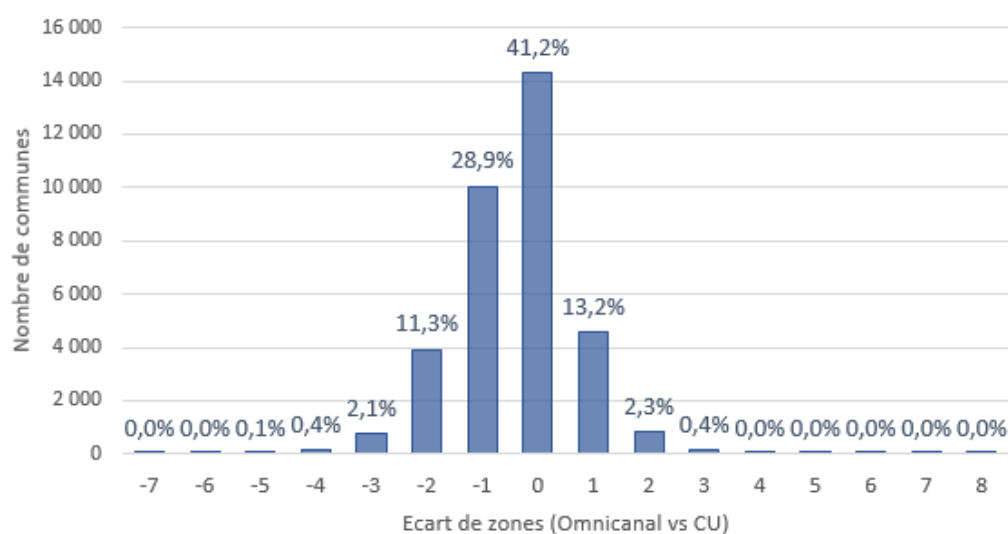


Figure 4.24 – Ecart de zones géographiques entre les zoniers omnicanal et CU

Le détail des zones CU impactées est donné ci-dessous sous la forme d'une matrice de passage du zonier CU vers le zonier omnicanal (graphique 4.25) qui permet de mettre en correspondance une zone CU donnée avec les zones omnicanals associées. Il apparaît que l'ensemble des zones CU sont concernées par cette réorganisation à des niveaux plus ou moins importants, la zone CU n°1 étant celle qui subit le moins de changements (correspondance exacte de 84% avec le zonier omnicanal) tandis que la zone CU n°10 est la plus touchée avec une baisse d'au moins une zone pour 72% de ses communes.

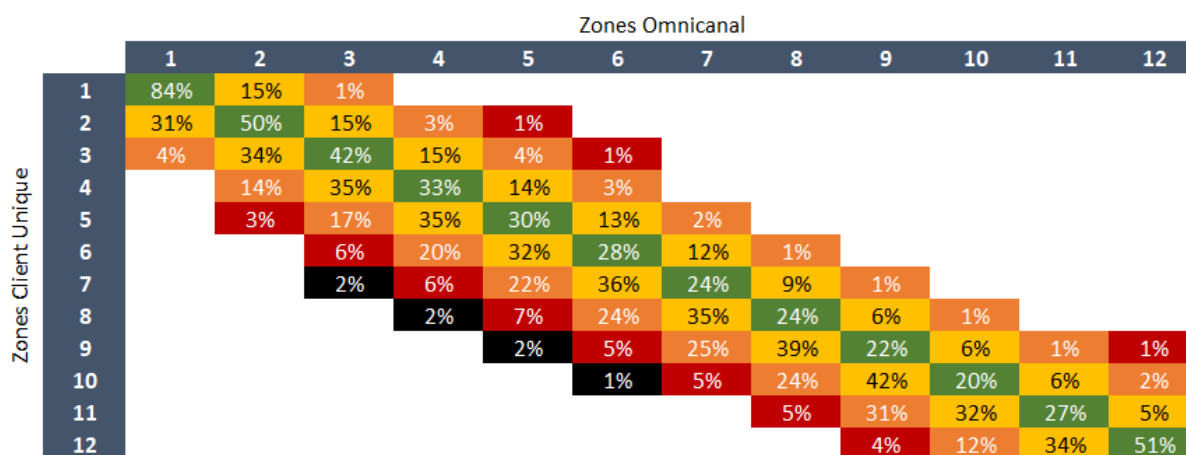


Figure 4.25 – Matrice de passage du zonier CU vers le zonier omnicanal (en % de communes)

4.4.4 Comparaison des distributions de prime pure

En préambule de l'étude de la performance des modèles selon le canal de distribution, une comparaison des distributions des primes pures est menée. Le graphique ci-dessous permet de mettre en évidence un comportement très similaire des modèles sur le canal intermédié laissant penser qu'un

nouvel équilibre technique est trouvé par le modèle Omnicanal. Ce constat ne se retrouve cependant pas sur le canal direct où la distribution des primes pures modélisées diffère significativement selon le modèle, le modèle Omnicanal se caractérisant par une distribution moins étendue (intervalle interquartile plus resserré) ainsi que des indicateurs synthétiques (moyenne et quartiles) inférieurs au modèle CU.

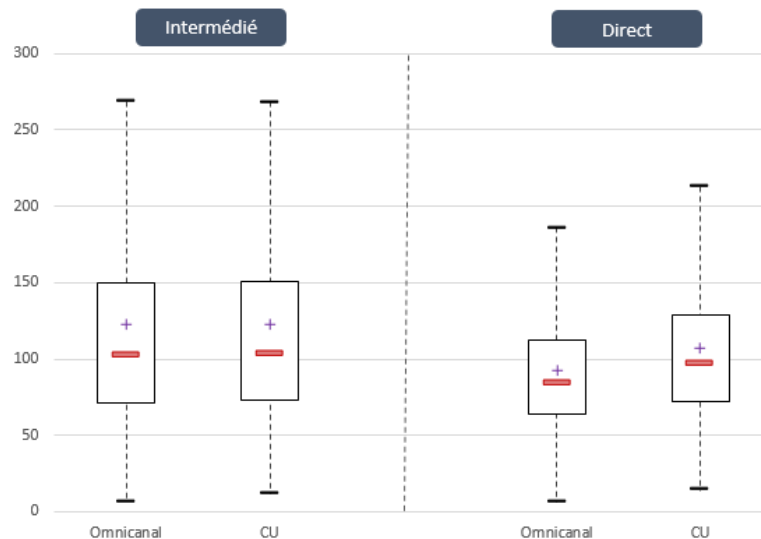


Figure 4.26 – Distribution des primes pures modélisées selon le canal de distribution

L'étude de la dispersion des écarts de prime pure modélisée selon le canal de distribution (graphique 4.27) comparant le modèle Omnicanal au modèle CU confirme ces résultats :

- **Canal intermédié** : La courbe est symétrique et centrée aux alentours de 0% (équidistribution des profils de risques avec une baisse ou une hausse tarifaire) ce qui indique une équivalence en moyenne des modèles sur ce canal.
- **Canal direct** : La courbe est également symétrique et centrée négativement autour de -13%, ce qui a pour conséquence une baisse tarifaire pour près de 77% des profils de risque par rapport au modèle CU. Ce dernier résultat apporte une première réponse quant à la capacité du modèle Omnicanal à s'ajuster sur le canal direct étant donné qu'une surestimation a été constatée avec le modèle CU.

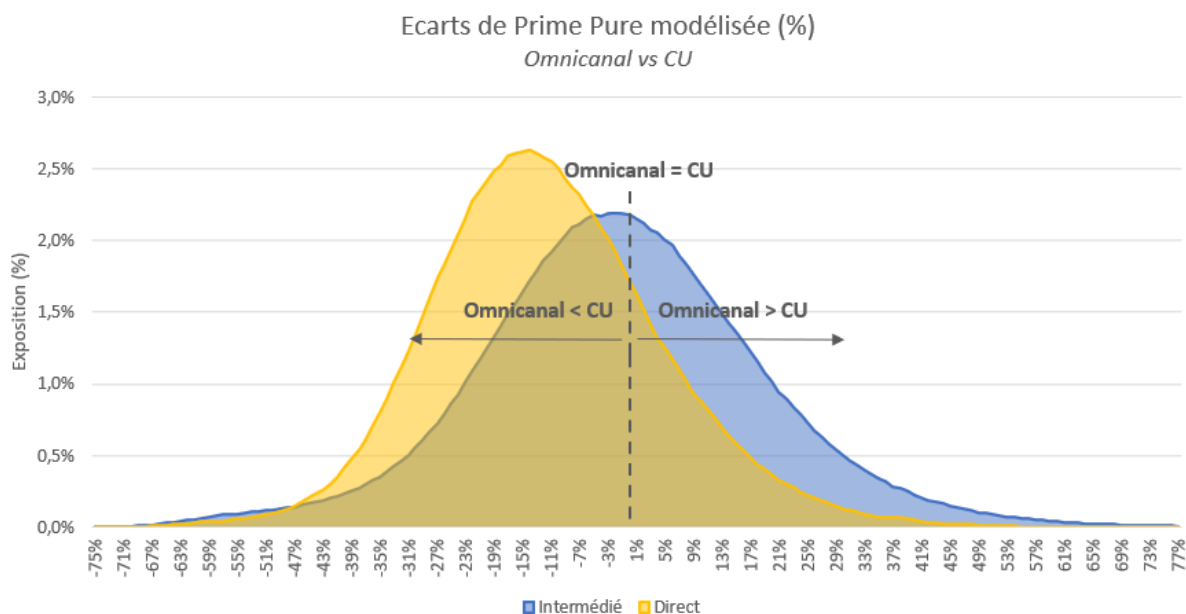


Figure 4.27 – Dispersion des écarts de prime pure modélisée (%) selon le canal de distribution (Omnicanal vs CU)

4.5 Performance des modèles selon le canal de distribution

L'objectif de cette dernière section est de comparer la performance des modèles et d'en mesurer la pertinence selon le canal de distribution. L'évaluation des modèles se base sur plusieurs indicateurs synthétiques (S/PPM, RMSE, MAE et indice de Gini) étudiés en parallèle. Une évaluation par critère tarifaire complète cette analyse.

En plus des deux modèles mentionnés (CU et Omnicanal), un troisième modèle est introduit. Il s'agit d'une adaptation du modèle CU en remplaçant son zonier par le zonier omnicanal (modèle nommé CU avec zonier omnicanal). Etant donné l'importance du zonier dans la tarification et de la similitude des coefficients tarifaires de cette variable entre les deux modèles, ce nouveau modèle apparaît assez pertinent à étudier.

4.5.1 Qualité d'ajustement global des modèles

Le ratio S/PPM est de nouveau utilisé ici pour mesurer la qualité de l'ajustement global des différents modèles selon le canal de distribution.

L'analyse du graphique 4.28 montre que les 3 modèles affichent un niveau de performance sensiblement différent selon le canal de distribution excepté pour le modèle Omnicanal qui apparaît le plus adapté sur l'ensemble des canaux, ce qui se traduit par un S/PPM très proche de 1 sur chacun d'eux. Le modèle CU est quant à lui parfaitement ajusté sur le canal intermédiaire mais beaucoup moins sur le canal direct puisqu'il surestime la prime pure observée de plus de 16% (S/PPM de 86,2%). Le modèle CU avec zonier omnicanal permet d'améliorer légèrement l'ajustement sur le canal direct sans pour autant être pleinement satisfaisant (S/PPM de 88,6% vs 86,2% avec CU) et au prix d'une dégradation de l'ajustement sur le canal intermédiaire (S/PPM de 103,1% vs 100% avec CU).

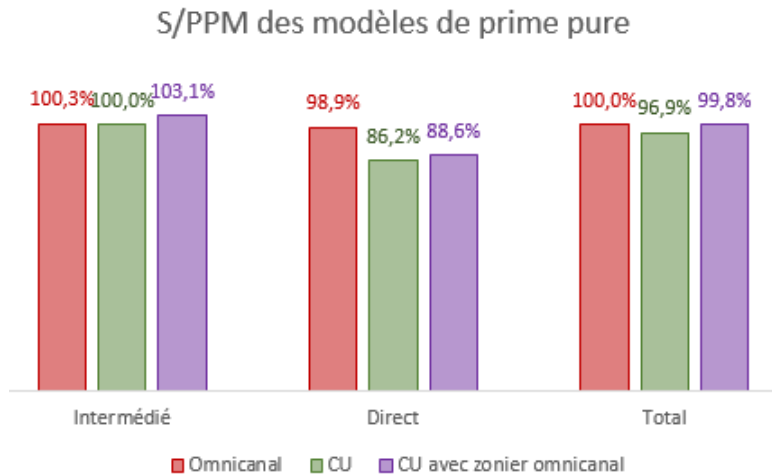


Figure 4.28 – Qualité d’ajustement des modèles selon le canal de distribution

4.5.2 Autres indicateurs de performance : MAE, RMSE et indice de Gini

Les critères MAE et RMSE sont des indicateurs couramment utilisés lorsqu’il s’agit de mesurer les écarts de prédiction générés par un modèle. Le premier critère fournit une indication sur l’erreur moyenne de prédiction et le second fournit une indication par rapport à la dispersion de la qualité de la prédiction. Une définition plus précise de ces deux critères est donnée ci-après.

Quant à l’indice de Gini, il est rappelé qu’il mesure la capacité de segmentation du modèle.

MAE (Mean Absolute Error)

Il correspond à la moyenne arithmétique des valeurs absolues des écarts entre les prédictions du modèle et les observations :

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i|$$

RMSE (Root Mean Square Error)

Il correspond à la racine carrée de l’erreur quadratique moyenne (moyenne arithmétique des carrés des écarts de prédiction) :

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

Aussi, le meilleur modèle sera celui qui minimise ces deux critères avec, en parallèle, un indice de Gini le plus proche de 1.

Le tableau 4.3 ci-dessous se concentre sur les deux modèles initiaux (CU et Omnicanal) et fait apparaître une performance similaire au sens du critère RMSE où des différences ne sont visibles que sur les dernières décimales en faveur du modèle Omnicanal. Avec le critère MAE, les différences sont plus marquées, principalement sur le canal direct où le modèle Omnicanal l’emporte assez nettement avec un gain de performance de 6,2% par rapport au modèle CU. Un meilleur indice de Gini est également obtenu avec le modèle Omnicanal quel que soit le canal de distribution (gain de 2,2pts sur le canal intermédié / 1,6pts sur le canal direct).

Modèle	Canal	MAE	RMSE	Gini
CU	Intermédié	110,21	6 461,36	38,6%
	Direct	79,71	2 852,14	25,4%
	Total	101,74	5 693,50	35,3%
Omnicanal	Intermédié	110,00	6 461,35	40,8%
	Direct	74,75	2 852,15	27,0%
	Total	100,21	5 693,49	36,5%

Table 4.3 – Indicateurs de performance des modèles

4.5.3 Analyse des résultats par critère et selon le canal de distribution

Pour finir, une étude plus fine de la performance des modèles est réalisée. Elle consiste à mesurer la qualité d’ajustement des modèles sur différents critères tarifaires et selon le canal de distribution.

L’âge du conducteur

Canal intermédié : De manière générale, les différents modèles présentent un niveau d’ajustement comparable sur la tranche d’âge des 26-74 ans (85% des expositions). Des différences apparaissent aux deux extrémités de la population, le modèle Omnicanal étant le plus performant sur les jeunes de moins de 25 ans (6% des expositions) tandis que les autres modèles sont un peu mieux ajustés sur les seniors de 75 ans et plus (6% des expositions) ainsi que sur les conducteurs non désignés (3% des expositions).

Canal direct : Sans équivoque, le modèle Omnicanal est le plus adapté sur l’ensemble des tranches d’âge excepté peut-être sur les 26-34 ans (14% des expositions) mais l’ajustement de ce segment est probablement dégradé du fait d’une charge de graves très importante.

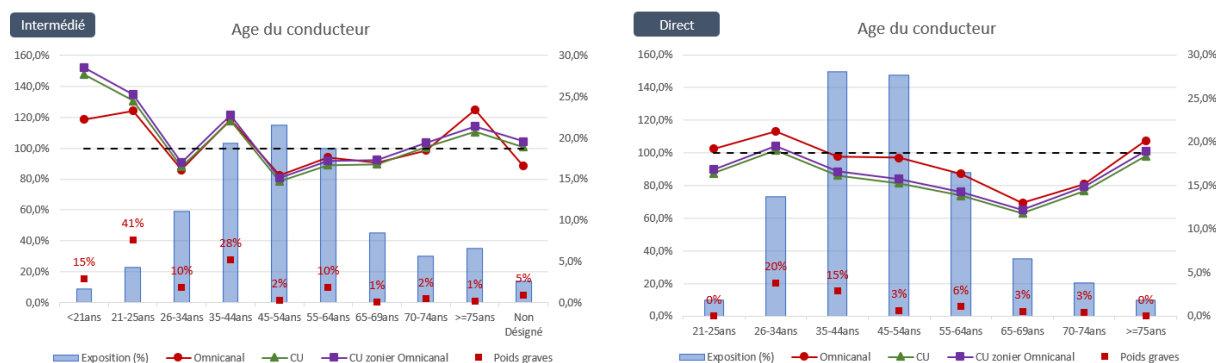


Figure 4.29 – Qualité d’ajustement des modèles selon l’âge du conducteur

L'ancienneté du véhicule

Canal intermédiaire : Le modèle Omnicanal fait clairement la différence sur les véhicules de plus de 20 ans (8% des expositions) et particulièrement sur les plus de 25 ans. Ce résultat semble indiquer un bon choix dans le retraitement de cette variable et notamment dans le regroupement des modalités (seuil de 30 ans pour le modèle Omnicanal vs 20 ans pour le modèle CU).

Canal direct : Les courbes d'ajustement des modèles sont parallèles entre elles, ce qui indique que les primes pures modélisées suivent exactement la même tendance. Mais le modèle Omnicanal s'en sort mieux sur la majorité des segments excepté sur les véhicules de 1-2 ans et 18-20 ans.



Figure 4.30 – Qualité d'ajustement des modèles selon l'ancienneté du véhicule

Le type de formule

Canal intermédiaire : Sur ce critère aucun des modèles ne se distingue véritablement l'un de l'autre.

Canal direct : Le meilleur ajustement est obtenu avec le modèle Omnicanal qui s'ajuste davantage sur les formules Tiers maxi et Tous risques (78% des expositions) et moins sur la formule Tiers mini qui est sous-tarifée à raison là encore d'un poids de graves bien plus important que sur les autres formules.



Figure 4.31 – Qualité d'ajustement des modèles selon la formule souscrite

La situation familiale

Canal intermédiaire : Cette variable n'est pas un critère tarifaire du modèle Omnicanal et pourtant il s'ajuste aussi bien voire mieux sur certains segments par rapport au modèle CU qui l'intègre dans sa tarification. Le modèle Omnicanal est plus performant sur les personnes seules (célibataires et divorcés/veufs) représentant 23% du portefeuille des intermédiaires. Comme vu précédemment, les non désignés (3% des expositions) sont surtarifés de près de 13% avec le modèle Omnicanal.

Canal direct : Le modèle Omnicanal est parfaitement ajusté sur le segment des mariés/pacsés représentant 79% du portefeuille du direct. Les célibataires sont également mieux approchés avec ce modèle. Concernant les divorcés/veufs (5% des expositions), ils sont surtarifés de 15% par le modèle CU et sous-tarifés de 8% par le modèle Omnicanal.

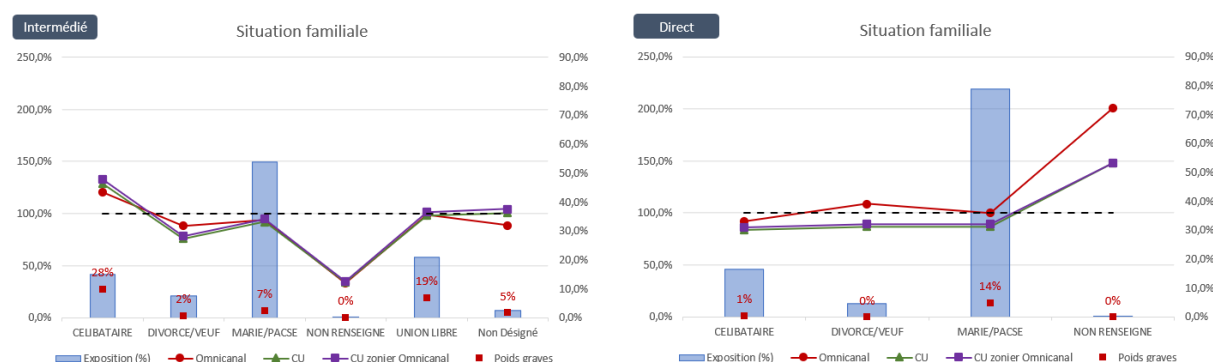


Figure 4.32 – Qualité d'ajustement des modèles selon la situation familiale

Les forfaits kilométriques

Canal intermédiaire : Le modèle Omnicanal surperforme les autres modèles sur le segment des petits rouleurs (5 000 km / 9 000 km) représentant 29% du portefeuille des intermédiaires et fait jeu égal sur les autres forfaits voire légèrement mieux sur le forfait illimité (écart par rapport à l'observé de -2,9% pour le modèle Omnicanal vs -6,5% pour le modèle CU).

Canal direct : A l'exception du forfait 7 000 km (8% du portefeuille) et du forfait illimité dans une moindre mesure, le modèle Omnicanal est celui qui s'ajuste le mieux au portefeuille du direct malgré une surtarification significative des forfaits 9 000 km et 19 000 km.



Figure 4.33 – Qualité d'ajustement des modèles selon le forfait kilométrique

L'usage du véhicule

Canal intermédiaire : Tous les modèles affichent une bonne qualité d'ajustement sur les usages « Vie privée » et « Vie privée / Trajet » représentant 87% du portefeuille des intermédiaires. L'usage « Déplacements professionnel » (13% des expositions) est assez nettement sous-tarifé par l'ensemble des modèles, le meilleur ajustement est cependant obtenu avec le modèle Omnicanal tandis que le troisième modèle est celui qui s'en éloigne le plus. Du fait d'une volumétrie trop insuffisante, les résultats sur l'usage « Tournées » ne sont pas significatifs.

Canal direct : Globalement, le modèle Omnicanal est celui qui présente le meilleur ajustement, en particulier sur l'usage « Vie privée » (42% des expositions). Des résultats équivalents entre les modèles sont constatés concernant l'usage « Vie privée / Trajet » avec une sous-tarification de 6,4% pour le modèle Omnicanal et une surtarification de 7,7% pour le modèle CU.

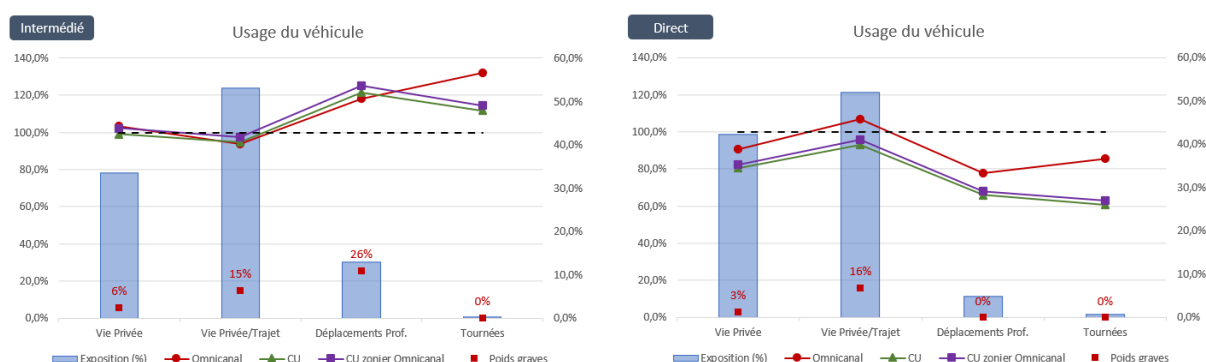


Figure 4.34 – Qualité d'ajustement des modèles selon l'usage du véhicule

Les enfants novices

La déclaration d'un enfant novice en tant que conducteur secondaire représente généralement un facteur aggravant du risque pour l'assureur. En effet, il induit un manque d'expérience de conduite et un comportement au volant souvent moins responsable du jeune conducteur (vitesse excessive, alcool, ...).

Canal intermédiaire : Le modèle Omnicanal se distingue des autres modèles par une meilleure qualité d'ajustement sur le segment à risque des enfants novices (4% des expositions) avec un S/PPM de 103% contre 111% pour le modèle CU et 115% pour le modèle CU alternatif.

Canal direct : Résultats non significatifs sur le segment des enfants novices compte-tenu d'une volumétrie insuffisante.

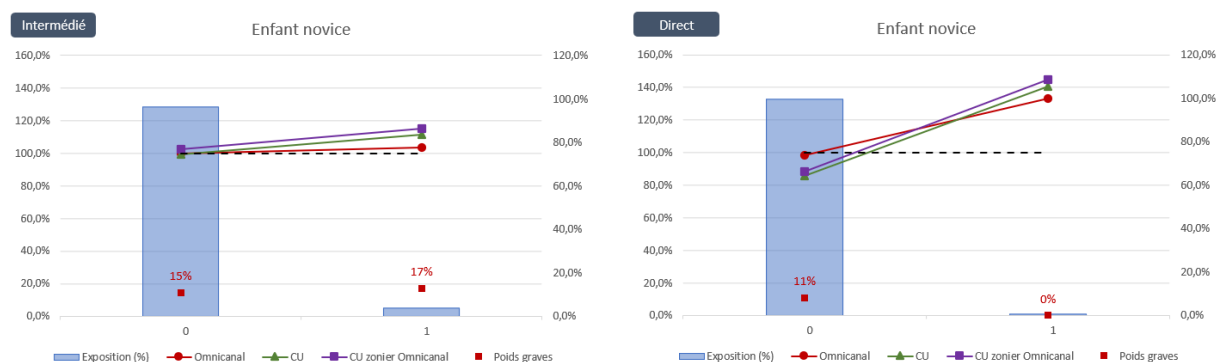


Figure 4.35 – Qualité d’ajustement des modèles selon la désignation d’un enfant novice

Le statut de la commune

La variable du statut de la commune (définition *Insee*) permet de catégoriser les communes en 4 grandes zones (Banlieue, Rurale, Ville-centre et Ville isolée) et fait référence à la composante géographique du risque intégrée dans les différents modèles via la variable du zonier.

Canal intermédiaire : Des performances similaires sont observées entre les différents modèles qui apparaissent parfaitement ajustés sur la modalité « Banlieue ». Des résultats plus contrastés sont constatés sur les autres modalités avec une sous-estimation de l’ordre de 15% sur la modalité « Rurale » représentant 37% des expositions.

Canal direct : Alors que le modèle Omnicanal est plus performant sur les modalités « Banlieue » (50% des expositions) et « Rural » (19% des expositions), les modèles CU et CU alternatif le sont davantage sur les modalités « Ville-centre » (24% des expositions) et « Ville isolée » (7% des expositions).



Figure 4.36 – Qualité d’ajustement des modèles selon le statut de la commune

Conclusion

La pression de la concurrence et les nouvelles attentes des clients qui ont parfaitement intégré le digital dans leurs modes de consommation obligent les assureurs à revoir leurs stratégies. S'inscrivant dans le cadre du projet « Client Unique » d'Aviva France et de la création d'une nouvelle offre automobile à destination du particulier, ce mémoire a été l'occasion d'aborder la problématique de la tarification de la garantie RC en assurance automobile et avait à cœur de proposer un modèle de prime pure ajusté sur chacun des canaux de distribution.

Une première analyse comparative des portefeuilles existants a permis de mettre en évidence des profils différents entre la clientèle du direct et celle des agents généraux. Alors que la clientèle du direct est plus sensible au prix et se concentre dans les grandes villes, celle des agents généraux s'attache davantage au niveau de couverture des garanties et réside majoritairement en zone rurale. Cette étude a démontré la nécessité de disposer d'une base de données globale afin de garantir l'exhaustivité des profils de risque et un meilleur apprentissage lors de l'élaboration du modèle de prime pure.

Aussi, le premier enjeu de ce mémoire a été d'unifier ces deux portefeuilles au sein d'une unique base de données étendue et fiabilisée sur laquelle devait s'appuyer la construction du nouveau modèle de prime pure. Mais l'utilité de ce datamart omnicanal dépasse largement le cadre de cette étude, il constitue un vrai plus pour Aviva France dans sa connaissance globale du risque de la branche automobile du particulier. L'exploration de cette base de données a conduit à l'étude séparée des composantes matérielles et corporelles du risque RC et à la détermination, via la théorie des valeurs extrêmes, d'un seuil d'écèlement adéquat à chacune de ces deux composantes.

Le second enjeu de ce mémoire a été de concevoir un nouveau zonier adapté à l'ensemble des canaux de distribution par une couverture complète du territoire grâce à la complémentarité des portefeuilles du digital et des agents généraux. Ce zonier omnicanal s'est avéré particulièrement pertinent et d'une importance majeure dans la tarification. En effet, le zonier fait partie des variables les plus influentes dans la tarification et a participé à améliorer l'ajustement du modèle sur l'ensemble des canaux.

Enfin, concernant l'objectif premier de ce mémoire qui était de proposer une nouvelle modélisation de la garantie RC adaptée à tous les modes de distribution : Celui-ci a été atteint et a permis de réconcilier les deux produits existants autour d'une même tarification. De plus, la comparaison avec le modèle existant a été riche d'enseignements et a permis de mieux comprendre pourquoi celui-ci avait échoué dans cet exercice. L'explication tient en 3 points : Le premier concerne la structure tarifaire du modèle CU qui, bien que très similaire au modèle Omnicanal, intègre certaines variables non pertinentes. Le second concerne la segmentation, la manière dont les variables tarifaires segmentent le risque et sur ce point les deux modèles divergent nettement sur certains critères. Enfin le dernier point concerne le zonier et bien que l'existant ait bénéficié d'un dézonage afin de l'adapter à l'exposition géographique très urbaine du portefeuille du direct, cela n'a pas véritablement suffi à améliorer sa compatibilité.

Maintenant il est clair que la tarification proposée n'est pas parfaite, elle est perfectible dans sa manière d'appréhender le risque corporel puisqu'elle néglige la modélisation des sinistres graves qui

dégradent significativement la qualité globale du modèle de prime pure. Dès lors, la modélisation des graves et plus particulièrement de leur fréquence apparaît comme un point crucial et les algorithmes d'apprentissage statistiques tels que les *Random Forest* ou le *XG Boost* peuvent représenter une solution pertinente.

Bibliographie

Articles et support universitaire

FFA (2021) – *Le marché de l'assurance automobile des particuliers en 2020*

https://www.franceassureurs.fr/wp-content/uploads/VF_Lassurance_automobile_2020.pdf

CCFA/AAA DATA (2021) – *Le marché automobile français*

<https://ccfa.fr/wp-content/uploads/2021/01/dp01012021-1.pdf>

LEGIFRANCE – *Code des assurances*

https://www.legifrance.gouv.fr/codes/texte_lc/LEGITEXT000006073984

F. PLANCHET, A. MISERAY (2017) – *Tarification IARD, introduction aux techniques avancées* – Cours enseigné à l'Institut de Science Financière et d'Assurance (ISFA)

<http://www.ressources-actuarielles.net/EXT/ISFA/fp->

[isfa.nsf/34a14c286dfb0903c1256ffd00502d73/457a36a8ecc541aec1257d740067eec4/\\$FILE/GLM_FP.pdf](http://www.ressources-actuarielles.net/EXT/ISFA/fp-isfa.nsf/34a14c286dfb0903c1256ffd00502d73/457a36a8ecc541aec1257d740067eec4/$FILE/GLM_FP.pdf)

Mémoires

F. BOUTTIER (2015) – *Modélisation de la prime pure RC Automobile*

B. CARLES (2017) – *Techniques de zonage en tarification IARD*

M. MAITI (2012) – *Modélisation de la prime pure en assurance automobile*

J. LE BLOA (2020) – *Tarification d'un produit responsabilité civile automobile mobilisant les tarifs concurrents*

R. LAILY (2014) - *Construction d'un Modèle Interne Partiel en Assurance non-vie*

Ouvrages

S. TUFFERY (2017) – *Data mining et statistique décisionnelle, la science des données*

G. SPORTA (2011) – *Probabilités, analyse des données et statistique*

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

Annexes

Annexe A – Détermination du seuil d'écèlement de la composante matérielle

Fonction moyenne des excès

Au vu du graphique ci-dessous, le seuil d'écèlement de la composante matérielle se situerait aux alentours de 50 000€.

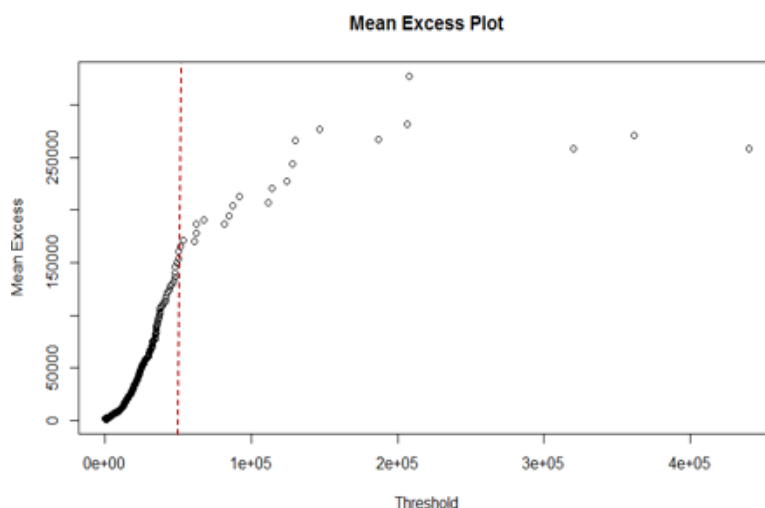


Figure 4.31 – *Mean Excess Plot* de la composante matérielle

Adéquation des excès par une GPD

L'ajustement des données à une loi de Pareto généralisée au seuil de 50 000€ est globalement bon même si le QQ-plot des excès fait apparaître quelques points qui s'écartent de la 1^{ère} bissectrice.

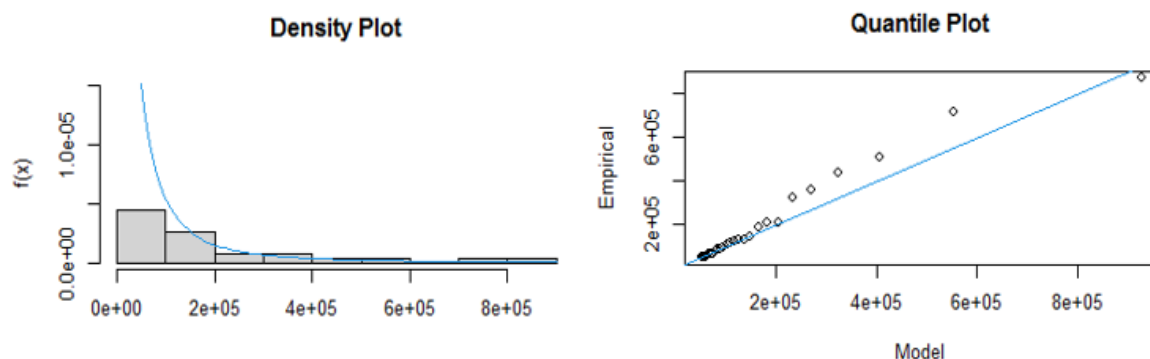


Figure 4.32 – Résultat de l'ajustement des excès par une GPD

Annexe B – Principaux résultats du modèle de fréquence RC matérielle non responsable

Détermination de la loi

Comme pour le modèle de fréquence RCM responsable, la loi de Poisson est parfaitement adaptée à la modélisation du modèle de fréquence RCM non responsable.

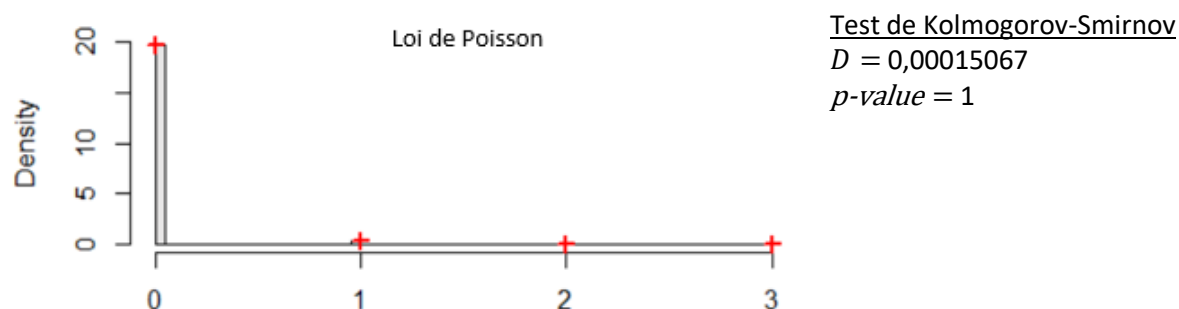


Figure 4.33 – Ajustement de la fréquence RCM non responsable à une loi de Poisson

Présentation du modèle

Alors que le modèle de fréquence RCM responsable s'explique en grande partie par la zone géographique, l'âge du conducteur et le forfait kilométrique (contribution cumulée de 56,1%), le modèle de fréquence RCM non responsable s'appuie davantage sur les caractéristiques du véhicule et notamment son ancienneté qui contribue à elle seule à hauteur de 34%. La zone géographique est également très importante dans le modèle puisqu'elle contribue à près de 30%.

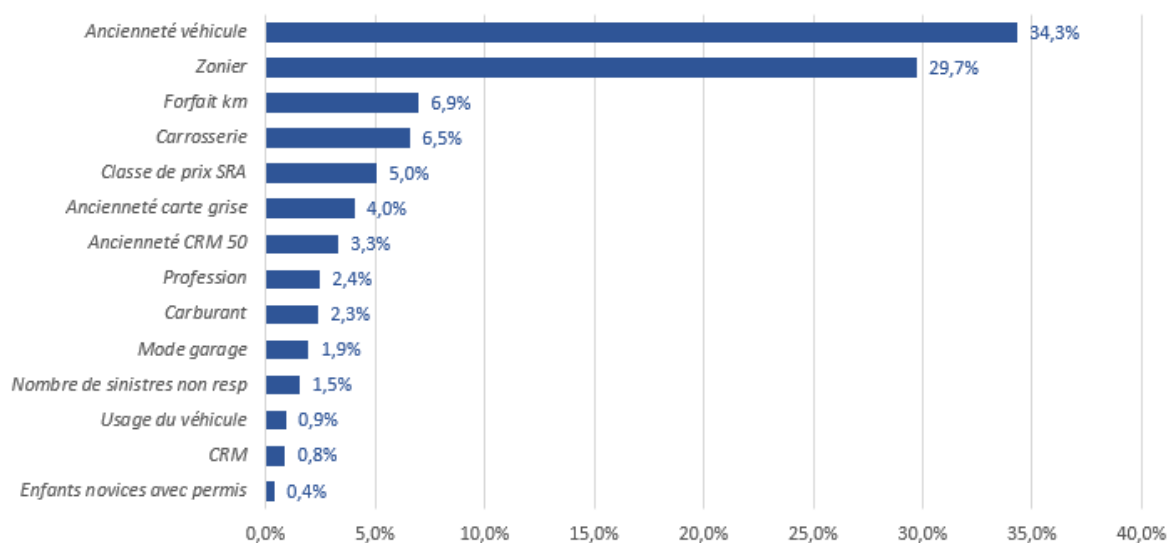


Figure 4.34 – Contribution des variables au modèle de fréquence RCM non responsable

Qualité d'ajustement global

Le tableau 4.4 montre clairement des écarts de prédiction très faibles ce qui traduit une bonne qualité d'ajustement en moyenne du modèle de fréquence RCM non responsable.

exercice	Apprentissage			Test		
	Fréquence estimée	Fréquence observée	écart (%)	Fréquence estimée	Fréquence observée	écart (%)
2013	3,43%	3,44%	-0,3%	3,47%	3,52%	-1,4%
2014	3,41%	3,38%	0,9%	3,44%	3,40%	1,2%
2015	3,36%	3,34%	0,6%	3,39%	3,35%	1,2%
2016	3,32%	3,36%	-1,2%	3,33%	3,36%	-0,9%
2017	3,30%	3,32%	-0,6%	3,30%	3,31%	-0,3%
2018	3,31%	3,28%	0,9%	3,29%	3,29%	0,0%
Total	3,35%	3,35%	0,0%	3,37%	3,37%	0,0%

Table 4.4 – Erreur moyenne d'ajustement du modèle de fréquence RCM non responsable

Courbe de lift

La courbe de lift fait apparaître une excellente qualité d'ajustement du modèle sur la base de test puisque la fréquence prédite est très proche de la fréquence observée sur l'ensemble de la distribution en vingtile.

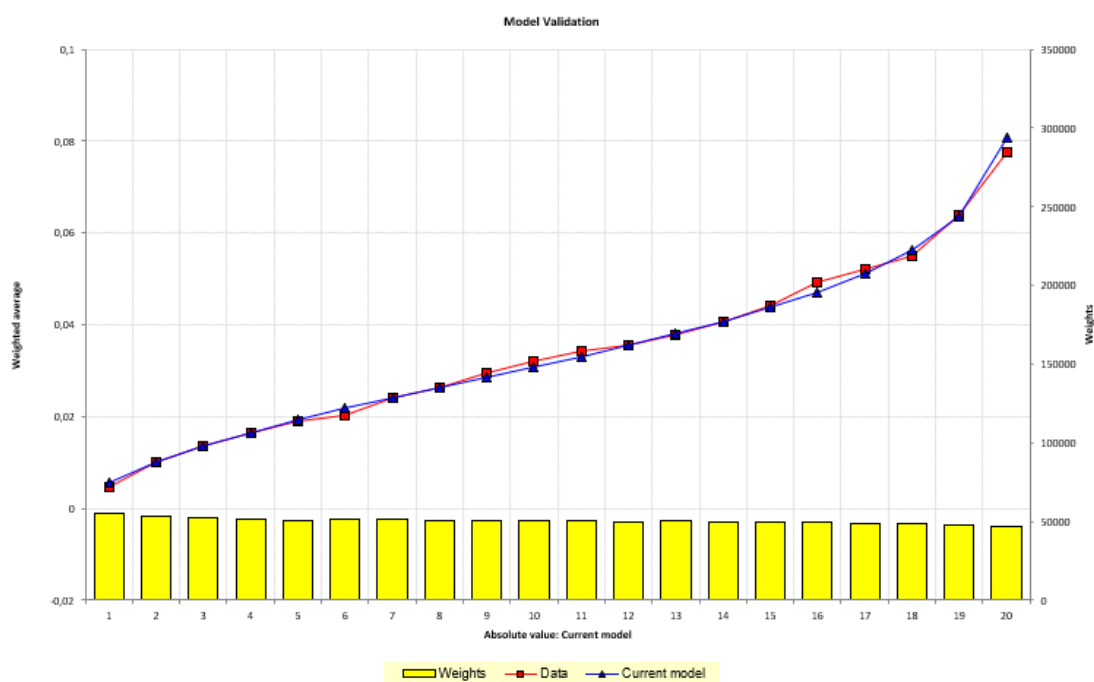


Figure 4.35 – Courbe de lift du modèle de fréquence RCM non responsable

Etude des résidus

Le nuage de points des résidus de la fréquence RCM non responsable est homogène, de forme elliptique et centré en 0. Les hypothèses du modèle sont donc valides.

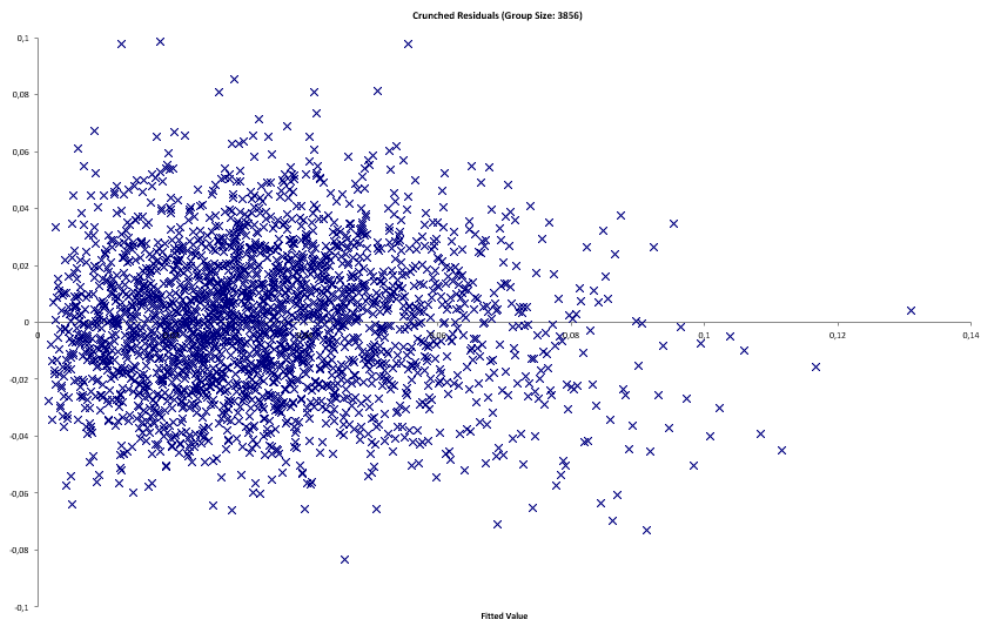


Figure 4.36 – Résidus standardisés de la déviance du modèle de fréquence RCM non responsable

Annexe C – Principaux résultats du modèle de coût moyen RC matériel non responsable

Détermination de la loi

La distribution du coût moyen RCM non responsable ne fait pas apparaître de pic de sinistres et s'ajuste assez bien par une loi Gamma.

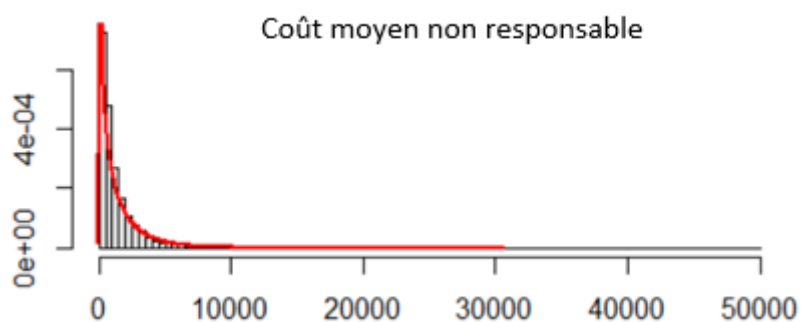


Figure 4.36 – Ajustement du modèle de coût moyen RCM non responsable à une loi Gamma

Présentation du modèle

Le modèle de coût moyen RCM non responsable se compose de 4 variables dont deux faisant référence à la valeur du véhicule et à sa récence (classe de prix SRA et ancienneté du véhicule), une autre variable liée à l'environnement (zone géographique) et la dernière à l'expérience d'assurance du conducteur (ancienneté d'assurance).

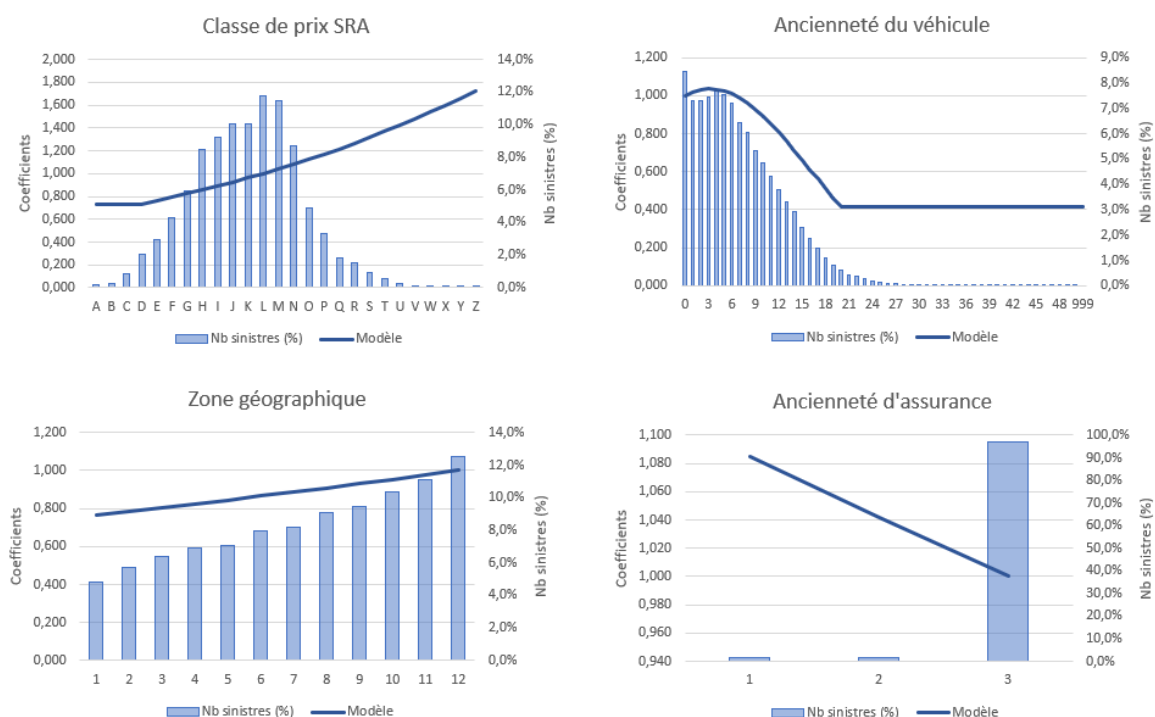


Figure 4.37 – Présentation du modèle de coût moyen RCM non responsable

De manière cohérente, l'évolution du coût moyen RCM non responsable est croissante suivant la classe de prix du véhicule et la zone géographique, décroissant suivant l'ancienneté du véhicule (courbe concave) et l'ancienneté d'assurance.

Qualité d'ajustement global

Malgré quelques écarts constatés par exercice de survenance, l'ajustement du modèle reste globalement satisfaisant.

exercice	Apprentissage			Test		
	Coût moyen estimé	Coût moyen observé	écart (%)	Coût moyen estimé	Coût moyen observé	écart (%)
2013	400 €	410 €	-2,5%	403 €	405 €	-0,7%
2014	415 €	407 €	2,0%	414 €	422 €	-1,8%
2015	430 €	427 €	0,6%	428 €	408 €	4,7%
2016	440 €	441 €	-0,3%	442 €	465 €	-5,0%
2017	456 €	434 €	4,9%	455 €	446 €	2,0%
2018	472 €	489 €	-3,6%	470 €	464 €	1,4%
Total	436 €	435 €	0,1%	436 €	435 €	0,1%

Table 4.6 – Erreur moyenne d'ajustement du modèle de coût moyen RCM non responsable

Courbe de lift

La courbe de lift établie sur la base de test ci-dessous fait apparaître des écarts ponctuels de prédiction sur certains vingtiles mais les deux courbes restent globalement très proches sur l'ensemble de la distribution, ce qui confirme la bonne qualité du modèle ainsi que sa robustesse.

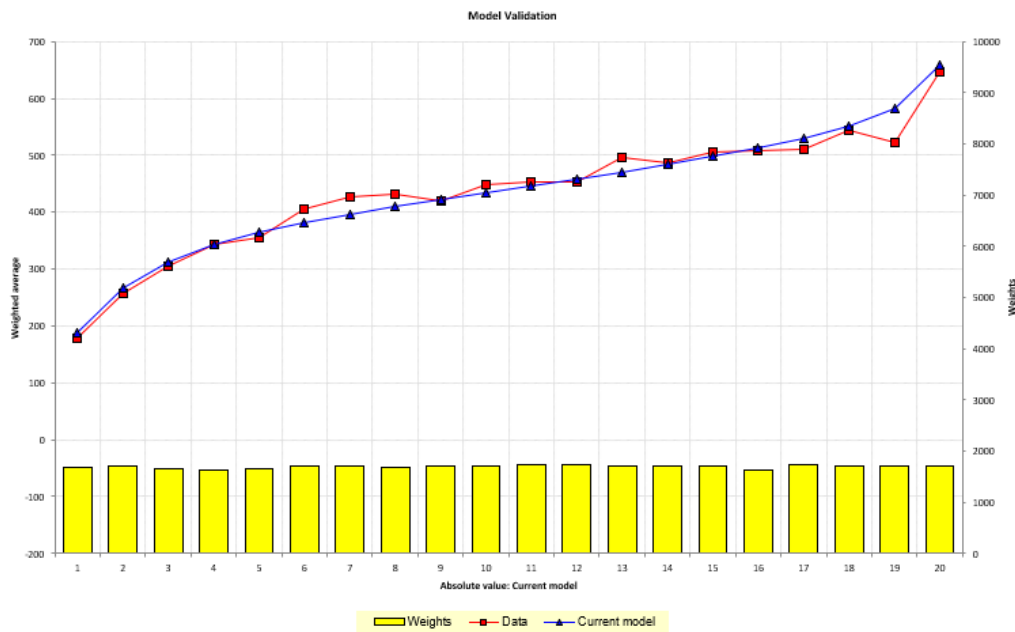


Figure 4.38 – Courbe de lift du modèle de coût moyen RCM non responsable

Etude des résidus

De même que pour la composante responsable, l'étude des résidus du modèle de coût moyen RCM non responsable (nuage de points concentré sur l'intervalle $[-2,2]$ et centré en 0) permet de valider les différentes hypothèses du modèle.

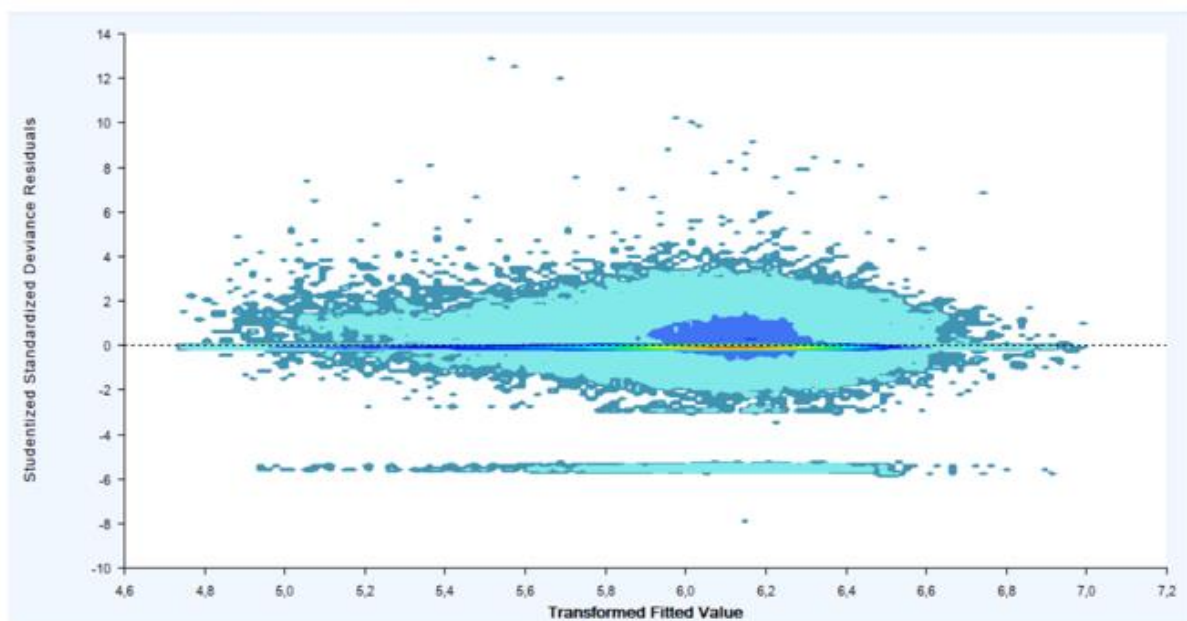


Figure 4.39 – Résidus de la déviance du modèle de coût moyen RCM non responsable

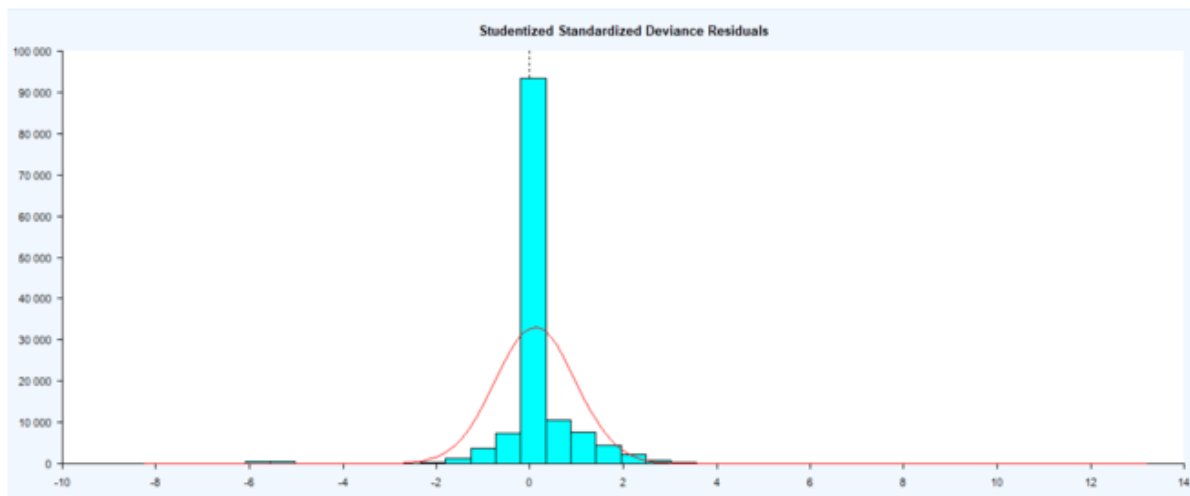


Figure 4.40 – Histogramme des résidus du modèle de coût moyen RCM non responsable