IRM | institut du risk management BY Institut.

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

Par : **Daria SERGEEVA**

Titre : **Recherche d'optimisation de tarifs d'assurance moto**

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 :

Jérôme SCHAEFFER

Michel NANG

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

Nicolas BARADEL

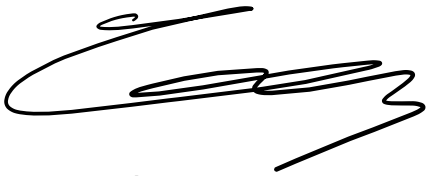
Secrétariat :

Bibliothèque :

Entreprise : Wakam

Nom : Christophe NEVES

Signature et Cachet :



Directeur de mémoire en entreprise :

Nom : Christophe NEVES

Signature :

Invité :

Nom :

Signature :

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

Signature du responsable entreprise

Signature(s) du candidat(s)

SAS au capital de 96 550 euros - Déclaration d'activité enregistrée sous le n° 11 75 44451 75 auprès du Préfet de Région IDF
4, rue Chauveau-Lagarde - 75008 Paris - Tél : 01 44 51 72 79 - Fax : 01 44 51 72 73
Email : formation@institutdesactuaires.fr
Siret : 512 264 441 000 19 - Code NAF : 8559 A - TVA intracommunautaire : FR 25 512264441 - RCS Paris : 512 264 441

TABLE DES MATIERES

REMERCIEMENTS

Je tiens à exprimer ma gratitude envers Alexis Pennes et Christophe Neves, pour l'aide et les encouragements.

Merci à l'équipe de la Direction Technique Wakam (Charlotte, Bertrand, Antoine, Hiba) pour le partage des connaissances et les conseils précieux.

Merci à Olivier Vermassen pour son aide sur l'utilisation d'Akur8.

Merci à mes amis de la promotion CEA 2019 – Josh, Cécile, Thibault et Béatrice, pour les échanges et la motivation.

Enfin, merci à ma famille pour sa patience.

RESUME

Mots clés : Assurance roulant, Tarification, Modèle linéaire généralisé, GLM, Modèle fréquence-coût, Optimisation d'un portefeuille d'assurance

Le but de ce mémoire est d'examiner les possibilités d'optimisation de rentabilité de plusieurs portefeuilles d'assurance deux-roues en France. En premier lieu à travers le changement du modèle de tarification, ensuite à travers l'adaptation du produit. Afin d'alléger la présentation, les résultats sont restreints à la garantie responsabilité civile. Une proposition de la tarification alternative repose sur une analyse ACM et segmentation de la population des assurées en profil types. Deux modèles sur la base de GLM sont explorés (fréquence x cout et prime pure Tweedie). Les modèles obtenus sont confrontés à la tarification actuelle et les critères de prise de décision de changement de la tarification. L'impact de la hausse tarifaire sur la rentabilité est évalué à travers une analyse de sensibilité de la rentabilité au taux de transformation et de renouvellement. Dans l'hypothèse où le changement de la tarification n'est pas justifié, nous dressons un panorama des adaptations du produit envisageables, toujours dans l'objectif d'améliorer la rentabilité.

ABSTRACT

Keywords : Motor insurance, Pricing, General linear model, GLM, Frequency-severity model, Insurance portfolio optimization.

The purpose of this dissertation is to examine the possibilities for optimizing the profitability of several two-wheel insurance portfolios in France. First through the change of the pricing model, then through the adaptation of the product. In order to lighten the presentation, the results are limited to the civil liability guarantee. A proposal for alternative pricing model is based on an ACM analysis and segmentation of the insured population into typical profiles. Two GLM-based models are explored (frequency x severity and pure premium Tweedie). The models obtained are confronted to the current pricing and the decision making criteria for pricing change. The impact of the tariff increase on profitability is assessed through an analysis of profitability sensitivity to the rate of conversion and renewal. In the event that the change in pricing is not justified, we provide an overview of possible product adaptations, always with the aim of improving profitability.

INTRODUCTION A LA PROBLEMATIQUE

Ce mémoire porte sur la recherche d'optimisation de tarification d'un produit moto. Dans un contexte de forte concurrence, cette gamme de produits classiques comporte une garantie obligatoire, la responsabilité civile, mais également une multitude de garanties facultatives. La prise en compte de ces garanties et des différences dans les limites d'indemnisation et les franchises rend une analyse transverse complexe et tout benchmark entre les assureurs difficilement réalisable pour les assurés et les assureurs eux-mêmes.

Une complexité supplémentaire s'ajoute lorsque le modèle de distribution de l'assureur prévoit un appel à des intermédiaires. Pourtant ce mode de distribution, selon l'analyse de la FFA, représente 52% du total des cotisations des assurances de bien et des responsabilités en France.

Lorsque le résultat technique d'un portefeuille historique n'affiche pas le niveau de rentabilité attendu, l'assureur peut appliquer une hausse tarifaire, avec le risque de faire fuir les « bons » assurés et provoquer de l'antisélection. Une adaptation du produit semble être une option plus complexe et coûteuse mais qui permet, en cernant mieux les besoins des assurés, de proposer un meilleur rapport qualité-prix.

L'objectif de ce mémoire est, à travers une étude de rentabilité de portefeuilles de plusieurs courtiers et une analyse segmentée des résultats, de proposer une tarification améliorée et des mesures de redressement des portefeuilles favorisant la rétention des assurés.

La première partie du mémoire consiste en une analyse des portefeuilles existants, dans l'objectif de répartir la population des assurés en un nombre de profils-types à l'aide d'une ACM. Puis, une étude segmentée permettra d'identifier les écarts de rentabilité éventuels entre les profils.

La deuxième partie propose une tarification alternative par profil-type en utilisant deux modèles GLM alternatives - fréquence x coût moyen et détermination de la prime pure directement en utilisant la distribution Tweedie. L'impact tarifant de la variable profil-type dans les deux modèles est testé sur le portefeuille global étudié. L'objectif étant de déterminer si la nouvelle tarification permet une meilleure adéquation au risque. Cette partie inclut également une validation de seuil de sinistralité attritionnelle et l'ajustement des données techniques pour tenir compte de l'inflation des coûts de sinistres par rapport aux données historiques utilisées.

Ces modélisations seront décrites sur l'exemple de la garantie responsabilité civile. Nous utiliserons cette garantie car, en tant que garantie obligatoire, elle est acquise par l'ensemble des assurés des portefeuilles étudiés. Cela permet d'avoir le maximum d'observations possible pour confirmer ou infirmer l'intérêt de notre analyse par profil-type. Si l'apport d'analyse par profil-type à la prédiction est confirmé, il conviendra de revalider qu'il est aussi important pour les autres garanties, en vérifiant garantie par garantie. En cas de résultats positifs, les autres garanties seront modélisées selon les mêmes principes pour définir le tarif du produit final.

Dans la troisième partie, la tarification obtenue est confrontée aux contraintes internes et externes et des méthodes alternatives de redressement de portefeuille sont explorées, telles que l'adaptation du produit ou la sélection des risques.

PREMIERE PARTIE :

ASSURANCE DEUX ROUES ET PRESENTATION DES PORTEFEUILLES

Dans cette première partie nous présentons d'abord le marché, le produit étudié ainsi que le contexte de l'entreprise. Nous nous positionnons du point de vue d'un assureur de taille moyenne, se situant juste derrière les 10 plus grands acteurs du marché d'assurance deux roues en France. Cela implique plusieurs contraintes métier comme la disponibilité limitée des données, la rareté des ressources et la relativement faible force de frappe commerciale pour faire concurrence aux leaders du marché. Dans ce contexte que nous pouvons qualifier de situation d'information imparfaite, nous analyserons les données disponibles pour l'étude d'optimisation du tarif et proposerons un moyen de segmentation plus fine.

1. Le marché de l'assurance deux roues en France

1.1. Chiffres clés

Selon la Fédération Française de l'Assurance (FFA), l'assurance automobile, segment dont l'assurance deux roues fait partie, représente presque la moitié des cotisations d'assurances de bien et de responsabilité. Sur un total de 22,8Md€ de cotisations d'assurance automobile en 2019. Les deux roues motorisés (hors flottes) représentent 4,8% de ce marché, soit 1,1Md€ de cotisations en 2019 (en hausse de 4% par rapport à 2018).

Selon les statistiques annuelles agrégées de la FFA pour les véhicules de 3ème catégorie [18], composée principalement de véhicules à deux roues motorisés, le ratio S/P hors commission en 2019 s'élève à 83%, niveau le plus bas depuis l'année 2014 (Figure 1).

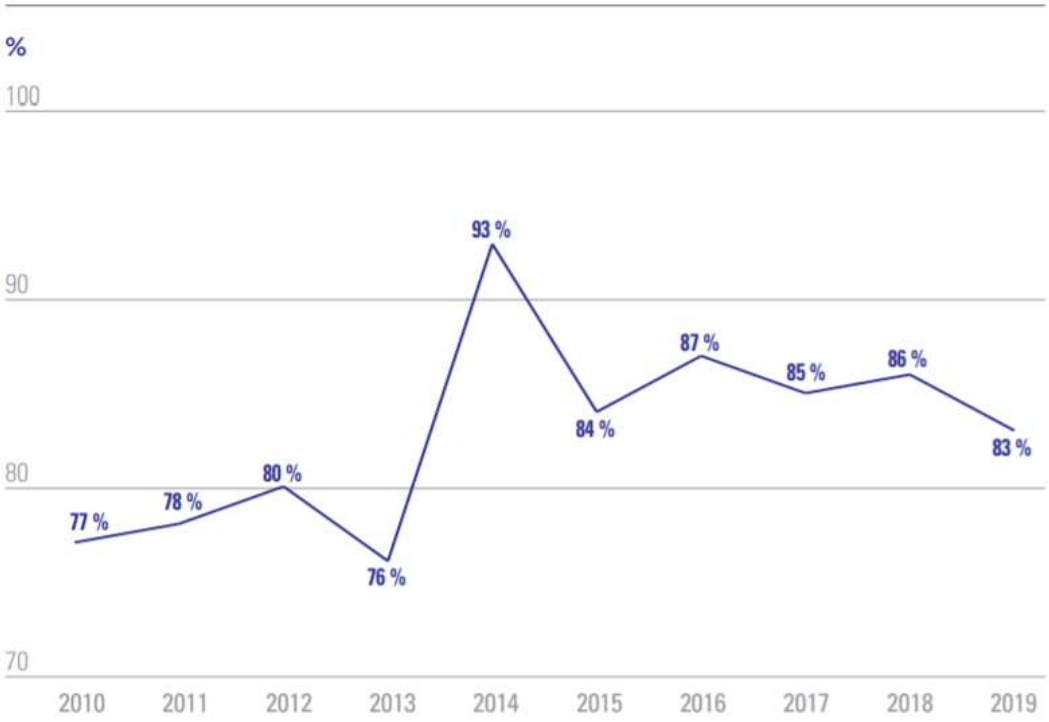


Figure 1 - Ratio sinistres à primes (Véhicules de 3e catégorie)

Au sein de la catégorie deux roues se trouvent les sous-catégories de cyclo, moto mais également ce qu'on appelle NVEI (Nouveaux Véhicules Electriques Individuels). Cette abréviation désigne les trottinettes électriques, gyroroues, gyropodes, hoverboards et tout autre bolide électrique et individuel. Notre étude se focalisera sur la sous-catégorie moto.

1.2. Garanties proposées et formules

De multiples garanties sont proposées sur le marché de l'assurance deux-roues en France.

La garantie responsabilité civile, la seule obligatoire pour les deux-roues à moteur, permet l'indemnisation des dommages causés aux tiers par la faute du conducteur du véhicule ou d'un de ses passagers :

- Blessures ou décès d'un passager, d'un piéton ou de l'occupant d'un autre véhicule…
- Dégâts matériels aux autres scooters, motos, deux-roues, voitures, immeubles

Cette garantie couvre les conducteurs autorisés ou non autorisés. Après avoir indemnisé les victimes, l'assureur peut disposer d'un recours à l'encontre des conducteurs non autorisés.

Dans sa publication « L'assurance des motos, scooters et autres deux-roues à moteur », la FFA dresse un panorama des garanties complémentaires, dont l'étendue peut varier selon les contrats et les sociétés d'assurances :

Les dommages subis par le deux-roues :

- La garantie dommages tous accidents (DTA) – cette garantie couvre tous les dommages matériels subis par le deux-roues, quel que soit le type d'accident ou la faute commise par son conducteur.
- La garantie dommages collision quant à elle ne joue qu'en cas de collision avec un piéton, un autre véhicule ou un animal dont le propriétaire est identifié.
- La garantie Accessoires permet d'assurer la prise en charge de certains accessoires hors-série.
- Les garanties incendie et vol permettent de recevoir une indemnité égale à la valeur du véhicule le jour de l'incendie ou du vol, ou à une valeur précisée dans le contrat. Comme le précise la FFA, en principe, la garantie incendie inclut aussi l'indemnisation des conséquences d'une explosion, de la chute de la foudre ou d'une combustion spontanée.

  Le contrat d'assurance définit les conditions d'application de la garantie vol ainsi que les modalités d'indemnisation. L'assureur peut exiger des mesures de prévention contre le vol (gravage des pièces, antivol « U », alarme, remise du deux-roues à moteur la nuit dans un garage fermé à clef …).

  Le vol d'accessoires et de pièces de rechange n'est, le plus souvent, pas couvert. Il est parfois couvert lorsqu'il est associé à une tentative de vol du deux-roues ou en cas d'effraction du local dans lequel le véhicule est garé.

Les garanties obligatoirement attachées aux garanties dommages facultatives :

Si le contrat comporte une garantie dommages au véhicule (tous accidents, dommages collision, vol, incendie…), le deux-roues est automatiquement couvert en cas de catastrophe naturelle, catastrophe technologique et attentat :

- Si le deux-roues est assuré par une garantie dommages au véhicule, les dégâts causés par une catastrophe naturelle (inondation, avalanche, tremblement de terre…), seront indemnisés après parution au Journal Officiel de l'arrêté interministériel constatant l'état de catastrophe naturelle. Une franchise de 380 euros est applicable.
- La garantie catastrophes technologiques couvre les dommages résultant des catastrophes technologiques ayant fait l'objet d'un arrêté au Journal Officiel. Les dommages sont alors réglés sans franchise.
- La garantie attentat couvre les dommages résultant d'actes de terrorisme et d'attentats commis sur le territoire national, la franchise applicable sera la même que celle prévue par la garantie incendie.

Les garanties non attachées aux garanties dommages facultatives :

- La garantie émeutes et mouvements populaires,
- La garantie forces de la nature joue en cas d'événements naturels non officiellement déclarés catastrophes naturelles. La garantie tempête, comme son nom l'indique, indemnise les dégâts causés la tempête mais la loi ne définit pas ce terme. Il revient à l'assureur de définir, dans son contrat, les conditions de mise en œuvre de cette garantie.

Les options spécifiques pour les motards :

- La garantie du casque ;
- La garantie des accessoires hors-série en cas de vol total du véhicule ou d'un dommage (top case, bulle…) ;
- La garantie contre la détérioration des équipements de protection du motard (gants, blouson, combinaison) à la suite d'un accident couvert par une garantie dommages ;
- La garantie des bris d'optiques (éléments vitrés du deux-roues) ;
- Une option intempéries est parfois proposée : elle tient compte de l'utilisation saisonnière de la moto. La période de non-utilisation, le plus souvent comprise entre trois et six mois, est fixée à la souscription.

La garantie du conducteur :

Lorsque le conducteur est blessé lors d'un accident de la circulation dans lequel il est fautif ou dans lequel aucun responsable n'est désigné, la garantie du conducteur lui permet d'être indemnisé. Par exemple, cette garantie prend en charge, selon les contrats d'assurance :

- Les frais médicaux, chirurgicaux, pharmaceutiques, d'hospitalisation et de prothèses qui ne sont pas pris en charge par la Sécurité sociale (dépenses médicales, frais d'hospitalisation, de transport, d'assistance à domicile...) ;
- Le préjudice financier lié à un arrêt de travail ou à une atteinte permanente à l'intégrité physique ;
- Le préjudice des ayants droit consécutif au décès.

Les assureurs proposent deux formules de garanties :

- De type forfaitaire, avec des capitaux fixés par le contrat en cas d'incapacité permanente ou de décès. Sans précision particulière au contrat, les prestations s'ajoutent aux indemnités qu'un tiers responsable peut être amené à verser ;
- De type indemnitaire, avec une indemnisation de l'ensemble du préjudice de l'assuré, avec, parfois, des plafonds de garantie et une franchise.

Les garanties de services :

- La garantie de protection juridique peut être proposée soit dans le contrat d'assurance automobile, soit dans un contrat autonome. Différents niveaux de garantie peuvent être proposés :
  - La garantie défense pénale et recours suite à accident. Cette garantie permet à l'assuré d'exercer un recours contre le responsable d'un accident lorsque l'assuré n'est pas responsable. Elle couvre également les frais d'avocat lorsque l'assuré se porte partie civile à un procès faisant suite à un sinistre garanti ;
  - La garantie de protection juridique segmentée, qui couvre un domaine d'intervention précis (avec une liste des risques couverts) en lien avec le deux-roues à moteur, comme les frais de procès engagé à la suite de mauvaises réparations par un garagiste ;
  - La garantie de protection juridique générale englobe les risques juridiques de toute nature qui relèvent du droit contractuel, du droit de la consommation, du droit administratif, du droit pénal, du droit immobilier, du droit numérique… ;

- La garantie assistance, pour le véhicule et pour les passagers, permet d'être dépanné et remorqué en cas de panne ou d'accident.

  De nombreux contrats d'assurance comprennent aussi l'envoi de pièces détachées, les frais d'hébergement pendant la durée de la réparation ou de conduite à destination, les frais de récupération du véhicule et le paiement d'une caution à l'étranger.

  L'assistance aux passagers inclut généralement le rapatriement en cas d'accident ou de maladie, le remboursement des frais médicaux engagés à l'étranger, le rapatriement du corps en cas de décès.

Dans la pratique, les garanties principales sont souvent analysées en 3 groupes ou formules – Responsabilité Civile, Dommages tous accidents et Vol (cette dernière garantie est souvent vendue avec la garantie Incendie), accompagnés par une ou plusieurs garanties incluses d'office ou en option.

Le tableau 1 présente des statistiques annuelles agrégées de fréquence et coûts moyens publiés par la FFA (L'assurance Française : Données clés 2019, [17]). Ces statistiques nous seront très utiles dans l'analyse, notamment pour vérifier la cohérence des résultats obtenus.

Variations		2015	2016	2017	2018	2019	Niveau 2019
RC Matériel	Fréquence	- 2,6 %	- 4,2 %	- 1,9 %	- 6,2 %	- 4,9 %	7,7 ‰
	Coût moyen	+ 3,1 %	+ 10,3 %	- 7,2 %	+ 5,8 %	- 1,3 %	1 270 €
RC Corporel	Fréquence	- 7,0 %	- 1,9 %	- 2,4 %	- 6,4 %	- 7,4 %	2,5 ‰
	Coût moyen	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.
Dommages tous accidents	Fréquence	- 1,6 %	- 1,4 %	+ 0,7 %	+ 1,1 %	+ 0,0 %	26,7 ‰
	Coût moyen	+ 5,8 %	+ 3,8 %	+ 3,6 %	+ 2,3 %	+ 4,4 %	2 000 €
Vol	Fréquence	- 4,5 %	- 5,4 %	- 4,3 %	- 9,5 %	+ 3,1 %	8,8 ‰
	Coût moyen	+ 0,7 %	+ 2,7 %	+ 6,5 %	+ 6,8 %	+ 7,0 %	2 540 €

Tableau 1 – Statistiques FFA : Fréquence et coût moyen des sinistres par garantie (Véhicules de 3ème catégorie)

1.3. Environnement réglementaire

Il est intéressant de considérer les contraintes règlementaires qui peuvent avoir une incidence sur le produit d'assurance, la modélisation et, par conséquent, sur la rentabilité de l'assureur.

Age du conducteur

Tout d'abord, en fonction de la motorisation du deux-roues, l'âge minimal du conducteur et le type de permis requis ne seront pas les mêmes.

- Il est possible de conduire dès l'âge de 14 ans des cyclomoteurs dont la cylindrée n'excède pas 50 cm³.
- A partir de 16 ans, il est possible de conduire une motocyclette légère d'une cylindrée maximale de 125 cm³, à condition d'avoir un permis A1.
- A partir de 18 ans, le conducteur peut accéder en plus à un deux-roues d'une puissance n'excédant pas 35 kW (ou 47,6 CV, ce qui correspond à une cylindrée de 750 à 900 cm³), à condition d'avoir un permis A2.
- A partir de 20 ans toute cylindrée est accessible, à condition d'avoir un permis correspondant.

Limites d'indemnisation

Certaines garanties sont plus strictement encadrées que d'autres :

Pour la garantie responsabilité civile, en règle générale, les dommages corporels de la victime sont indemnisés sans aucune limite tandis que les dommages matériels sont soumis à un plafond de remboursement établis par l'assureur. La procédure et les droits sont encadrés par la loi Badinter (loi n°85-677 du 5 juillet 1985).

Pour les garanties optionnelles, les plafonds de garanties et les franchises sont établis librement lors de la création du produit par l'assureur. Cependant, l'environnement concurrentiel joue un rôle non négligeable dans ces considérations.

Possibilité de résilier le contrat d'assurance

Depuis 2015, la loi sur la consommation, dite « loi Hamon », a entériné la possibilité pour l'assuré de résilier un contrat d'assurance, à tout moment et sans justification, passé un an de contrat.

Fiscalité des produits d'assurance :

Les cotisations d'assurance ne sont pas soumises à la TVA. Elles incluent en revanche une taxe fiscale, dont le taux légal varie selon la nature des contrats d'assurance. Les montants des taxes collectées sont reversés au Trésor Public. La cotisation comprend également des contributions qui servent à financer divers fonds ou organismes de solidarité nationale auxquels elles sont reversées.

Par exemple, la taxe fiscale pour la garantie responsabilité civile obligatoire s'élève à 33% de la cotisation nette, avec 15 % réattribués à la Sécurité sociale. A ce taux s'ajoutent 2% de contribution Fonds de garantie des assurances obligatoires de dommages.

Les autres garanties relatives aux véhicules (dommages, assistance aux véhicules, pannes mécaniques...) sont soumises à une taxe de 18% à laquelle s'ajoute une contribution fixe de 5,9€ par contrat à destination du fonds de garantie des victimes des actes de terrorisme et d'autres infractions (FGTI).

Coefficient bonus/malus (CRM) :

L'article A121-1 du Codes des Assurances prévoit que les contrats d'assurance garantissant des véhicules terrestres à moteur (sauf exceptions prévues dans l'article R. 311-1) doivent comporter la clause de réduction ou de majoration des primes. Avec ce système, le conducteur qui ne cause pas d'accident bénéficie d'un bonus : sa prime de référence diminue. A l'inverse, l'automobiliste responsable d'un accident est pénalisé d'un malus : sa prime de référence augmente.

Les modalités de calcul du bonus-malus sont prévues par le Code des assurances. Comme le précise la FFA dans son article [20], bonus et malus sont exprimés par des coefficients de réduction ou de majoration, compris entre 0,50 et 3,50. L'avis d'échéance mentionne le montant de la prime de référence qui correspond au tarif de base de la société d'assurances applicable au véhicule concerné compte tenu de ses caractéristiques

techniques : type de véhicule, zone géographique de circulation ou de garage, usage socio-professionnel, kilométrage parcouru…

Chaque année, cette cotisation est multipliée par un coefficient de bonus-malus. Le coefficient d'origine est de 1 ; il est inférieur à 1 en cas de bonus et supérieur à 1 en cas de malus.

Chaque année sans sinistre engageant la responsabilité de l'assuré entraîne une réduction de 5 % du coefficient de l'année précédente. Pour obtenir le nouveau coefficient, le coefficient de l'année précédente est multiplié par 0,95. Le maximum est fixé à 0,50, ce qui correspond à un bonus de 50 %.

Tout accident dont l'assuré est totalement responsable entraîne une majoration de 25 % du coefficient précédemment appliqué. On obtient le nouveau coefficient en multipliant le précédent par 1,25.

Si un automobiliste provoque plusieurs accidents au cours de la même année, le coefficient de son bonus ou de son malus est multiplié par 1,25 autant de fois qu'il y a eu d'accidents, sans pouvoir excéder 3,50. En cas de partage de responsabilité et quel que soit le pourcentage de responsabilité retenu, on réduit la majoration de moitié (12,5 % au lieu de 25 %). Le coefficient de l'année précédente est alors multiplié par 1,125. Le malus disparaît après deux années d'assurance consécutives sans accident.

2. Présentation de l'entreprise et des portefeuilles

2.1. Présentation de l'entreprise

Anciennement appelée La Parisienne Assurances, Wakam est une société d'assurances qui distribue ses produits sous marque blanche, à travers un réseau des partenaires-courtiers. Dans le cadre de son business model, Wakam délègue la distribution et la gestion de la production et des sinistres à ses partenaires-courtiers. La plupart du temps, les produits sont créés sur mesure et peuvent être adaptés plusieurs fois au cours de la vie du partenariat. Wakam conserve systématiquement la main sur la tarification initiale et les mesures de renouvellements tarifaires proposées aux partenaires. La création de produits sur mesure, gérés et reportés depuis des outils hétérogènes, engendre des

difficultés pour l'analyse en transverse ainsi que pour la prise de décisions tarifaires et leur déploiement. Les portefeuilles qui vont être présentés dans la suite de ce document sont issus de partenariats distincts, lancés sur la période allant de 2012 à 2017.

2.2. Chiffres clés et présentation des portefeuilles pour analyse

Afin d'avoir un périmètre comparable, nous nous focaliserons sur le marché de la France et sur les deux roues de cylindrée strictement supérieure à 50 cm^3, ce qui exclut notamment les cyclomoteurs.

Les tableaux 2-4 ci-après présentent une analyse de rentabilité de 6 portefeuilles, toutes garanties et formules confondues sur les exercices de 2018 à 2020, vue au 31 décembre 2020. Les données sont exprimées en S/P dossier/dossier et n'incluent pas les sinistres survenus mais non déclarés (IBNR). Ceci explique un S/P plus faible pour l'exercice de survenance 2020, une partie significative des sinistres n'étant pas encore déclarée. La modélisation globale va devoir tenir compte des IBNR et des analyses des bonis-malis sur les exercices antérieurs. Il s'agit également d'une année marquée par le début de la crise COVID-19 et de plusieurs périodes de confinement. La fréquence des sinistres de cette année restera donc anormalement basse et la modélisation globale va également devoir tenir compte de ce biais. Pour ce faire, le delta du changement de la fréquence sera estimé sur la base des paramètres observables, comme le nombre de véhicules en circulation, et les coefficients du modèle seront à ajuster.

La structure de chaque portefeuille en détail sera analysée par la suite, afin d'identifier les impacts des particularités de chaque population des assurés sur la fréquence et les coûts dont il va falloir tenir compte lors de la modélisation.

Portefeuilles	A	B	C	D	E	F	TOTAL
Toutes Générations			Survenance 2020				
S/P TOTAL	**50,3%**	**21,7%**	**16,1%**	**194,9%**	**75,3%**	**45,0%**	**40,8%**
S/P < 30K	*41,4%*	*18,2%*	*14,7%*	*62,7%*	*75,3%*	*35,1%*	*34,5%*
S/P de 30K à 100K	*6,6%*	*3,5%*	*1,4%*	*31,8%*	*0,0%*	*9,9%*	*6,3%*
S/P > 100K	*2,3%*	*0,0%*	*0,0%*	*100,3%*	*0,0%*	*0,0%*	*0,0%*
Fréquence TOTAL	**1,8%**	**3,7%**	**4,8%**	**6,1%**	**7,7%**	**6,3%**	**3,3%**
Coût Moyen TOTAL	**3 232**	**1 478**	**962**	**8 652**	**2 625**	**1 953**	**2 760**
CM < 30K hors graves	*2 300*	*1 102*	*752*	*1 666*	*2 625*	*1 126*	*1980*

Tableau 2 - Rentabilité dossier/dossier - Exercice de survenance 2020

Portefeuille Toutes Générations	A	B	C	D	E	F	TOTAL
			Survenance 2019				
S/P TOTAL	**75,9%**	**47,0%**	**30,7%**	**76,7%**	**237,4%**	**98,1%**	**95,5%**
S/P < 30K	*47,9%*	*44,3%*	*30,7%*	*61,0%*	*131,2%*	*72,3%*	*63,3%*
S/P de 30K à 100K	*6,8%*	*2,6%*	*0,0%*	*15,7%*	*11,5%*	*25,8%*	*7,1%*
S/P > 100K	*21,1%*	*0,0%*	*0,0%*	*0,0%*	*94,7%*	*0,0%*	*25,0%*
Fréquence TOTAL	**2,0%**	**6,0%**	**6,9%**	**6,9%**	**11,4%**	**8,6%**	**4,9%**
Coût Moyen TOTAL	**3 838**	**2 025**	**1 337**	**2 878**	**5 595**	**3 254**	**3 584**
CM < 30K hors graves	*2 085*	*1 710*	*1 337*	*1 262*	*2 916*	*1 550*	*2 097*

Tableau 3 - Rentabilité dossier/dossier - Exercice de survenance 2019

Portefeuille Toutes Générations	A	B	C	D	E	F	TOTAL
			Survenance 2018				
S/P TOTAL	**103,8%**	**150,5%**	**77,4%**	**74,6%**	**174,1%**	**192,3%**	**126,9%**
S/P < 30K	*52,9%*	*56,0%*	*52,3%*	*55,6%*	*144,9%*	*69,9%*	*77,9%*
S/P de 30K à 100K	*11,3%*	*9,7%*	*9,2%*	*19,1%*	*21,8%*	*32,0%*	*14,8%*
S/P > 100K	*39,6%*	*84,8%*	*15,9%*	*0,0%*	*7,4%*	*90,4%*	*34,3%*
Fréquence TOTAL	**1,6%**	**7,6%**	**8,8%**	**8,2%**	**12,7%**	**10,7%**	**4,7%**
Coût Moyen TOTAL	**5 664**	**5 355**	**3 312**	**2 418**	**3 547**	**5 399**	**4 254**
CM < 30K hors graves	*2 555*	*1 775*	*1 929*	*1 338*	*2 533*	*1 599*	*2 250*

Tableau 4 - Rentabilité dossier/dossier - Exercice de survenance 2018

La difficulté d'analyse transverse consiste en le fait que chaque portefeuille concerne un ou plusieurs produits créés sur mesure, avec ses propres critères de souscription, formules incluant une ou plusieurs garanties optionnelles, avec ses propres plafonds de garantie.

2.3. Présentation des données

Pour chacun des portefeuilles, il existe une base de données de quittances et une base de sinistres. Les bases quittances contiennent les informations sur les assurés, les véhicules et les durées de couverture, ainsi que les montants de quittances par garantie. Les bases de sinistres contiennent les informations sur la date du sinistre, la garantie concernée et la charge. Les bases sont hétérogènes entre elles quant au nombre et au format des variables. Nous présenterons ci-après les variables d'intérêt ainsi que les variables explicatives, la structure des données ainsi que les statistiques descriptives relatives à ces bases. Les bases contenant un grand nombre de variables, ne seront présentées que celles qui seront intéressantes pour cette étude.

Certaines bases de quittances contiennent également l'indication de la formule souscrite (RC, VOL-Incendie, DTA). Les formules et leurs dénominations ne sont pas homogènes entre les portefeuilles, chacun des produits ayant été créé sur mesure. Les tableaux 5 et 6 font un état des lieux de la présence et du poids relatif de chacune des garanties dans les portefeuilles analysés, tout exercice confondu. La commercialisation de chacun des portefeuilles a débuté à une période différente, ce qui explique la différence de poids entre les portefeuilles.

Il en ressort, que les portefeuilles C, D et F sont assez petits – tous les trois combinés ils représentent 14% du total des données. Nous pouvons nous attendre à la volatilité plus importante sur ces portefeuilles. Par exemple, le portefeuille D présente un S/P de 195% en 2020, mais cette performance, du fait de la faible taille de ce portefeuille, est fortement impactée par quelques sinistres graves. Le portefeuille A, au contraire, représente 62% des données. Les conclusions de l'étude seront fortement impactées par le mix des assurés de ce portefeuille.

Afin d'avoir un socle commun, l'analyse se fera au niveau de garanties principales, présentes dans chacun des portefeuilles, à savoir la responsabilité civile (RCX), défense pénale recours (DRX), protection du pilote (PROPIL), vol (VLX) et dommages tout accident (DTA). Les définitions de l'ensemble des garanties sont présentées dans le glossaire.

Notons le poids particulièrement faible de la garantie responsabilité civile dans le portefeuille D, la moyenne dans les autres portefeuilles étant supérieure à 45%. Ce même portefeuille comporte presque toutes les garanties disponibles. Il s'agit du

portefeuille d'un courtier grossiste qui permet à ses courtiers-partenaires des produits modulables.

CodeGarantie	Tous Portefeuilles	Portefeuille A	Portefeuille B	Portefeuille C	Portefeuille D	Portefeuille E	Portefeuille F
EQUIMOT	0,6%	0,3%	0,6%	1,6%	2,4%	1,0%	1,4%
VLX	16,1%	18,1%	11,0%	8,0%	6,6%	21,3%	8,4%
PJ	2,6%	4,1%	0,1%	1,0%			0,0%
VNEUF24	0,0%	0,0%			0,1%		
CATECH	0,3%	0,3%	0,4%	0,0%	0,8%	0,3%	0,3%
PROPIL	8,0%	8,7%	5,7%	7,8%	7,5%	7,6%	6,3%
OBJTR	0,0%				0,2%		
DRX	11,6%	7,9%	24,1%	25,0%	13,3%	4,2%	25,1%
VNEUF36	0,0%				0,3%		
CATNAT	1,0%	1,1%	0,7%	0,6%	0,4%	1,3%	0,5%
CIR	0,0%				0,0%		
VNEUF6	0,0%				0,0%		
REMAC	0,0%				0,1%		
PFX	0,0%	0,0%					
RCX	45,8%	46,7%	44,5%	45,7%	29,5%	47,6%	46,5%
BDG	0,1%	0,1%			1,3%		
RFRA	0,0%				0,2%		
INC	0,0%				1,0%		
ACCESS	0,6%	0,8%			1,2%	0,8%	
CASQ	1,5%	1,5%	1,3%	1,1%	6,9%	0,3%	1,2%
VNEUF12	0,1%				0,9%	0,9%	
PTX	0,0%				0,0%		
DEMFR	0,1%				2,0%		
ATTEN	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
OPPROPIL	0,5%		0,6%	0,5%	6,8%	1,2%	0,7%
DCX	0,0%		0,2%	0,0%			0,1%
DTA	10,9%	10,4%	10,7%	8,6%	18,4%	13,5%	9,5%
TOTAL	100%	100%	100%	100%	100%	100%	100%

Tableau 5 - Présence et poids relatif des garanties par portefeuille

CodeGarantie	Tous Portefeuilles	Portefeuille A	Portefeuille B	Portefeuille C	Portefeuille D	Portefeuille E	Portefeuille F
TOTAL	100%	62%	14%	7%	4%	11%	3%

Tableau 6 - Poids relatif des portefeuilles étudiés, tout exercice confondu

Il existe également une base de données de zones géographiques (Zonier), qui présente les caractéristiques de statut urbain et socio-économique des communes françaises.

Il est à noter que dans le respect de la règlementation RGPD les données utilisées sont anonymisées et ne contiennent pas d'éléments permettant d'identifier les individus.

2.4. Préparation des données et analyse des variables explicatives

Pour pouvoir comparer les portefeuilles et identifier les critères de segmentation, plusieurs retraitements des bases ont été nécessaires :

- A partir des données de quittances par garantie, construction d'une base des polices par garantie, identifiant l'année d'effet et la durée d'exposition par police. Une attention particulière a été portée au calcul de l'exposition des polices résiliées en cours d'année, en s'appuyant sur la date de résiliation,
- Retraitement des variables explicatives, en particulier des dates, pour homogénéiser leur format et les présenter en nombre d'années (âge du conducteur, ancienneté du permis et du véhicule),
- Rattachement à chaque base de polices obtenue les données des sinistres survenus sur les mêmes garanties et la bonne période de couverture, à partir des données sinistres,
- Rattachement des données de statut urbain de la commune à partir du Zonier.

Les variables explicatives retenues sont les suivantes :

Pour le véhicule :

- Code postal de stationnement
- Marque et modèle
- Code, Groupe et Classe SRA
- Puissance et cylindrée
- Ancienneté de mise en circulation

Pour le conducteur :

- Coefficient bonus-malus
- Ancienneté du permis
- Age du conducteur

Cet ensemble de variables sera utilisé pour expliquer les différences de rentabilité et construire les tarifs.

Il est à noter que la totalité des portefeuilles ne dispose pas de l'ensemble des variables. Aucune des bases ne contient de données concernant les modalités de stationnement (garage/voie publique/autre) et l'usage du véhicule (privé/professionnel). Nous ne disposons également pas de données sur le sexe du conducteur. Pour le portefeuille E, la base de données ne dispose pas de CRM des assurés – une variable obligatoire règlementairement mais non transmise dans le cadre du reporting des polices souscrites envoyé par le courtier.

Tout ceci introduit un biais de sélection dans les données utilisées. En effet, nous sommes conscients que les données dont nous disposons et qui seront utilisées pour notre étude tarifaire ne sont pas une représentation fiable du marché. Un exemple donné par Rain et Jacques (2013) [7] illustre cet effet : une compagnie qui a historiquement utilisé une variable tarifaire de moins que ses concurrents et de ce fait n'était pas compétitive sur les bonus 50. En l'absence d'une variable, cette compagnie a eu tendance à vendre des contrats aux plus mauvais risques parmi les bonus 50. Une modélisation tarifaire réalisée à partir d'une base de données des assurés d'une telle compagnie ne reflètera pas complétement l'ensemble des comportements possible pour des assurés ayant un bonus 50, mais seulement une partie.

Le tableau 7 fait état de la disponibilité des données des variables pour chacun des portefeuilles.

Portefeuille	Formule Souscrite	Ville Souscripteur	Code Postal Souscripteur	VEHICULE CODE POSTAL GARAGE	VEHICULE MARQUE	VEHICULE MODELE	VEHICULE CODE_SRA	VEHICULE GROUPE_SRA	VEHICULE CLASSE SRA	VEHICULE PUISSANCE REELLE	VEHICULE CYLINDREE	CONDUCTEUR PRINCIPAL CRM SOUSCRIPTION	CONDUCTEUR PRINCIPAL CRM AUTRE	CONDUCTEUR PRINCIPAL DATE PERMIS	VEHICULE DATE MISE EN CIRCULATION	CONDUCTEUR PRINCIPAL DATE NAISSANCE
A	OUI	OUI	OUI	OUI	OUI	OUI	NON	OUI	OUI	OUI	OUI	OUI	OUI	OUI	OUI	OUI
B	OUI	NON	NON	NON	OUI	OUI	OUI	NON	NON	NON	OUI	OUI	OUI	OUI	OUI	NON
C	OUI	NON	NON	NON	OUI	OUI	OUI	NON	NON	NON	OUI	OUI	OUI	OUI	OUI	NON
D	OUI	OUI	OUI	OUI	OUI	OUI	OUI	OUI	OUI	OUI	OUI	OUI	OUI	OUI	OUI	OUI
E	NON	OUI	OUI	OUI	OUI	OUI	OUI	OUI	OUI	NON	OUI	NON	NON	OUI	OUI	OUI
F	OUI	NON	NON	NON	OUI	OUI	OUI	NON	NON	NON	OUI	OUI	OUI	OUI	OUI	NON

Tableau 7 - Disponibilité des données des variables descriptives par portefeuille

Lorsqu'une seule modalité d'une variable est présente dans la base, les données de cette variable sont considérées comme absentes dans le tableau 8 mais ne sont pas éliminées.

Plusieurs anomalies et valeurs aberrantes ont été mises en évidence au fil de l'analyse :

- Quelques données manquantes (âge du conducteur, date de mise en circulation du véhicule, coefficient de bonus-malus) systématiquement ou aléatoirement – ces données ont été maintenus et attribués la catégorie « non disponible ». L'objectif est d'analyser les coefficients tarifants des sorties GLM pour cette catégorie et pouvoir ajuster les coefficients des autres catégories en fonction. Il est entendu que la tarification définitive n'autorisera pas de données manquantes.
- Données non homogènes (présence de motos à très faible cylindrée dans un des portefeuilles) – compte tenu de la proportion importante de ces contrats (25% du portefeuille concerné) ces données ont été écartées. Nous avons considéré que ces faibles cylindrées doivent faire l'objet d'un modèle spécifique (nature de risque différente).
- Sinistres sans numéro de police (6% du coût total des sinistres) – ces données n'ont pas pu être rattachées aux données contrats. Ils ont été écartés. Le montant global de ces sinistres sera affecté sous forme de chargement additionnel fixe à l'ensemble des polices à la fin de la tarification.
- Données incohérentes sur certaines polices (Age du conducteur ou de permis dépassant 80 ans) - afin de ne pas fausser l'analyse, les données de ces variables sont remplacées pour apparaître comme non disponibles – 0,04% des polices concernés.

Nous avons vérifié la cohérence des dates de permis par rapport à la date de naissance du conducteur et la cohérence de l'ancienneté de la mise en circulation du véhicule (3% des observations ont été écartées par suite de ce contrôle). Il est important de mentionner qu'il y a certainement d'autres incohérences dans les données, non détectées, liées à la saisie manuelle.

Little et Robin (1987) [13] ont développé une typologie répartissant les données manquantes en trois catégories distinctes :

Missing Completely At Random – la probabilité qu'une observation de ce type soit manquante ne dépend ni de variables observées ni de variables non-observées. Ce type

de valeurs manquantes, si leur nombre est restreint, peut être ignoré sans entrainer de biais significatif dans le modèle.

Missing At Random - La probabilité qu'une observation soit manquante ne dépend que des valeurs observées. Ce type de données entraîne un biais conséquent dans l'analyse et pourrait dans certains cas être traité par des méthodes d'imputations i.e. en estimant ces observations.

Missing Not At Random - La probabilité qu'une observation soit manquante dépend des valeurs non observées. Ce type de données manquantes implique un biais et une perte de précision du modèle, nécessitant une analyse de sa sensibilité.

Trois possibilités s'offrent pour le traitement des données manquantes :

- Eliminer l'observation – une solution qui peut être appliquée lorsque la proportion des données manquantes est très faible. Dans le cas contraire elle entrainera des biais dans le modèle
- Remplacement des données manquantes par les valeurs estimées comme proches. Plusieurs méthodes plus ou moins complexes existent : remplacement par la moyenne, par la dernière observation. Des méthodes d'intelligence artificielle peuvent également être utilisées, et notamment la méthode des k plus proches voisins. Cette méthode consiste à affecter une classe à une nouvelle donnée à partir d'un ensemble de données labellisées et de sa position par rapport à ces données labélisées. La nouvelle donnée est affectée à la classe la plus fréquente parmi les *k* données les plus proches.
  Les approches de remplacement des données ont été critiquées notamment dans l'article de Schafer et Graham (2002) [15] pour le manque de précision qu'entraînent ces règles.
- Utilisation des algorithmes spécifiques, pas sensibles aux valeurs manquantes, pour modélisation tarifaire, par exemple les algorithmes d'arbres de régression, dont CART.

Afin de déterminer le traitement adapté pour chacun des types de données manquantes dans l'analyse qui nous concerne, le nombre de valeurs présentes a été comparé au nombre total des observations dans la base (Tableau 8). Les cellules en surbrillance rouge mettent en évidence les variables avec moins de 95% de données disponibles. Les cellules en police rouge montrent les variables ayant une seule modalité et donc de facto également absentes. Les cases rouges sont des variables absentes (0%).

Portefeuille	Nombre d'observations total du Portefeuille	Poids du Portefeuille dans le total des observations	VILLE_SOUSCRIPTEUR	CODE_POSTAL_SOUSCRIPTEUR	CODE_POSTAL_GARAGE	MARQUE	MODELE	CODE_SRA	GROUPE_SRA	CLASSE_SRA	PUISSANCE_REELLE	CYLINDREE	CRM_SOUSCRIPTION	CRM_AUTRE	ANC_PERMIS	ANC_MEC	AGE
A	393 261	58%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
B	101 557	15%				100%	100%	58%	100%	100%		100%	100%	100%	100%	100%	
C	45 776	7%				100%	100%	56%	100%	100%		100%	100%	100%	100%	100%	
D	35 427	5%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
E	82 476	12%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%		100%	72%	100%	72%
F	19 375	3%				100%	100%	55%	100%	100%		100%	100%	100%	100%	100%	
TOTAL	677 872	100%															

Tableau 8 – Disponibilité des données par variable

Les variables géographiques (Ville souscripteur, Code postal garage), d'âge et de véhicule ne peuvent pas être remplacées sur la base des autres variables. Du fait du poids de ces variables manquantes, leur remplacement aurait engendré un biais très important. Nous garderons donc les observations manquantes en les identifiant par « NonCommuniqué ».

2.5. Statistiques descriptives

La multitude de variables ne pouvant pas être présentée de manière exhaustive dans ce mémoire, il est proposé de se focaliser sur les principales, à savoir l'âge, l'ancienneté du permis et le coefficient bonus-malus (CRM) du conducteur, la cylindrée et l'ancienneté du véhicule, à travers la garantie responsabilité civile, présente par défaut dans tous les portefeuilles. Dans la deuxième partie de ce mémoire, nous étudierons également l'impact sur la modélisation des variables de zoning (i.e. code INSEE et le regroupement associé).

Nous utiliserons cette garantie car, en tant que garantie obligatoire, elle est acquise par l'ensemble des assurés des portefeuilles étudiés. Cela permet d'avoir le maximum d'observations possible pour confirmer ou infirmer l'intérêt de notre analyse par profil-type. Si l'apport d'analyse par profil-type à la prédiction est confirmé, il conviendra de revalider qu'il est aussi important pour les autres garanties, en le vérifiant garantie par garantie. En cas de résultats positifs, les autres garanties seront modélisées selon les mêmes principes pour définir le tarif du produit final

Après retraitements des données par exercice de rattachement, cette base comporte les informations sur 136 676 polices, 225 473,7 années-risques étalées sur les années d'effet de 2012 à 2020 et 1685 sinistres dont la charge est supérieure à 0.

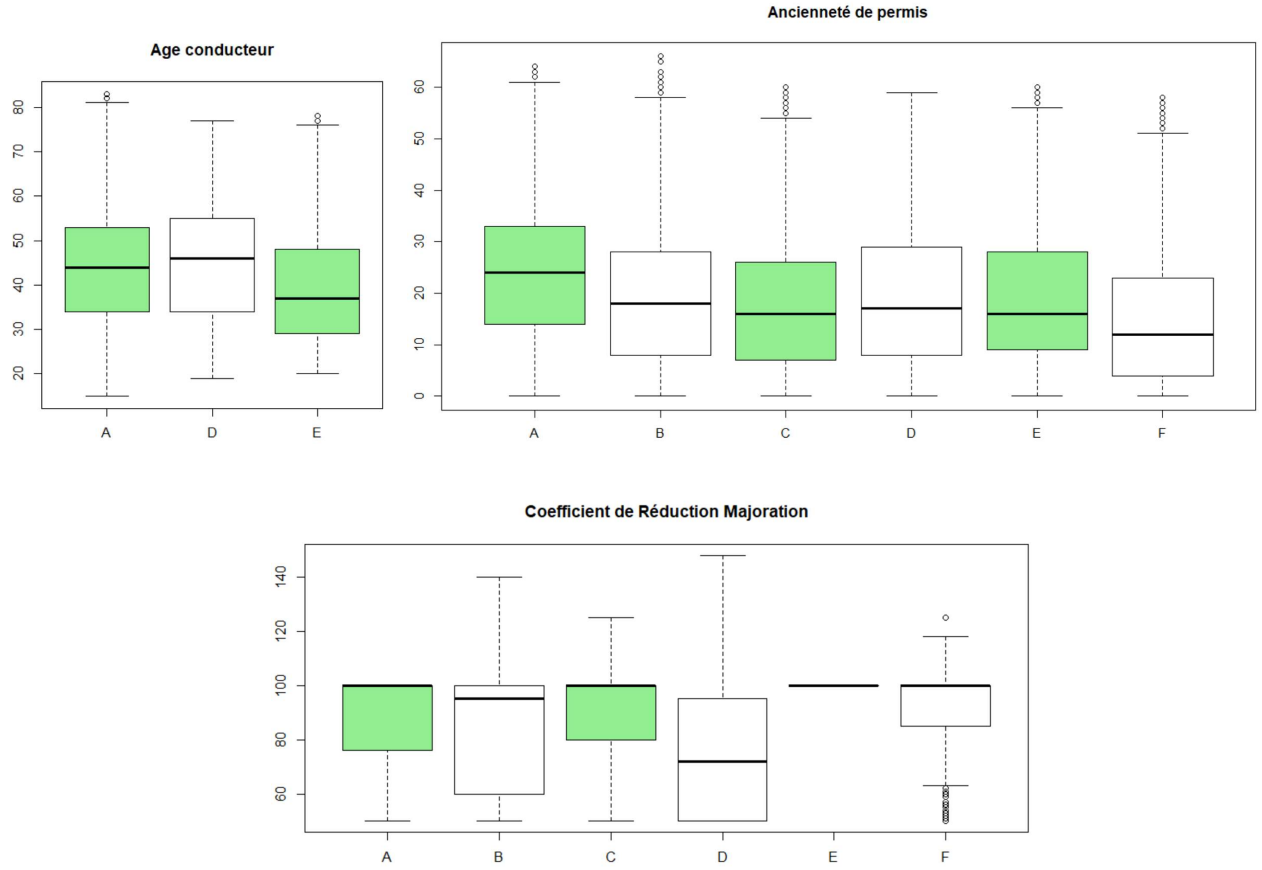


Figure 2 - Statistiques descriptives des variables explicatives du conducteur

La figure 2 présente les statistiques des conducteurs. La valeur centrale de chaque graphique est la médiane (il existe autant de valeurs inférieures que de valeurs supérieures à cette valeur dans chaque portefeuille). Les bords des boîtes sont de quartiles. 50% des observations se retrouvent à l'intérieur de la boite. Notons que portefeuille D se distingue par le CRM médian beaucoup plus faible, ce qui semble cohérent avec la médiane d'age des conducteurs plus élévée. Pour rappel, le portefeuille E ne dispose pas de données de CRM des conduteurs.

L'ancienneté de permis est la seule variable de conducteur présente dans tous les portefeuilles. Elle s'établit entre 15 et 18 ans, sauf pour le portefeuille F qui est plus

jeune, avec la médiane de l'ancienneté de permis légerement supérieure à 10 ans, ce qui est cohérent avec un CRM médian proche de 100.

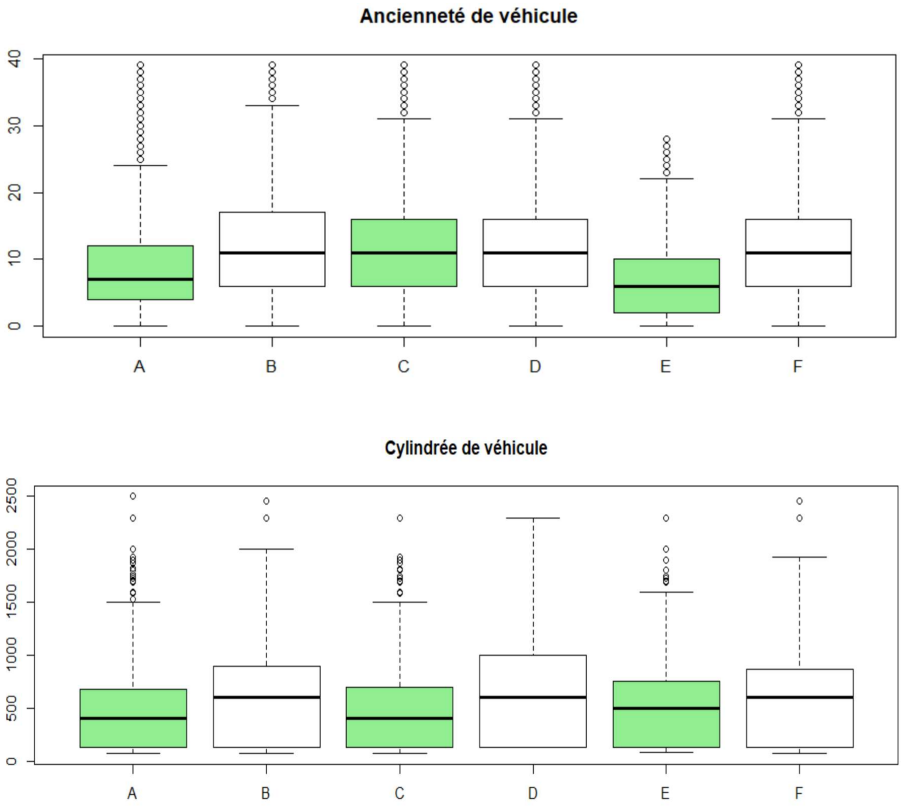


Figure 3 - Statistiques descriptives des variables explicatives du véhicule

La figure 3 compare les données d'ancienneté et de cylindrée par portefeuille. Notons un poids plus élevé des faibles cylindrées dans les portefeuilles A et C. Notons également qu'il y a assez peu de véhicules à très forte cylindrée. La modélisation sur ces segments sera donc moins fiable et l'effet des variables nécessitera un lissage. Ce point sera à prendre en compte dans l'analyse du modèle et son ajustement.

3. Identification des segments (ACM)

3.1. Méthodologie

L'identification des segments est réalisée sur la base des données séparée en deux parties – d'apprentissage et de validation, selon les modalités décrites dans le chapitre 3.

L'analyse par correspondances multiples (ACM) est une méthode statistique qui est généralement applicable à de vastes tableaux avec en ligne les individus ou les observations et en colonnes des modalités des variables nominales et descriptives de ces individus. Ces diverses modalités s'excluent mutuellement et une modalité est obligatoirement présente.

Une variable continue, comme l'âge ou l'ancienneté de permis est transformée en variable nominale par le découpage en classes des valeurs.

Le principe de la méthode ACM consiste en une analyse d'un tableau disjonctif obtenu en transformant les données en profils lignes (individus) et profils colonnes (modalités), avec le code 1 pour la modalité observée et 0 pour toute autre modalité. Chaque modalité individuelle correspond à une colonne (voir Figure 4).

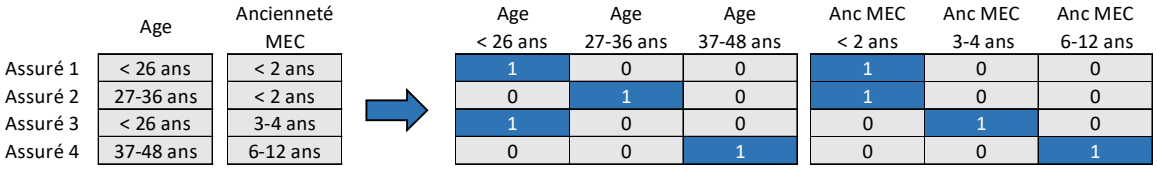


Figure 4 - Transformation des données ACM – tableau disjonctif

Ainsi, il est possible de déterminer des proximités entre individus pour en tirer des enseignements. Par exemple, un profil particulier de l'assuré peut émerger à partir de divers critères observés.

La procédure est sensiblement la même que pour une analyse factorielle des correspondances (AFC) dont le principe consiste à comparer le tableau des observations avec un tableau théorique de totale indépendance. Les écarts entre les deux sont mesurés avec la distance du χ^2. Une distinction entre l'AFC et l'ACM est que l'AFC se fonde sur des fréquences marginales alors que l'ACM est réalisée à partir d'un tableau disjonctif [21].

La formule de la distance du χ^2 est ainsi adaptée à la distance entre deux individus. La somme des distances individuelles obtenue est rapportée au nombre des variables :

$$d^2(i,i') = \frac{1}{v}\sum_{\mu=1}^{p}\frac{(x_i - x_{i'})^2}{\frac{m}{n}}$$

Avec :

v – le nombre des variables,

n – l'effectif,

μ - une modalité et m – son poids.

La somme des valeurs d'un tableau disjonctif est égale à nv.

Pour chaque modalité, on observe une distance entre deux individus (soit 1, soit 0 puisque le tableau ne comporte que des 0 et des 1). Cet écart est rapporté au poids de cette modalité par rapport à l'effectif. Donc, si le poids est élevé, la distance est faible. Si pour une variable tout le monde bénéficie du « 1 » alors tout le monde est proche. C'est aussi l'une des limites de la méthode car si presque personne ne présente telle modalité, alors ceux qui lui sont rattachés se situent à une distance considérable des autres. Il faut veiller à ce qu'aucune modalité ne soit que très peu représentée car les distorsions que cette rareté entraîne risquent de masquer des distances plus faibles mais plus intéressantes car concernant un effectif plus nombreux.

Il existe donc pour chaque modalité un nuage de points-individus dont le centre de gravité est $\frac{m}{nv}$. Il est aussi possible de représenter un nuage de points-modalités global présenté dans la Figure 5.

L'ACM met en évidence les profils d'individus semblables quant aux attributs qui servent à les décrire : 1) deux individus se ressemblent s'ils possèdent les mêmes modalités 2) la proximité entre modalités des variables différentes s'explique par le fait que ces modalités concernent globalement les mêmes individus ou des individus semblables 3) la proximité entre deux modalités d'une même variable s'interprète en termes de ressemblance entre les groupes d'individus qui les possèdent (par rapport aux autre variables actives de l'analyse).

Les modalités de faible effectif sont éloignées des autres modalités sur le plan factoriel – c'est-à-dire, plus le groupe est excentré, plus son effectif est faible (le cas du profil Expert). Ainsi le nombre de véhicule à cylindrée supérieure à 1000, à droite sur le graphique, sera plus faible que le nombre de véhicules dans la catégorie de cylindrée 125-600, qui se trouve quasi au croisement des axes.

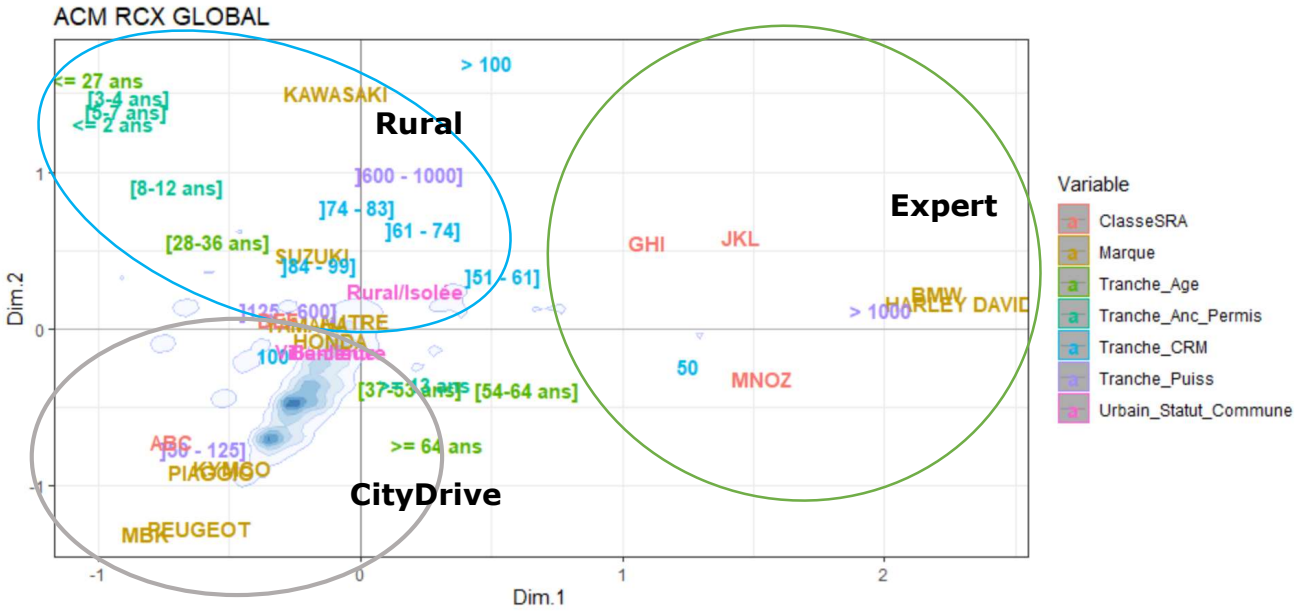


Figure 5 - ACM de population des assurés RCX (Tous portefeuilles)

Une étape importante avant de réaliser une ACM consiste à discrétiser les variables en les découpant en classes homogènes. Le résultat d'analyse est sensible à cette segmentation. Deux questions principales se posent : en combien de classes découper et quelles valeurs regrouper au sein d'une classe. Pour déterminer le nombre de classes, nous avons utilisé une méthode des arbres de régression – Classification ascendante hiérarchique. Il s'agit d'un algorithme CAH qui propose le nombre des classes à retenir en construisant la subdivision qui maximiserait la dispersion entre les classes. Les autres algorithmes de regroupement hiérarchique demandent à l'utilisateur de définir le nombre de classes souhaité. CAH présente toutes les combinaisons possibles et laisse le choix à l'utilisateur. L'inconvénient de CAH est qu'il n'est pas adapté à des volumes de données importants car le temps de calcul augmente sensiblement.

L'exemple d'application de cette méthode pour la détermination du nombre de classes d'âge du conducteur est présenté dans la figure 6 ci-dessous. Selon cet algorithme, le nombre optimal de classes est 5 (les zones entourées en couleur en bas du graphique). Ce nombre est déterminé à partir du diagramme des indices de niveau (en haut à droite de la figure). Nous observons une grosse perte d'inertie jusqu'à 5 classes, l'ajout des classes supplémentaires n'ayant qu'une perte d'inertie marginale.

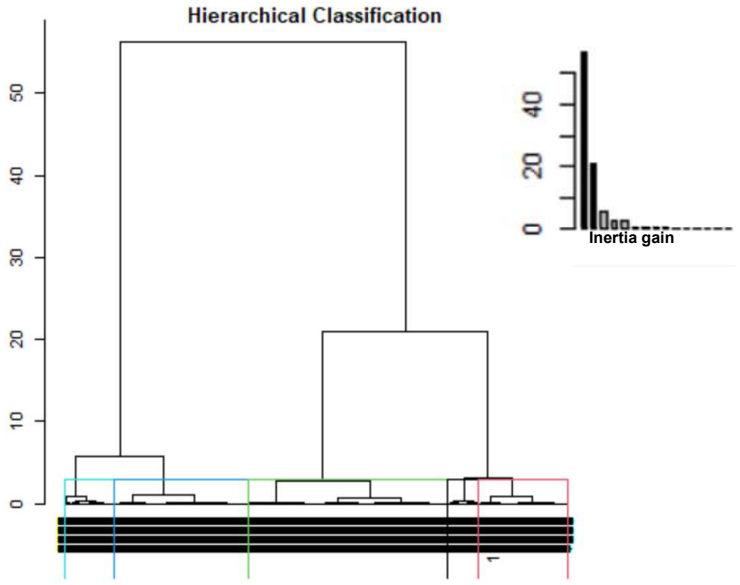


Figure 6 - Classification Ascendante Hiérarchique de l'âge du conducteur

Nous retenons les classes suivantes :

1) Moins de 27 ans
2) 28 – 36 ans
3) 37-53 ans
4) 54-64 ans
5) Plus de 64 ans

Les variables ancienneté de permis, CRM et cylindrée ont été discrétisées de la même manière, les classes suivantes ont été retenues :

Ancienneté de permis - Moins de 2 ans, 3-4 ans, 5-7 ans, 8-12 ans, supérieure à 12 ans ;

CRM – 50, 51-61, 62-74, 75-83, 84-99, 100, supérieur à 100 ;

Cylindrée – 51-125, 126-600, 601-1000, supérieure à 1000.

Ces discrétisations sont cohérentes par rapport à la volumétrie des observations au sein de chaque classe et aux discrétisations appliquées traditionnellement par le métier.

3.2. Application sur les portefeuilles étudiés

En réalisant une ACM sur l'ensemble des portefeuilles (voir Figure 5), 4 profils-types des conducteurs se dessinent :

- **CityDrive** : Le conducteur circulant en centre-ville sur des véhicules de puissance faible (51-125) et ayant un CRM de plus de 84,
- **Rural** : Le conducteur circulant dans le milieu rural sur un véhicule de cylindrée de 125 à 1000, hors expert (CRM50),
- **Expert** : Le conducteur âgé de 54 ans et plus circulant en véhicule de cylindrée et prix forts et ayant un CRM de 50,
- **ProfilAutre** (n'apparait pas sur la Figure 4) : le conducteur avec des métriques non rentrées dans les trois profils précédents.

Il est à noter que l'ACM présentée a été réalisée sur la base des données contrats, c'est-à-dire sans distinction si l'assuré a eu un sinistre ou pas. Les profils obtenus en faisant une ACM sur la population des assurés ayant eu au moins un sinistre n'auraient pas d'application lors de la modélisation car c'est bien tous les profils, sinistrés et non sinistrés, que nous cherchons à tarifer. Une telle analyse peut cependant comporter des informations intéressantes. Les ACM obtenus sur les données sinistrées sont présentés dans l'Annexe 3. Il est à noter que l'utilisation uniquement des données des assurés sinistrées réduit significativement la taille de l'échantillon et par conséquent la fiabilité de l'analyse. Nous ne poursuivrons donc pas cette voie.

3.3. Etude de rentabilité segmentée

Afin d'obtenir une première validation intuitive du modèle, les profils obtenus ont été comparés en termes de fréquence et coût écrêté à 30K€ et non revalorisé.

Le Tableau 9 met en évidence, comme attendu, une fréquence des sinistres moins élevée parmi les conducteurs expérimentés et dans le milieu rural et une fréquence plus élevée en centre-ville. Quant au coût moyen, il est au contraire moins élevé en centre-ville, où la vitesse est limitée, et beaucoup plus fort au milieu rural.

Profil	Poids dans le portefeuille	Coût moyen, €	Coût moyen écrêté, €	Fréquence	S/P	S/P écrêté
Expert	5%	3 642	3 525	0,8%	21%	20%
CityDrive	12%	2 662	2 373	2,2%	41%	37%
Rural	10%	18 973	4 453	0,6%	72%	17%
Autre	73%	8 349	3 028	1,6%	77%	28%

Tableau 9 - Fréquence et coût moyen des profils identifiés

Cependant, ces résultats ne démontrent pas s'il y a un gain de prédictivité suffisamment important pour justifier la complexité apportée dans le modèle tarifaire.

Nous rencontrons également un problème de représentativité des données - le poids des profils obtenus s'avère trop faible pour l'analyse. Les trois profils sont représentés par 27% de la population analysée.

Nous avons donc procédé au relâchement des critères comme suit :

- CityDrive : Le conducteur circulant en centre-ville sur des véhicules de puissance faible (51-125), hors la catégorie Expert ;
- Rural : Le conducteur circulant dans le milieu rural, hors catégorie Expert ;
- Expert : les conducteurs ayant un CRM inférieur à 61 ;
- ProfilAutre : le conducteur avec des métriques non rentrées dans les trois profils précédents.

Profil	Poids dans le portefeuille	Coût moyen, €	Coût moyen écrêté, €	Fréquence	S/P	S/P écrêté
Expert	18%	1 826	1 632	0,8%	19%	17%
CityDrive	13%	1 520	1 360	0,6%	41%	37%
Rural	15%	7 349	1 948	1,6%	55%	14%
Autre	54%	6 048	1 829	2,2%	96%	29%

Tableau 10 - Fréquence et coût moyen des profils (critères assouplis)

Les profils représentés (Tableau 10) correspondent désormais à 39 % de population (en nombre de lignes), mais l'affectation au profil s'effectue sur la base d'une seule variable pour les profils Expert et Rural, ce qui ne constitue pas de changement par rapport à la modélisation classique GLM. Seul le profil CityDrive présente un croisement des deux variables.

Nous conservons cependant les profils définis et analysons dans les parties suivantes l'apport de cette variable à la prédiction.

3.4. Implication sur la construction du tarif

À la suite de la définition des profils, plusieurs hypothèses se posent :

- La modélisation par profil type fera ressortir des variables explicatives différentes pour chacun des profils par rapport à la modélisation au global
- La modélisation par profil type sera plus performante que la modélisation au global

Si ces hypothèses se confirment, il serait pertinent de gérer la tarification de manière distincte par segment, aussi bien pour les affaires nouvelles que dans le cadre de des travaux de revalorisations tarifaires. Les S/P cibles pourront également se définir par profil.

Néanmoins, il convient d'évoquer tout de suite les limitations d'une telle étude - en splittant notre base d'étude en 4 sous-ensembles, nous divisons notre l'échantillon global des données par 4, car uniquement les données relatives aux profils analysés seront utilisées pour la régression. Il en résulte une perte de fiabilité et une plus forte dépendance des résultats sur les éventuels écarts de notre échantillon de la population globale que nous essayons d'estimer.

Une autre limitation consiste dans le fait que l'ensemble des critères tarifants « standard » n'est pas disponible dans notre base de données. Dans les critères standard sont entendus :

- Sinistralité
- Profil de l'assuré (date du permis, coefficient bonus-malus, date de naissance)
- Données du deux roue (la puissance, valeur, date de premier immatriculation)
- Usage
- Niveau de couverture

Notre liste des variables est limitée. Sont absentes les données d'usage (type de trajet, zone, mode de stationnement), de valeur de véhicule, de la sinistralité passée. L'apport de ces variables pourrait changer les résultats.

3.5. Conclusion

Dans cette section, nous avons présenté les données et identifié les profils-type des assurés qui présentent les caractéristiques distinctes pour le risque responsabilité civile.

Pour définir les profils, une base de données a été construite et cette même base de données sera utilisée pour tester le modèle de tarification en segmentée et au global. Les 80% des observations de cette base sont utilisés pour la modélisation et la définition des profils, les 20% restants pour validation des résultats.

En l'absence des données régulières sur certaines variables des portefeuilles (âge, CRM, positionnement géographique), une catégorie « Non communiqué » est créée et nous permettra d'analyser le comportement de ces données et leur impact sur la modélisation.

Pour définir les profils-type, les données ont été discrétisées et analysées en appliquant la méthode statistique d'analyse des correspondances multiples (ACM). La spécificité de la méthode d'ACM, comme de toute méthode non supervisée, est qu'elle dépend largement de l'interprétation. Le choix de retenir 4 profils-types est motivé par la simplification et l'interprétabilité de ces profils. En contrepartie, le fait de limiter le nombre de profils permet de conserver un nombre d'observations suffisamment important par profil. Chaque ajout d'un profil supplémentaire réduit le nombre d'observations dans chacune des bases individuelles car toute observation peut être affectée à un seul profil.

Les résultats qui seront obtenus sur la base des données compilées peuvent également présenter de nombreux biais inhérents à l'analyse actuarielle sur les bases de données de tailles limitées. Il reste donc possible que les résultats soient affectés par le mix particulier des portefeuilles individuels, le poids et le nombre des sinistres graves et/ou corporels, la concentration géographique (la densité n'est pas la même entre Paris et Chamonix, qui sont toutes les deux dans la catégorie ville-centre)…

DEUXIEME PARTIE :

GLM EN TARIFICATION MOTO

Dans ce chapitre, nous poserons d'abord les bases théoriques essentielles de la tarification en assurance. Ensuite, nous réaliserons plusieurs modélisations séparées afin de projeter le nombre de sinistres, le coût du sinistre et la prime pure directement, en appliquant les modèles linéaires généralisés. Ces modélisations seront décrites sur l'exemple de la garantie responsabilité civile. Nous utiliserons cette garantie car, en tant que garantie obligatoire, elle est acquise par l'ensemble des assurés des portefeuilles étudiés. Cela permet d'avoir le maximum d'observations possible pour confirmer ou infirmer l'intérêt de notre analyse par profil-type. Nous ne nous attarderons pas à réaliser les mêmes étapes pour les autres garanties à ce stade. Si l'apport d'analyse par profil-type à la prédiction est confirmé sur la garantie responsabilité civile, une analyse similaire sera à réaliser pour les autres garanties. En cas de résultats positifs, les autres garanties seront modélisées selon les mêmes principes pour définir le tarif du produit final.

Devenus classiques dans les compagnies d'assurance, les modèles linéaires généralisés sont intégrés dans de nombreux logiciels statistiques, comme SAS, et sont réalisables en différents langages de programmation (dont R et Python sont les plus connus). Les modèles qui seront présentés dans ce chapitre sont élaborés sur le logiciel Akur8.

En particulier, le modèle obtenu en utilisant la segmentation des profils proposée dans le chapitre précédent sera comparé au modèle sur la base des variables initiales sans regroupement.

Pour chacun des tests, les bases sont divisées en deux, selon la méthodologie qui sera décrite plus loin : une base d'apprentissage de 80% des données et une base de test avec les 20% restants.

1. Modélisation du risque en assurance moto

Le tarif de l'assurance est composé de la prime pure, ou prime de risque, du chargement de gestion et de la marge attendue par l'assureur. La prime de risque inclut en général un chargement de sécurité qui est plus ou moins important en fonction de la volatilité du risque. Dans l'approche collective, les contrats ne sont pas distingués et les

charges individuelles ne sont pas forcément connues. La prime pure est le résultat du produit de l'espérance du nombre moyen des sinistres et de leur coût moyen.

L'application de la seule approche collective, dans un contexte concurrentiel, confronte l'assureur au risque de l'antisélection. En effet, l'application d'un tarif uniforme aux bons et mauvais risques ferait fuir les bons assurés qui se verront proposer un meilleur tarif ailleurs, alors que les mauvais assurés seraient intéressés de rester, car ils obtiennent un tarif plus bas qu'ailleurs.

A l'opposé, il est possible d'individualiser la prime en se fondant sur l'historique de chaque contrat. L'inconvénient est que le risque n'est plus mutualisé, alors même que la sinistralité passée n'est pas toujours représentative de la sinistralité future.

L'introduction de la segmentation vise à pallier ce problème, en sous-divisant le risque en classes plus ou moins fines sur la base des informations connues a priori, comme l'âge ou l'ancienneté de permis du conducteur.

L'assureur peut également partiellement utiliser l'historique de chaque assuré pour calculer une prime qui correspond à son risque. Il s'agit de la tarification a posteriori : le tarif initial de l'assuré est adapté, au cours de la vie de son contrat, à sa sinistralité individuelle. C'est l'objet de la théorie de crédibilité, dont le modèle de crédibilité linéaire est largement utilisé dans l'assurance (Bühlmann, 1967).

La théorie de crédibilité retrouve son application dans le coefficient CRM créé par l'Etat (Code d'assurance : Articles A121-1 à A121-2) pour récompenser les bons conducteurs ou, inversement, pour punir les conducteurs les plus malheureux. Il est compris entre 0,50 et 3,50 : plus le coefficient est petit, plus le bonus est grand et meilleur est le tarif ; plus le coefficient est élevé, plus le malus et le tarif d'assurance sont élevés.

2. Tarification selon le modèle linéaire généralisé

2.1. Notions théoriques du modèle linéaire généralisé

Les modèles linéaires généralisés ont été introduits par Nelder and Wedderburn (1972) [14] et, depuis, leur application a été suffisamment étudiée par la communauté actuarielle. Pour réaliser la modélisation GLM, une solution s'appuyant sur l'outil Akur8

est utilisée. Cette solution permet une visualisation des données et simplifie l'application du GLM tout en laissant le choix à l'actuaire quant au paramétrage à utiliser. Le fait d'automatiser la modélisation permettra de focaliser la recherche sur le côté pratique d'optimisation de tarifs. L'outil Akur8 permet également de comparer les modèles entre eux, un avantage non négligeable lors de l'analyse et le choix des paramètres à retenir, comme on pourra le constater ci-après.

Le modèle GLM permet de relier des variables explicatives à une variable à expliquer au travers d'une fonction appelée « fonction lien » qui linéarise la relation entre ces deux types de variables.

Ce modèle permet de dépasser les limites du modèle linéaire dues à ses hypothèses non réalistes pour l'assurance : normalité de la variable expliquée, constance de sa variance (homoscédasticité), additivité. En assurance non-vie nous nous intéressons à des valeurs non négatives, car la fréquence des sinistres et le coût sont des valeurs nulles ou positives. La distribution de la fréquence et du coût moyen n'a donc pas la forme en cloche et symétrique de la distribution normale et par conséquent, l'hypothèse de normalité ne peut donc pas être maintenue. Également, la variance de la variable de fréquence et/ou de coût au sein des différents segments de la clientèle ne sera pas la même.

Le modèle GLM se compose de trois éléments [7] :

- Une variable à prédire (Y) suivant une certaine loi de distribution de la famille des exponentielles. Chaque composant de y est indépendant et suit la forme de l'une des distributions de la famille exponentielle ; la même forme pour chaque composant, mais avec des paramètres différents
- Des facteurs prédictifs qui se combinent pour produire une prédiction linéaire $\eta = X.\beta$
- Une fonction de lien servant à linéariser le lien entre la variable de réponse (Y) et les variables explicatives (X). Cette fonction est du type $E(Y) = \mu = g^{-1}(\eta)$. Il est à noter que pour les variables explicatives il s'agit des corrélations et non des liens de causalité. Ce n'est pas parce qu'un conducteur est jeune qu'il représente un plus grand risque, mais parce qu'il a potentiellement moins d'expérience. Son risque est donc corrélé à son âge, sans que l'âge en soit la cause.

Les deux fonctions de lien les plus souvent utilisées en tarification sont les suivantes :

Lois	Lien	Fonction	Application
Poisson	Log	$g(\mu)= \log(\mu)$	Nombre de sinistres, fréquence de sinistres
Gamma	Réciproque	$g(\mu)=-1/\mu$	Coût moyen des sinistres

Tableau 11 - Fonctions de lien usuelles

Voici un exemple d'application via une transformation du lien du type ln :

$$g(x)= \log(x) \text{ alors } g^{-1}(x) = e^{x}$$

$$\text{Si } \ln[E(Y)] = \beta X \text{ et } E(Y) = \mu, \text{ alors } \mu_i = g^{-1}(\beta_i x_{i1}+\ldots.+ \beta_i x_{ip}) = e^{\beta i x i1} \ldots e^{\beta i x ip}$$

Avec :

x_1 *à* x_p – les facteurs explicatifs (ex. âge, voiture, localisation)

β - les paramètres estimés par maximum de vraisemblance.

Le modèle n'est donc plus additif, comme le cas du modèle linéaire, mais multiplicatif.

Enfin, pour valider le tarif GLM, les paramètres du modèle de la prime pure obtenus sont testés et analysés à travers les intervalles de confiance.

La prime pure est obtenue en multipliant une prime pure de base par le coefficient représentant la modalité de l'individu pour chaque variable tarifaire. Pour chaque variable tarifaire, le coefficient correspondant à la modalité de référence est 1, et pour chaque modalité différente, le coefficient diffère afin de représenter l'impact sur le risque :

Prime pure de base	50

Age	Coefficient
25	1,5
30	1,45
35	1,3
40	1

Ancienneté du véhicule	Coefficient
0	1,3
2	1,1
4	1
6	0,8

Tableau 12 - Calcul de la prime pure en utilisant des coefficients

Ainsi une prime pure d'un conducteur de 25 ans avec un véhicule de 4 ans est égale à 50x1,5x1 = 75 et une prime d'un conducteur de 30 ans avec le même véhicule est égal à 50x1,45x1= 65.

L'hypothèse sous-jacente du modèle multiplicatif est que le changement d'un niveau sur une variable est indépendant du niveau des autres variables, ce qui n'est pas toujours vérifié. Cette hypothèse créée des distorsions dans l'évaluation de la prime pure. Les distorsions seront plus fortes lorsque le risque s'éloigne du risque de référence.

Un moyen de limiter les distorsions consiste à estimer une prime pure de chaque croisement des variables explicatives, mais dans la pratique il est quasi impossible d'avoir une base de données assez large pour que chaque combinaison des variables soit suffisamment représentée pour que la loi des grands nombres puisse s'appliquer et que le calibrage du modèle soit fiable. L'utilisation des profils-types définis dans le premier chapitre constitue une tentative de réduire ces distorsions.

Dans la pratique, la démarche pour la modélisation GLM est donc la suivante [10] :

1. Analyse descriptive, visualisation des données (étape réalisée dans le premier chapitre)

2. Choix du modèle et du fonction du lien (gamma, poisson etc.) : transformation de variables, variables à introduire dans le modèle

3. Estimation des paramètres du modèle, tests, intervalles de confiance

4. Sélection des variables explicatives, choix de leur paramétrisation (continues/discrétisées)

5. Interprétation du modèle

6. Validation et comparaisons de modèle(s)

Akur8 permet d'automatiser certaines tâches des étapes 3 à 6 tout en gardant la main sur les choix des variables et les arbitrages souhaités entre les modèles les plus pertinents parmi les modèles acceptables.

2.2. Utilisation de la validation croisée

Les critères statistiques classiques pour la validation du modèle sont les critères AIC et BIC.

Le critère d'information d'Akaïke (AIC) est un critère défini par la formule :

$$AIC= -2log\ L+2k$$

Où *L* représente la vraisemblance maximisée et k le nombre de paramètres dans le modèle. Ce critère vise une recherche de compromis entre la parcimonie dans le nombre de variables et la minimisation du biais. Le meilleur modèle est celui ayant un AIC le plus faible.

Il est à noter que la plupart des propriétés avantageuses des estimés du maximum de vraisemblance, dont l'absence de biais, sont valides dans la limite où la taille de l'échantillon est suffisamment grande. Néanmoins ce critère de taille de l'échantillon n'est pas absolu et dépend du modèle et en particulier du nombre de paramètres à estimer.

En pratique, le maximum de vraisemblance est obtenu par un algorithme numérique recherchant le maximum par un processus itératif. Une fonction de vraisemblance complexe pourrait avoir plusieurs maximums locaux (des points où la fonction est maximisée par rapport aux valeurs proches des paramètres), dans lequel cas il n'est pas garanti que l'algorithme trouve le maximum global (celui avec la vraisemblance la plus élevée).

Le Bayesian Information Criterion (BIC) est plus parcimonieux que le critère AIC puisqu'il pénalise le nombre de variables présentes dans le modèle.

Avec le développement de la science de données, ces critères cèdent leur place aux techniques de validation croisée, qui ont un spectre d'application plus large. C'est cette dernière qui sera utilisée. Issue d'apprentissage supervisé, la validation croisée vise à réduire le biais introduit par la base de données elle-même. Tout modèle est construit en optimisant les paramètres afin de correspondre le mieux aux données utilisées. En prenant ensuite un échantillon de données indépendant, mais issu de la même population, il peut s'avérer que le modèle ne modélise pas aussi bien le comportement de cet échantillon indépendant – cet effet est appelé surapprentissage. D'où l'importance

d'avoir des mesures permettant de qualifier le comportement du modèle sur les données non utilisées lors de l'apprentissage.

La méthode K-fold, schématisée dans la Figure 7, consiste à séparer les données en deux parties - la première à utiliser pour la construction du modèle (la base d'apprentissage) et la deuxième pour sa validation (base de test). La base d'apprentissage est répartie en *k* groupes, dans l'exemple ci-dessous - 4. Au sein de chaque groupe les trois quarts des données sont utilisées pour l'apprentissage et le quart restant pour le test. La moyenne des performances des *k* tests peut être considérée comme un estimateur robuste.

La valeur standard du paramètre *k* est de 10, cette valeur ayant démontré de fournir un meilleur équilibre entre le temps de calcul et le biais de l'estimateur de la performance du modèle [21]. Cette valeur sera retenue pour nos tests.

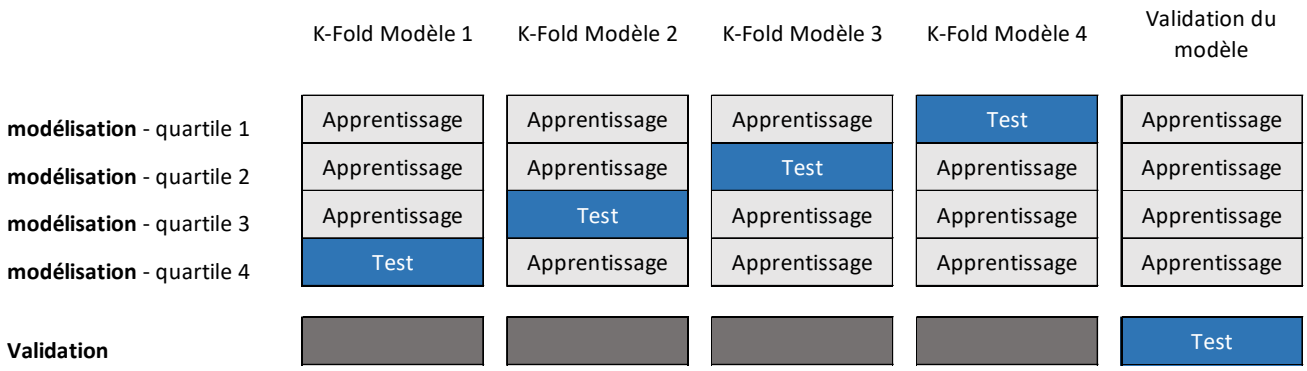


Figure 7 - Méthode de validation croisée K-Fold

Pour appliquer cette méthode, nous affectons à chaque observation de notre base un numéro aléatoire de 1 à 10 qui permettra de distinguer les données en deux sous-ensembles – numéros de 1 à 8 pour le jeu d'entrainement et les numéros 9 et 10 pour le jeu de validation.

2.3. Modélisation de la fréquence des sinistres

Choix du nombre de variables et du modèle :

Pour modélisation de la fréquence, les modèles possibles ont été comparés en utilisant les critères de R^2 ajusté et de coefficient de Gini (voir la section *Critères de validation*

d'un modèle pour l'explication des critères). Chacun des points sur la figure 8 représente un modèle. Il n'y a pas de gain significatif entre les modèles à partir de 8 variables. Nous choisissons donc le modèle parmi ceux à 8 variables dont les coefficients R^2 ajusté et Gini sont les plus importants. La significativité statistique est automatiquement contrôlée par Akur8, l'ensemble des modèles proposés pour considération répond aux critères de validité statistique.

Compte tenu de la faible présence de sinistres corporels dans notre base de données, les sinistres sont modélisés sans distinction entre matériels et corporels.

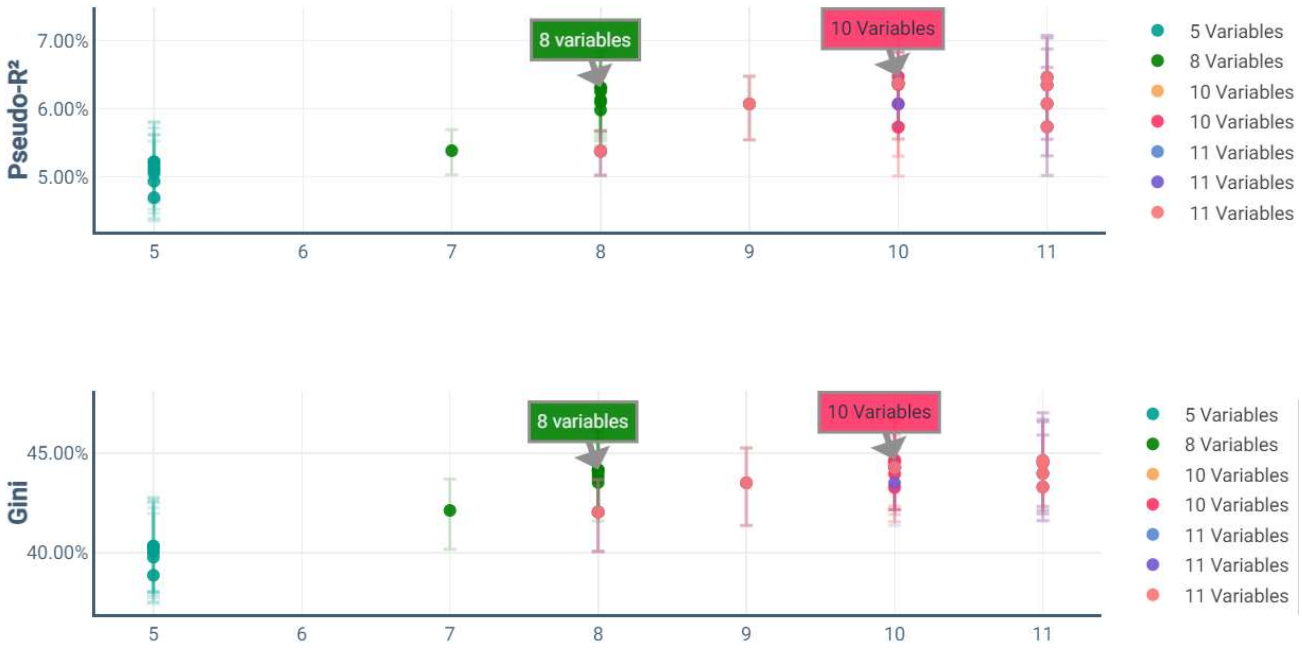


Figure 8 - Performance des modèles de fréquence par nombre des variables

Une fois le nombre de variable défini, nous allons nous attarder sur les variables qui ont été retenues. La figure 9 présente ainsi les variables retenues et leur impact sur la tarification, exprimé en spread de coefficients. L'ancienneté du permis ressort de loin comme le critère le plus diversifiant, suivi du statut urbain de commune. La variable du profil a été retenue mais son impact tarifaire est très léger.

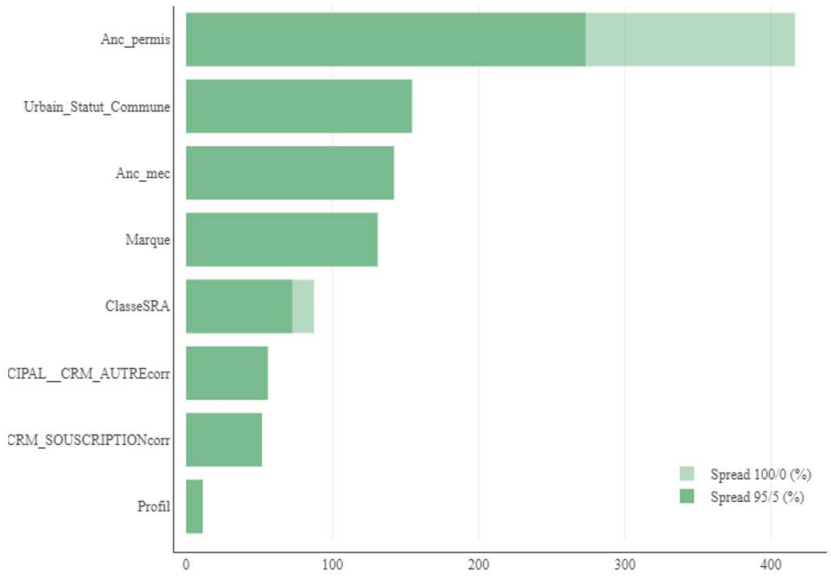


Figure 9 - Impact tarifant des variables retenues pour la fréquence

La métrique du spread est utilisée pour évaluer l'impact tarifaire de chacune des variables. Le spread est calculé selon la formule suivante :

$$Spread = \frac{coefficient\ de\ tarification\ maximal + 1}{coefficient\ de\ tarification\ minimal + 1} - 1$$

Le spread à 95/5 est calculé selon la même formule mais en excluant les coefficients des 5% de l'échantillon aux coefficients plus bas et plus élevés.

L'impact de chaque variable tarifaire sur le tarif peut ensuite être analysé afin de s'assurer de la cohérence de la modélisation. Les coefficients des données observés de la variable ancienneté du permis (ligne violette dans la figure 10) sont très volatiles. Les coefficients retenus pour le modèle (ligne verte) ont été lissés, réduits pour les jeunes conducteurs et maintenus stables pour les conducteurs à partir de 40 ans. Ces modifications visent à pallier le manque de données sur certains segments qui engendrent une volatilité accrue et intégrer un avis métier. Par exemple, nous savons que le nombre d'observations pour les conducteurs ayant un permis de plus de 40 ans est très faible et que les fluctuations observées n'ont pas d'explications en dehors d'un manque de données.

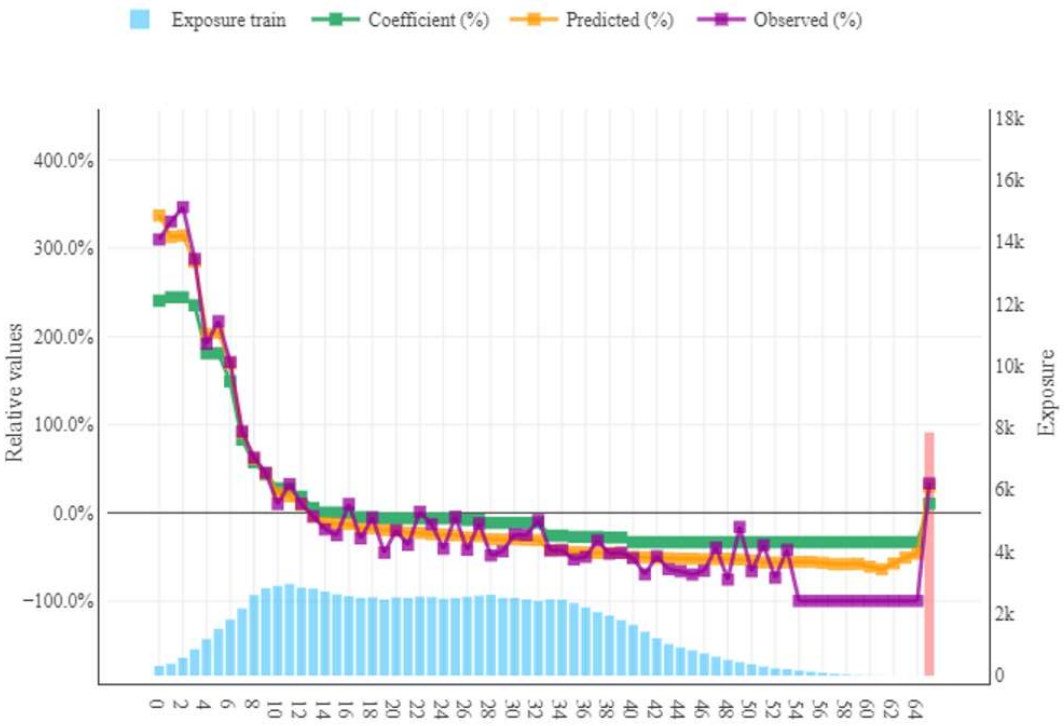


Figure 10 - Analyse du spread du coefficient Ancienneté du permis

La deuxième variable la plus discriminante est le statut urbain de la commune. Comme indiqué dans la première partie de l'étude, la logique actuarielle voudrait que le risque ne soit pas le même entre le centre-ville dans une zone à faible population par rapport au centre-ville d'une grande agglomération par exemple de Paris. Or, les coefficients proposés par le modèle initial (Figure 11) ne tiennent pas compte du positionnement de la commune et pénalisent ainsi les habitants des régions à faible population. La prime d'un habitant dans le centre d'une petite commune du centre de la France, à tout critère équivalent, sera la même que la prime d'un parisien.

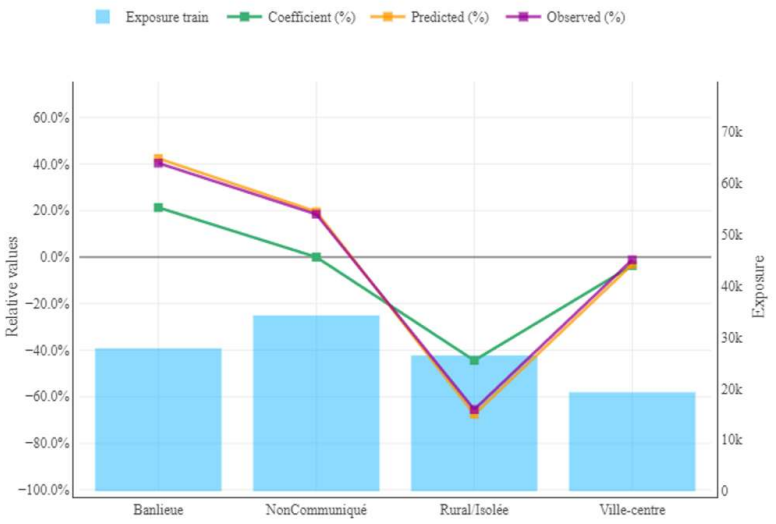


Figure 11 - Spread et exposition de la variable Statut urbain de la commune

Nous avons donc décidé d'enrichir le modèle en utilisant le positionnement GPS de la commune concernée. L'analyse graphique présentée dans la Figure 12 confirme notre intuition en attribuant aux zones de l'Ile-de-France, du sud de la France et de la Corse des coefficients majorants.

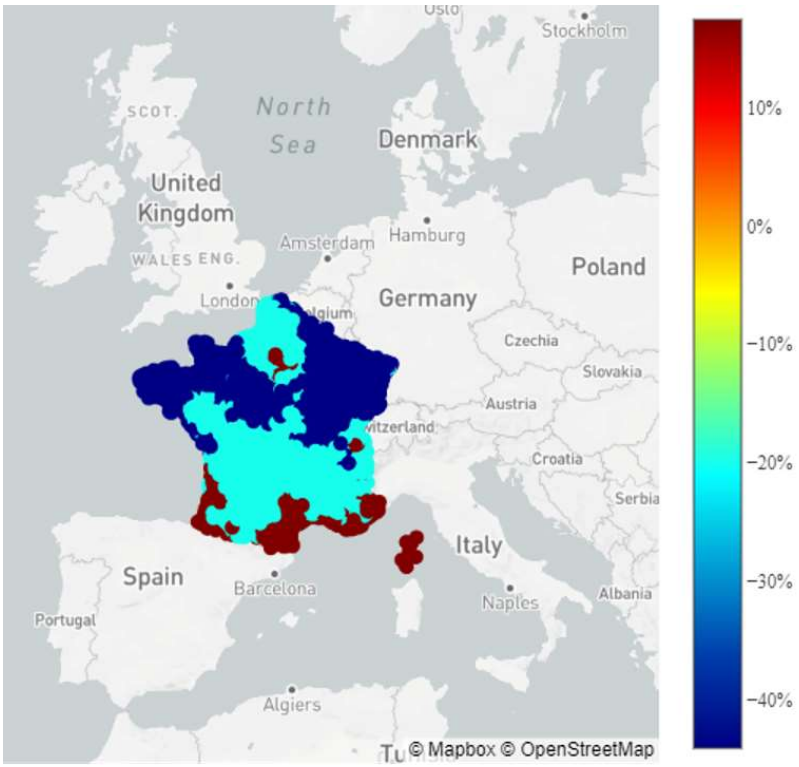


Figure 12 - Spreads des coefficients de tarification (Positionnement GPS)

Quant aux profils-type proposés dans le premier chapitre, leur impact tarifant, présenté dans la figure 13, ne semble pas suffisamment important pour justifier la complexité d'analyse diversifiée. Nous ne poursuivrons donc pas cette piste pour la modélisation de la fréquence des sinistres. Cette variable est fortement corrélée aux autres variables du modèle, ce qui crée du bruit dans la modélisation. Le remplacement des variables qu'elle résume ne semble pas avoir un effet escompté.

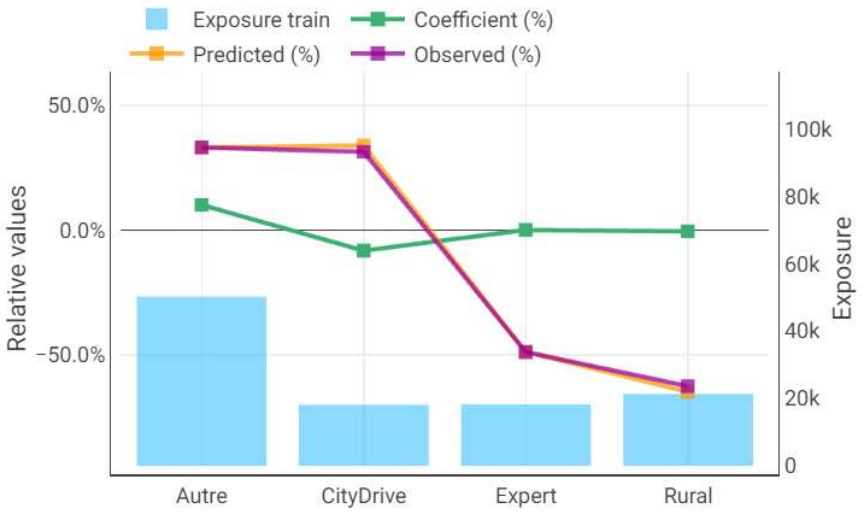


Figure 13 - Impact tarifant du Profil-type du conducteur

2.4. Modélisation de coût moyen

2.4.1. Seuil de la sinistralité attritionnelle

La modélisation du coût moyen GLM est réalisée sur les sinistres dits « attritionnels », autrement dit, normaux et récurrents. Cette nomination est donnée pour les distinguer des sinistres graves, qui ne pourront pas être modélisés aussi facilement, du fait de leur nature plus rare et donc leur faible présence dans la base de données. Il est donc communément accepté, lorsque l'assureur ne dispose pas d'études plus fiables, de repartir la charge des sinistres graves sur l'ensemble des assurés, sur la base des prédictions du ratio de S/P budgété dans le business plan.

Il convient d'abord de définir le seuil jusqu'auquel la sinistralité est considérée comme attritionnelle. Il n'existe pas d'approche unique à ce sujet, les approches existantes

démontrent plusieurs seuils possibles. Notre objectif est donc de nous assurer que le seuil retenu par l'assureur reste cohérent par rapport à nos données et ne pénalise pas la modélisation.

C. SCARROTT et A. MACDONALD (2012) ont dressé un panorama des outils pour détermination du seuil des valeurs extrêmes dans une population. Ce seuil est celui à partir duquel la queue de la distribution peut être approximée par un modèle de valeurs extrêmes (dans la plupart des cas par la loi de Pareto Généralisé). La théorie des valeurs extrêmes s'intéresse à ce seuil pour étudier la queue de la distribution mais ses méthodes sont applicables pour définir le seuil de la sinistralité attritionnelle, car ce seront les valeurs qui se retrouveront en dessous de ce seuil.

Pour les populations à queues de distribution lourdes, c'est-à-dire avec une probabilité plus importante que celle de la loi normale d'avoir des sinistres extrêmes, l'approche graphique de Hill est utilisée. Avant de pouvoir l'appliquer, il convient de vérifier si le comportement de la distribution de la charge des sinistres supérieurs à un seuil ressemble graphiquement à une distribution à queue lourde.

Le diagramme Quantile-Quantile est un outil graphique permettant d'évaluer la pertinence de l'ajustement d'une distribution donnée à un modèle théorique. La Figure 14 présente le diagramme Quantile-quantile du type Pareto appliquée sur les données des sinistres supérieurs à 1000 euros. Plusieurs seuils ont été testés pour choisir un seuil à partir duquel la distribution prend une forme de loi Pareto. La forme de droite du diagramme démontre que la distribution peut être approximée par cette loi. Le seuil se trouve donc forcément au-dessus de 1000 euros.

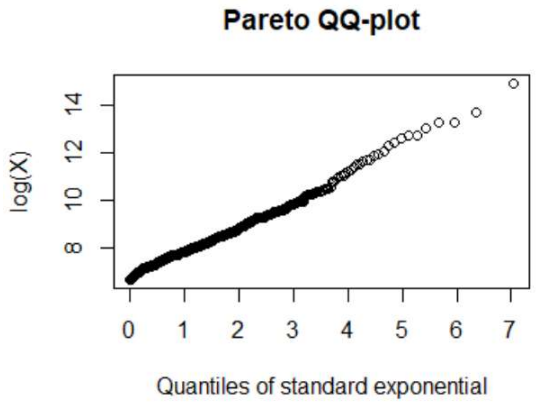


Figure 14 - Diagramme Quantile-quantile Pareto (sinistres au dessus de 1K€)

Le graphique de Hill présenté dans la Figure 15 permet d'identifier les zones possibles pour le seuil. Pour réaliser ce graphique les montants des sinistres doivent être triés du plus grand au plus petit. Le seuil possible se situe sur la plus grande valeur possible de la zone de graphique où le paramètre α se stabilise. La première zone stable se trouve vers α=0,9 et la statistique d'ordre 40-50 (difficilement lisible entre la statistique d'ordre 15 et 268 indiqués sur la Figure 15). Cette zone stable correspond à un seuil entre 20 000 et 30 000 maximum.

Figure 15 - Graphique de Hill de la base des sinistres

Pour valider les analyses graphiques, les sinistres ont été classés par tranche de charges (Tableau 13). La zone optimale qui permettra de capter les sinistres de fréquence, à partir de 95% en nombre, se trouve sur les tranches de 10 000 à 30 000.

Tranche de charge, K€	Nombre de dossiers	Coût moyen, K€	% Cumul en charge	% Cumul en nombre
0-1	634	0,5	2%	38%
1-1,5	297	1,3	5%	56%
1,5-3	368	2,1	11%	78%
3-5	145	3,8	16%	87%
5-10	100	6,8	21%	93%
10-20	61	13,5	27%	97%
20-30	15	24,7	30%	98%
30-60	15	39,2	34%	99%
60-200	14	105,7	46%	99%
> 200	10	710,7	100%	100%
Total général	**1659**	**7,9**	**100%**	**100%**

Tableau 13 - Répartition des sinistres par tranche de charge

Il est à noter que, pour des raisons pratiques, un assureur choisit un seul seuil d'écrêtage pour l'ensemble de ses garanties. Ainsi, nous conservons un seuil d'écrêtage à 30 000€ pour le risque RC moto. Les analyses qui viennent d'être présentées démontrent que ce seuil reste acceptable, même s'il se situe en haut de la fourchette.

2.4.2. Revalorisation des coûts matériels

L'historique des bases de données sinistres remonte à 2012, il est nécessaire de revaloriser les sinistres anciens afin de ne pas fausser la projection des coûts à aujourd'hui. La garantie responsabilité civile couvre les sinistres matériels et les sinistres corporels et la structure des coûts de ces deux types des sinistres n'est pas la même. Pour les sinistres matériels il s'agira principalement de coût des pièces détachées et de la main d'œuvre, alors que pour les sinistres corporels il s'agira des frais médicaux, de perte de gains professionnels et d'assistance par une tierce personne et même éventuellement de l'évolution du taux de mortalité.

Pour l'inflation pour la partie dommage matériel, il est possible d'utiliser l'indice INSEE d'entretien et réparation de véhicules particuliers (Identifiant 001764110). L'évolution de cet indice sur les années 2012-2021, présentée dans la figure 16, démontre une hausse supérieure à 20% et nous confirme la nécessité de revalorisation des coûts de sinistres matériels préalable à la modélisation.

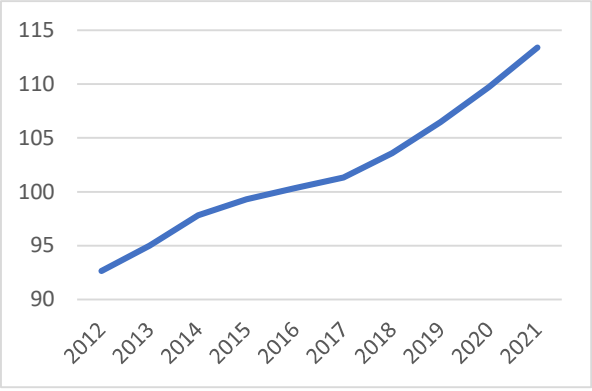


Figure 16 - Indice d'entretien et réparation de véhicules particuliers

La revalorisation est à appliquer uniquement sur les règlements passés, en fonction de la date effective de règlement, mais pas sur la réserve résiduelle, cette dernière tient déjà compte de l'inflation.

Il est à noter que plusieurs conventions entre les assureurs existent sur le marché français dans l'objectif d'accélérer l'indemnisation des victimes et faciliter la gestion des recours. Ces conventions prévoient des montants forfaitaires ou des barèmes d'indemnisations. Notons les deux plus applicables - la convention d'Indemnisation et de Recours Corporel Automobile (IRCA) et la convention d'indemnisation et de recours entre sociétés d'assurance (IRSA). La modélisation des coûts d'un assureur reflètera l'application ou pas de ces éventuelles conventions. Lorsqu'une des compagnies d'assurance impliquées dans le sinistre n'entrent pas en conventions concernées, l'indemnisation se fait en droit commun. Nous nous positionnons dans le cas d'une compagne adhérente à l'IRCA mais pas à l'IRSA. Cette particularité se reflètera dans l'estimation des coûts de sinistres dossier dossier de notre base de données et, par conséquent, sur les résultats de modélisation qui ne seront applicables que pour une compagnie qui appliquera le même schéma de conventions.

2.4.3. Revalorisation des coûts des sinistres corporels

En termes d'évaluation des coûts, les sinistres corporels de la garantie responsabilité civile présentent plusieurs spécificités par rapport aux sinistres matériels :

- La durée de vie du sinistre est plus longue
- Absence de plafond de garantie
- Evaluation aux dires d'expert, bien qu'encadrée par la règlementation.

La difficulté de détermination du taux de revalorisation pour ces sinistres est expliquée par la variété des postes de préjudices indemnisables, établis par un médecin expert.

Le rapport d'un groupe de travail dirigé par M. Jean-Pierre Dintilhac a élaboré une nomenclature commune des préjudices corporels. Les postes indemnisables en sont (DINTILHAC, 2005) :

Préjudices patrimoniaux

a) Préjudices patrimoniaux temporaires (avant consolidation)

- Dépenses de santé actuelles (D.S.A.)
- Frais divers (F.D.)
- Pertes de gains professionnels actuels (P.G.P.A.)

b) Préjudices patrimoniaux permanents (après consolidation)

- Dépenses de santé futures (D.S.F.)
- Frais de logement adapté (F.L.A.)
- Frais de véhicule adapté (F.V.A.)
- Assistance par tierce personne (A.T.P.) assistance dont a besoin la victime dans la vie courante
- Pertes de gains professionnels futurs (P.G.P.F.) perte de revenus liée à l'état de santé de la victime
- Incidence professionnelle (I.P.)
- Préjudice scolaire, universitaire ou de formation (P.S.U.)

Préjudices extrapatrimoniaux

a) Préjudices extrapatrimoniaux temporaires (avant consolidation)

- Déficit fonctionnel temporaire (D.F.T.)
- Souffrances endurées (S.E.), sur une échelle de 1 - très léger, à 7 - très important.
- Préjudice esthétique temporaire (P.E.T.)

b) Préjudices extrapatrimoniaux permanents (après consolidation)

- Déficit fonctionnel permanent (D.F.P.) taux d'atteinte permanente à l'intégrité physique et psychique qui est évalué entre 0 et 100 %.
- Préjudice d'agrément (P.A.) suppression ou diminution définitive ou temporaire des activités de loisir de la victime
- Préjudice esthétique permanent (P.E.P.), estimé sur une échelle de 1 à 7
- Préjudice sexuel (P.S.)
- Préjudice d'établissement (P.E.)
- Préjudices permanents exceptionnels (P.P.E.)

c) Préjudices extrapatrimoniaux évolutifs (hors consolidation)

- Préjudices liés à des pathologies évolutives (P.E.V.)

Une partie seulement des coûts liés aux sinistres corporels, aménagements ou dépenses de santé par exemple, peut être indexée sur l'inflation ou indices spécifiques. Les autres postes, comme A.T.P., P.G.P.F. et D.S.F. peuvent impliquer un provisionnement sous forme de rentes, avec des hypothèses sous-jacentes de taux et de revalorisation spécifiques. Ainsi, certains coûts de sinistres corporels encore ouverts à la date d'analyse se retrouveront revalorisés (à travers le mécanisme de révision des sinistres), d'autres non.

Le barème officiel de l'IRCA fixe les seuils planchers et plafonds d'indemnisation en fonction du taux de déficit fonctionnel permanent :

- D.F.P. de 0% : utilisation d'un forfait de base de 1 480€ (depuis 2018),
- D.F.P. compris entre 0% et 5% : application du barème officiel IRCA,
- D.F.P. supérieur à 5% : évaluation au coût réel, uniquement si la compagnie adverse ne dénie pas sa garantie.

En présence d'une base de sinistres suffisamment large et fiable, il aurait été possible de calculer des triangles de coût moyens attritionnels corporels et constater l'évolution d'une année sur l'autre. Dans notre cas, nous nous contentons d'utiliser un taux 3%, issu d'un benchmark auprès des autres assureurs, sur les produits similaires, obtenu auprès de l'équipe de tarification. En France il est d'usage de considérer que la partie matérielle représente 66% de la prime pure.

2.4.4. Variables significatives du coût moyen

Avant de réaliser la modélisation, nous avons éliminé les sinistres avec une charge négative pour le calcul de la charge moyenne (167 sinistres). Une charge négative apparaît lorsque la somme de recours obtenus et/ou à obtenir pour un sinistre donné est supérieure à l'indemnisation versée et/ou provisionnée pour ce même sinistre. Les causes d'apparition d'une charge négative peuvent être multiples – par exemple, un reliquat de recours obtenu en attente de reversement à un assuré ou une saisie erronée d'un montant de recours ou de l'indemnisation.

La base de données sinistres ne contient pas l'immatriculation du véhicule. Ainsi, lorsqu'un assuré possède deux véhicules, nous faisons l'hypothèse que le sinistre se rapporte au premier véhicule.

Pour la modélisation du coût moyen, nous avons utilisé la loi Gamma, couramment utilisée par le métier dans cet objectif. La loi log-normale a également été testée mais aucun modèle n'a abouti à un coefficient de la prédictivité R^2 supérieur à 0%.

L'application du GLM Gamma s'est relevée moins impactante pour le calcul du coût moyen que pour le calcul de la fréquence. L'apport de l'information des 4 variables à la prédictivité du coût moyen n'explique pas plus de 3,5% de variance. La figure 17 démontre que l'ajout des variables supplémentaires a même un impact négatif sur la

prédictivité du modèle (R^2 négatifs pour certains modèles à 10 variables). Nous retenons donc le modèle à 4 variables pour analyse plus détaillé.

Les variables retenues et leur impact tarifant sont présentés dans la figure 18. L'impact de la cylindrée ressort en plus important, avec le spread des coefficients jusqu'à 65%. Notons qu'aucune variable en lien avec la géographie n'a été retenue (statut commune, profil-type..). L'ajout de la variable code INSEE engendre une baisse de prédictivité du modèle, exprimé en R^2 ajusté, de 3,29% à 3,07%.

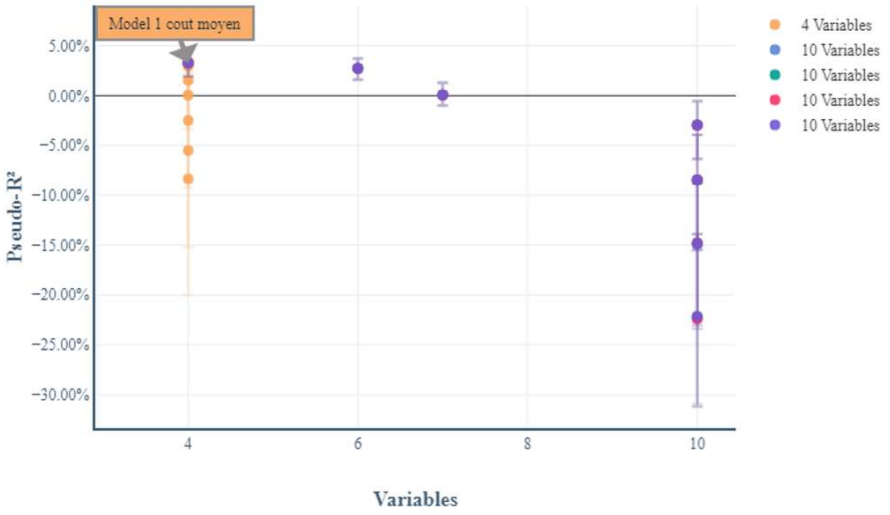


Figure 17 - Performance prédictive des modèles coût moyen Gamma

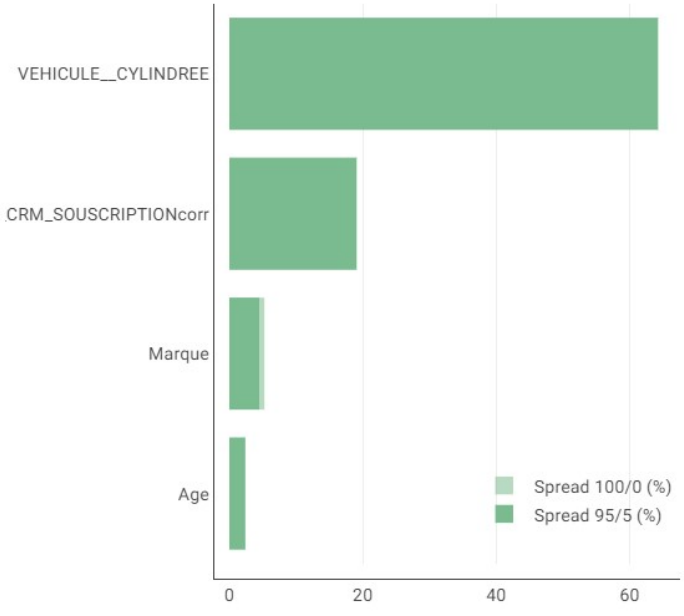


Figure 18 - Impact tarifant des variables retenues pour le coût moyen

A cause de la faible volumétrie des sinistres de la base, l'exposition à différentes cylindrées n'est pas uniforme, ce qui explique une forte volatilité des coefficients observés (Figure 19). Néanmoins, nous avons obtenu des résultats cohérents avec des coefficients qui augmentent avec la hausse de la cylindrée.

Il est à noter que pour la même raison (faible volumétrie des sinistres dans la base), nous avons pris la décision de ne pas distinguer les sinistres corporels des sinistres matériels dans le modèle proposé. La volumétrie des sinistres corporels étant encore moindre.

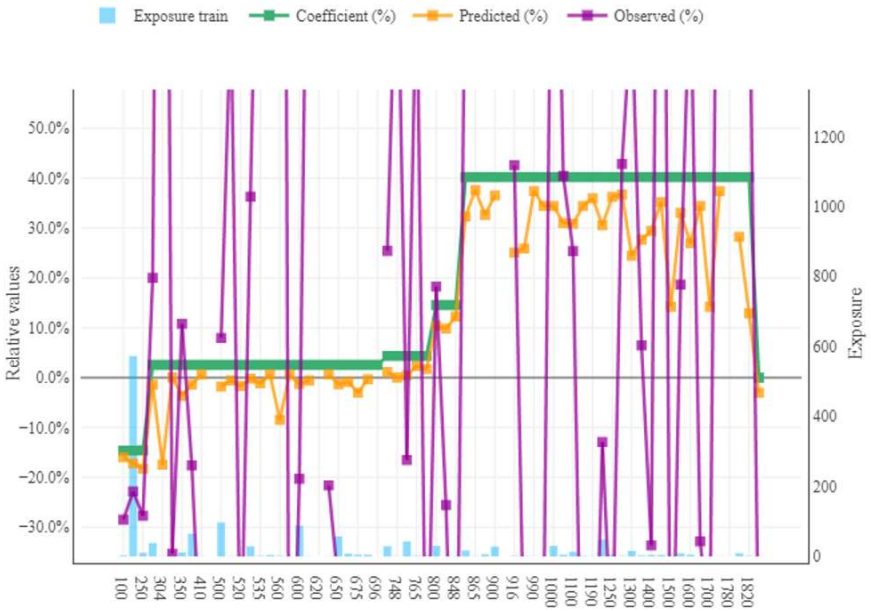


Figure 19 - Impact tarifant de la variable Cylindrée

La prime pure obtenue en application des deux modèles exposés ci-dessus (fréquence x coût) sera comparée avec une prime pure issue de la modélisation en utilisant la distribution Tweedie.

2.5. Modélisation de la prime pure (Tweedie)

Tweedie est une famille de distributions de probabilité qui permet de modéliser la prime pure directement, sans passer par l'estimation de la fréquence et du coût moyen. Cette distribution est particulièrement applicable lorsqu'une grande partie des variables de prédiction Y sont des valeurs à zéro, ce qui la rend utile lors de la modélisation des sinistres.

La distribution Tweedie est en fait une famille de distributions appartenant à la classe des modèles exponentiels de dispersion avec une variance de la forme $Var(Y) = \varphi\mu^{\rho}$, où $\varphi > 0$ représente le paramètre de dispersion/échelle et μ représente la moyenne. ρ doit appartenir à l'intervalle $(-\infty, 0] \cup [1, \infty)$.

La famille Tweedie comprend des distributions continues telles que la distribution Normale et Gamma, la distribution de Poisson exclusivement discrète, et la classe de distributions composées mixtes Poisson-Gamma. Ces distributions sont un cas spécial de la distribution Tweedie, défini par le paramètre ρ : $\rho = 0$ la distribution normale, $\rho = 1$ poisson, $\rho = 2$ gamma, $\rho = 3$ normale inverse.

Pour d'autres valeurs de ρ non remarquables, les distributions sont toujours définies mais ne peuvent pas être écrites dans une forme finie. Quand $1 < \rho < 2$, les distributions sont continues pour Y supérieur à zéro, avec une quantité positive pour $Y = 0$. Pour $\rho > 2$, les distributions sont continues pour Y supérieur à zéro. Le choix de ρ peut se faire simplement en analysant les résidus. La variable dépendante doit être numérique, avec des données supérieures ou égale à zéro. La valeur fixe du paramètre de la distribution Tweedie peut être n'importe quelle valeur supérieure à 1 et inférieure à 2.

La figure 20, présente une sélection des modèles de prédiction de la prime pure sur la base de la distribution Tweedie à paramètre ρ =1,5. Le modèle le plus prédictif comporte 11 variables initialement (12 avec l'enrichissement par code INSEE).

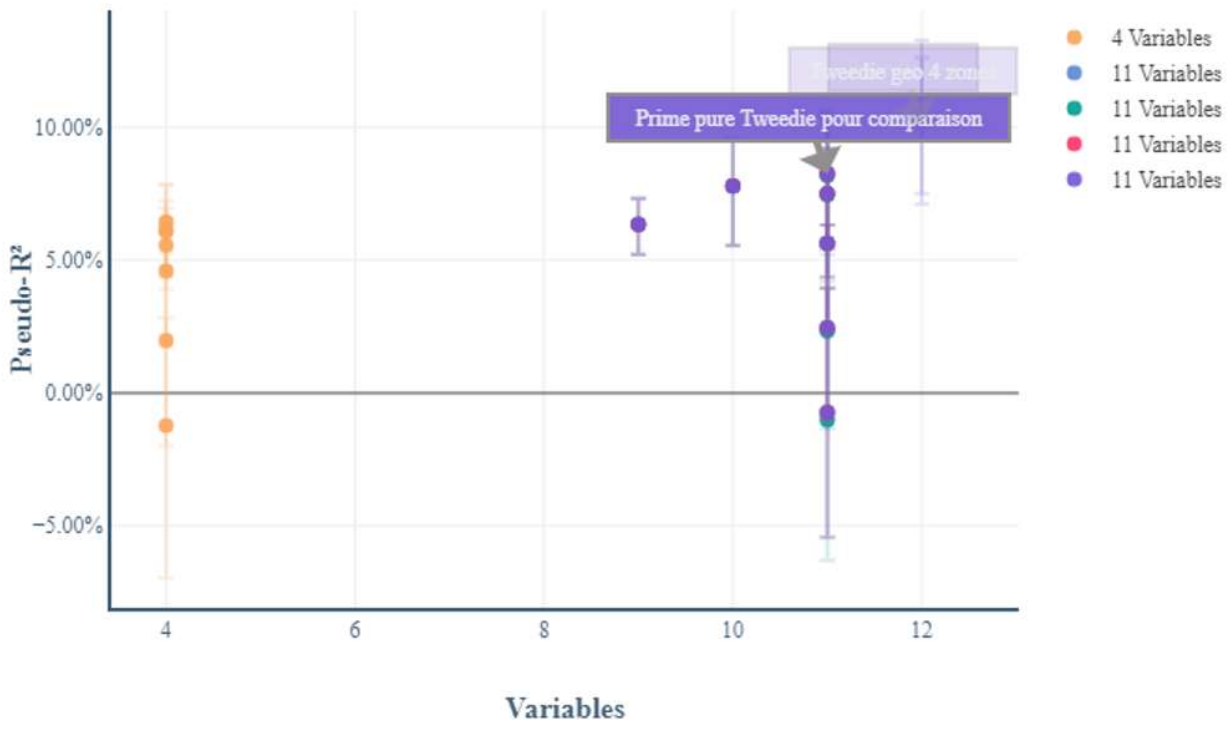


Figure 20 - Performance des modèles Tweedie

Les variables retenues sont significativement les mêmes que celles obtenues lors de la modélisation par fréquence et coût moyen (voir Figure 21). Uniquement la variable Tranche d'Age de conducteur est ajoutée. Notons également que la variable Profil a un impact tarifant.

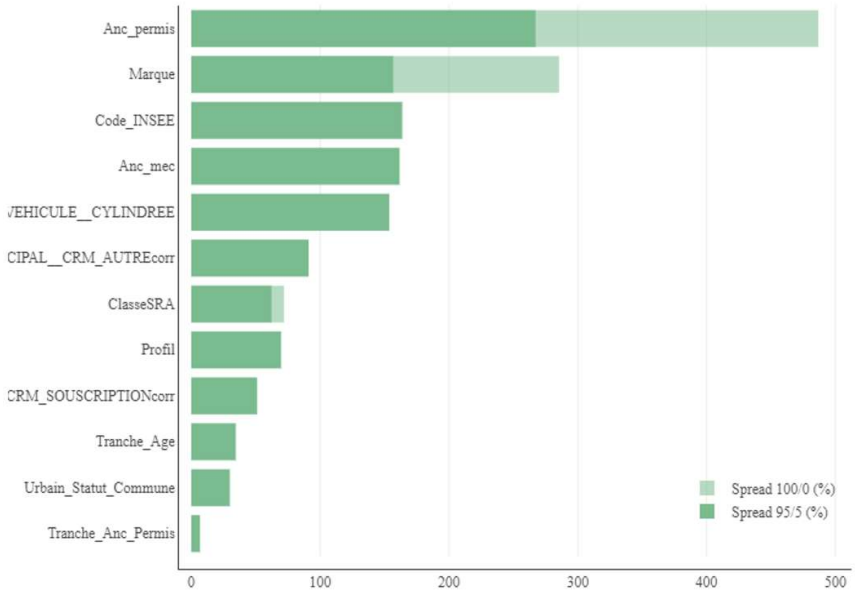


Figure 21 - Variables significatives prime pure Tweedie

Le résultat de la modélisation de l'impact tarifaire de la variable Ancienneté de permis présente des incohérences, les zones 18-26 ans d'ancienneté de permis et 36-42 ancienneté de permis se relèvent moins chères que les zones à 28-34 ans. Il sera donc nécessaire de lisser les coefficients dans le cas où le modèle Tweedie sera retenu.

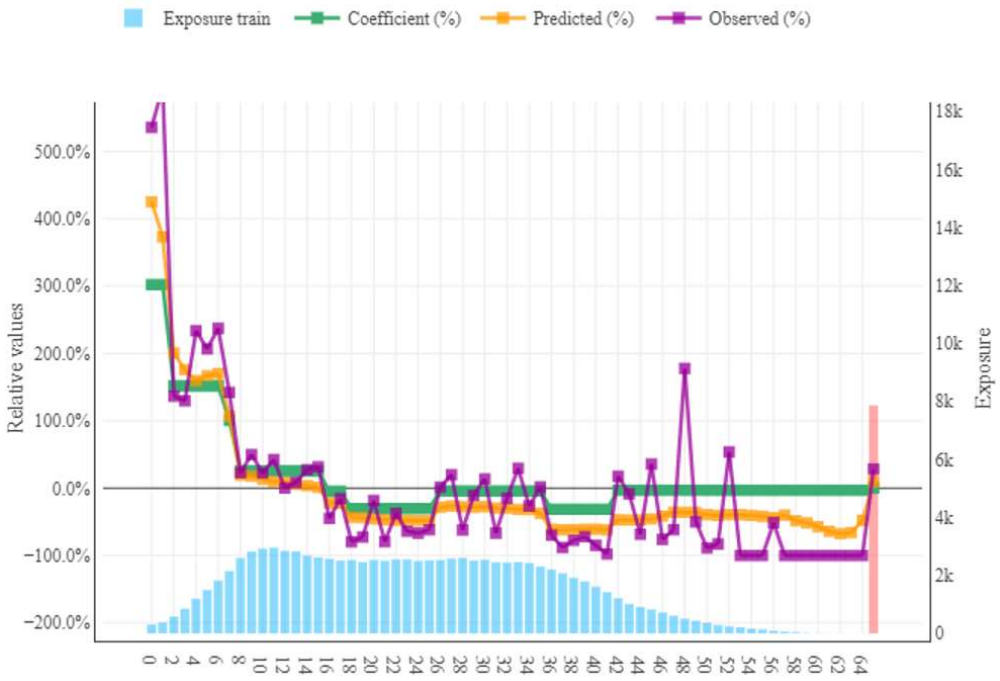


Figure 22 - Impact tarifant de l'Ancienneté du permis Tweedie

2.6. Critères de validation d'un modèle

Les deux modèles sont obtenus - un modèle classique fréquence x coût moyen et un modèle de prime pure obtenue directement en appliquant la distribution Tweedie. Se pose la question des critères du choix entre les deux. Plusieurs critères entrent en jeu, d'ordre statistique ou technique, qui correspondent à des méthodes de validation théoriques, mais également les critères opérationnels, comme logique actuarielle ou les contraintes du marché.

Les critères techniques les plus répandus sont les suivants :

Le coefficient de détermination linéaire de Pearson (R^2) – correspond au pourcentage de variation dans la variable prédite expliqué par la régression. Pour comparer deux modèles ayant le même nombre de variables explicatives il s'agit de comparer les R^2 obtenus et choisir le modèle pour lequel le R^2 est le plus grand. Le R^2 représente

également la corrélation entre les valeurs prédites et les valeurs observées. Le R^2 est proportionnel à la variance de la variable de réponse. A pourcentage d'erreur égal dans les prédictions, une base de tests avec une variance plus faible aura un R^2 plus faible. Le coefficient R^2 ajusté pénalise la métrique R^2 pour un nombre croissant des variables prédictives dans le modèle.

L'erreur quadratique moyenne (RMSE) – est une erreur moyenne de prédiction. La comparaison des modèles se fait sur la base des indicateurs d'erreurs - l'erreur quadratique moyenne, l'erreur absolue moyenne et l'erreur euclidienne. Un modèle est plus pertinent si son indicateur est plus faible.

L'indice Gini est une métrique qui permet de quantifier la performance du modèle pour segmenter les observations. Un modèle à un coefficient Gini plus élevé permettra le tri des risques plus fin et donc une tarification au plus près du risque.

Les statistiques présentées dans le Tableau 14 sont obtenues en application de la méthode de validation croisée k-fold. Cette méthode est appliquée pour évaluer la performance des modèles sur un échantillon de données limité. Les données d'étude ont été divisées en 80%/20% en attribuant un nombre aléatoire de 1 à 10 à chaque ligne, les lignes de 1 à 8 servent ont servi de base d'apprentissage et les lignes 9 et 10 ont servi pour le test des modèles.

Nous nous focaliserons sur les statistiques obtenues pour le test. Les deux modèles, fréquence Poisson et prime pure Tweedie ont des indices Gini très proches, de l'ordre de 47%. Les coefficients obtenus par les deux modèles résulteront en des inégalités fortes entre les assurés. Les assurés les moins risqués auront un tarif très bas, alors que les assurés plus risqués auront un tarif beaucoup plus élevé. Notons que le modèle de prime pure Tweedie présente une qualité de prédiction (R^2) légèrement plus importante mais également une volatilité plus importante.

Il est donc impossible de faire le choix entre ces modèles sur la base seule des statistiques présentées. Une analyse des coefficients produits par rapport aux critères de logique actuarielle, au niveau de mutualisation attendue et au niveau des primes observées sur le marché reste déterminante. La simplicité d'implémentation opérationnelle peut, dans certains cas, également entrer en ligne de compte.

Tweedie Pure Premium				Poisson Frequency			Gamma Average Cost
METRIC	TRAIN FULL	TRAIN K-FOLD	TEST K-FOLD	TRAIN FULL	TRAIN K-FOLD	TEST K-FOLD	TEST K-FOLD
GINI	59.15 %	61.5 %	46.82 %	52.23 %	54.04 %	46.42 %	13.82 %
NORM. GINI	59.21 %	61.56 %	46.87 %	52.44 %	54.25 %	46.6 %	21.31 %
PSEUDO-R²	16.1 %	17.39 %	10.16 %	9.01 %	9.58 %	7.05 %	3.29 %
RMSE	4967	4824	4112	11.73	11.44	10.2	5949
DEVIANCE	3779000	2792000	1012000	17490	13040	4467	586.8
AVG. DEVIANCE	35.05	34.52	37.54	0.1622	0.1612	0.1657	1.757
MAE	56.37	55.98	56.01	0.03219	0.03218	0.03223	10110000000000
NB. VARIABLES	12	11.75	11.75	9	9	9	3.5
MISSING ZIP	32.28 %	32.28 %	32.28 %	32.17 %	32.17 %	32.17 %	
OBSERVED TARGET AVERAGE	29.77	29.77	29.77	0.01632	0.01632	0.01632	3262
PREDICTED TARGET AVERAGE	27.01	26.63	26.63	0.01632	0.01632	0.01632	3166

Tableau 14 - Tests statistiques des modèles au global

2.7. Résultats de modélisation par profil-type

Les performances de prédiction des meilleurs modèles de fréquence par profil-type ont été comparées à la performance du modèle de fréquence au global, sans regroupement des variables. Le Tableau 15 présente les statistiques obtenues qui ne valident pas l'hypothèse d'une performance plus importante d'une modélisation par profil-type.

En effet, le modèle de fréquence au global semble proposer un meilleur compromis de prédictivité et différenciation des assurés avec le R^2, à 7%, une déviance moyenne à 16% et l'indice Gini à 46%.

Fréquence modèle Global		Expert	CityDrive	Rural	Profil Autre
METRIC	TEST K-FOLD	TEST K-FOLD	TEST K-FOLD	TEST K-FOLD	TEST K-FOLD
GINI	46.42 %	38.33 %	35.76 %	23.39 %	38.66 %
NORM. GINI	46.6 %	38.43 %	35.95 %	23.42 %	38.86 %
PSEUDO-R^2	7.05 %	4.05 %	3.98 %	-2.23 %	5.32 %
RMSE	10.2	0.1931	8.3	0.1943	13.61
DEVIANCE	4467	452.9	1036	420.8	2762
AVG. DEVIANCE	0.1657	0.09935	0.2286	0.07912	0.22
MAE	0.03223	0.01776	0.04569	0.01221	0.04389
NB. VARIABLES	9	9	9	8	9

Tableau 15 - Tests statistiques des modèles Fréquence par Profil-type

2.8. Conclusion

Ce chapitre a d'abord présenté les bases théoriques de la tarification en application de GLM et les principales étapes dans une modélisation.

Nous avons ensuite cherché à déterminer si la variable profil-type, définie dans la première partie de l'étude, a un apport significatif pour la modélisation de fréquence et de coût moyen et si la modélisation des risques séparément par profil-type sera plus performante que la modélisation sur l'ensemble des profils sans segmentation. Nous avons utilisé la méthode de validation croisée (K-Fold) en tant que critère de performance des modèles alternatifs testés.

Pour la modélisation de la fréquence, il s'avère que les profils-types n'apportent pas de performance supplémentaire. La modélisation à travers des bases séparées par profil-type diminue même la performance du modèle. En revanche, les tests d'ajout des

données géographiques se sont avérés concluants. Le zoning a un impact positif sur la performance du modèle de la fréquence.

Concernant le coût moyen, ni la variable du profil-type ni les données géographiques ne se retrouvent parmi les variables significatives. Le coût moyen étant uniquement impacté par la cylindrée, le CRM, la marque et l'âge de l'assuré.

La modélisation de la prime pure directement, en utilisant la distribution Tweedie, présente une alternative acceptable à la modélisation traditionnelle fréquence x coût.

Maintenant que le modèle de tarification optimisé est défini, se pose la question des critères de prise de décision pour l'assureur pour procéder au changement de méthode, les conséquences d'une telle décision et les autres options disponibles.

TROISIEME PARTIE :

METHODES ALTERNATIVES DU REDRESSEMENT DU PORTEFEUILLE

Comme précisé dans le chapitre précédent, le modèle GLM est sensible aux paramètres et aux données utilisées. Ainsi, le changement d'un de ces deux éléments donne lieu, sans surprise, à un changement de la tarification. L'assureur doit-il la changer à chaque fois qu'il réalise une nouvelle étude actuarielle ?

Plusieurs contraintes métier l'en empêchent, tout en sachant que les décisions tarifaires ont un vrai impact stratégique sur la vie de l'assureur de niche. Entrer en guerre de prix avec les leaders du marché n'est pas une option. L'objectif est donc d'augmenter la rentabilité tout en gardant ou augmentant le chiffre d'affaires.

Dans ce chapitre, nous explorons les critères d'évaluation qui justifieront un changement de tarification. Lorsqu'un tel changement ne s'avère pas justifié, nous considèrerons d'autres stratégies possibles (avec ou sans changement de la tarification) pour améliorer la rentabilité.

1. Impact des décisions tarifaires sur la rentabilité du portefeuille

1.1. De la prime pure à la prime commerciale

Rappelons que le prix du risque n'est pas le seul élément déterminant la prime commerciale. La figure 23 montre ses autres composantes, notamment les charges fixes et variables ainsi que la marge attendue par l'assureur.

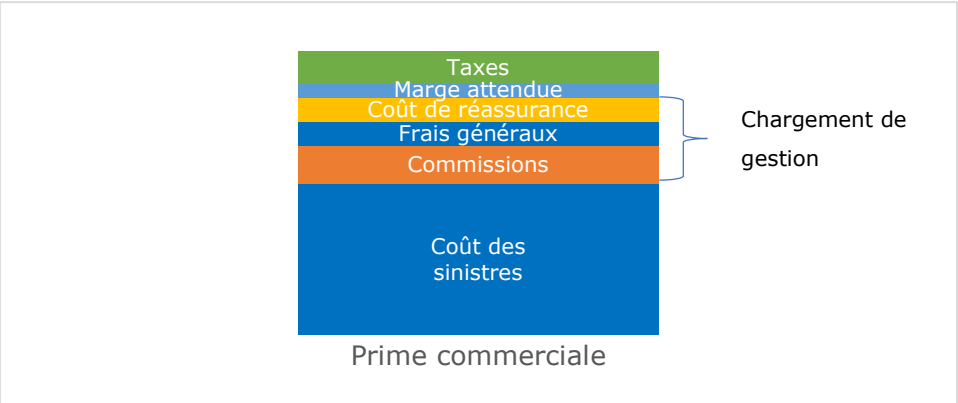


Figure 23 - Composantes de la prime commerciale

Le classement entre les coûts fixes et les coûts variables dépend du business modèle de l'assureur. Dans un modèle de ventes directes et de gestion de sinistres internalisée la majorité des coûts de l'assureur seront des coûts fixes. Analysons les trois principales catégories des dépenses d'un assureur avec la gestion externalisée :

Commissions :

Dans un modèle où la distribution et la gestion des contrats et sinistres sont externalisées, les coûts de gestion prendront majoritairement la forme de commissions versées aux partenaires indexées sur le volume des primes et donc des coûts variables. Dans le modèle de vente indirecte, les commissions jouent ainsi un rôle non négligeable. La commission versée au partenaire va tenir compte des actes de gestion qui sont attendus du partenaire – commercialisation, gestion des contrats, gestion des sinistres, traitement des réclamations etc.

Frais Généraux :

Les frais généraux correspondent aux coûts de fonctionnement d'une entreprise au quotidien – les dépenses de loyer, des salaires, de location du matériel et des logiciels etc. Malgré l'externalisation des certaines fonctions, les frais généraux restent une ligne de dépense importante.

Coût de réassurance :

La réassurance est l'assurance des assureurs pour faire face notamment aux sinistres graves. A l'équivalent de l'assurance, la partie qui cède le risque, appelée cédante, va payer une prime d'assurance, appelée dans ce cas une prime de réassurance, pour couvrir une partie ou l'intégralité des risques qu'elle porte. La réassurance permet aux

assureurs d'accepter des risques pour lesquels, sans appel à la réassurance, ils n'auraient pas suffisamment de capitaux propres. La nature du risque RC, en particulier des dommages corporels d'un montant illimité, rend l'appel à la réassurance indispensable et représente un coût pour l'assureur. Les principales formes de la réassurance sont rappelées en Annexe 4.

Le schéma de réassurance et donc son coût va dépendre de l'appétit au risque de l'assureur mais également de l'équilibre de ses besoins des capitaux propres requis par les exigences de la Solvabilité 2.

Les sujets d'optimisation des chargements de gestion sont multiples et mériteraient une discussion à part entière. Pour notre étude, il est important de comprendre que le niveau de tarification retenu doit tenir compte des hypothèses de ces chargements qui sont considérées comme fixes.

1.2. Démarche de validation des conclusions tarifaires

Afin d'évoluer et projeter la performance d'un portefeuille, deux ratios sont utilisés en assurances : sinistres sur primes (S/P) et ratio combiné. Le ratio combiné désigne le rapport entre les coûts totaux (dont le coût des sinistres et de gestion) et les primes.

Lors de la validation des modèles tarifaires, l'actuaire doit confronter ses résultats à un ratio S/P cible, au-dessus duquel la commercialisation du produit ne sera pas rentable pour la société, compte tenu des charges évoquées précédemment.

Il est également nécessaire de confronter les résultats de l'étude aux autres travaux menés par les actuaires de l'assureur tels que le calcul du résultat technique, les analyses des hypothèses de provisionnement, les analyses des résultats des populations couvertes pour le renouvellement, la cohérence du S/P obtenu avec le niveau de rentabilité vu par les souscripteurs.

1.3. Facteurs externes en jeu lors du changement de la tarification

Le prix commercial est fondé sur l'ensemble des éléments ci-avant mais également contraint par la réglementation et le prix du marché. Dans un monde idéal pour un

assureur, ces contraintes pourraient être accompagnées des informations sur la sensitivité du client au prix et la valeur estimée de l'apport de ce client en portefeuille. Ainsi, l'assureur aurait pu proposer à chaque prospect intéressant le maximum du prix que le prospect pourrait accepter. Une telle approche risque de ne pas être appréciée par le régulateur, mais le mécanisme du marché fera le nécessaire d'un point de vue éthique.

Afin d'évaluer l'impact du changement tarifaire sur la rentabilité, le concept d'élasticité de la demande doit être évoqué. L'élasticité correspond à la sensibilité de l'assuré au changement du prix et varie de sensibilité forte (demande élastique) à sensibilité faible (demande non élastique). Un client au comportement non élastique renouvèlera une police avec une hausse tarifaire plus forte alors qu'un client au comportement élastique résiliera, à la suite d'une hausse, même faible, du tarif.

L'élasticité en fonction du prix peut être formulée comme suit :

$$E = -\frac{(\mu_{P1} - \mu_{P0)}/\mu_{P0}}{(P1 - P0)/P0}$$

Avec :

μ_{P1} – *demande au prix P1 (nouveau prix)*

μ_{P2} - *demande au prix P0 (prix initial).*

La modélisation de l'élasticité consiste à proposer un modèle permettant de prédire si le client accepte ou refuse une proposition d'affaire nouvelle ou de renouvellement en fonction des facteurs relatifs au prix. Ces facteurs peuvent être internes à la société (politiques tarifaires) ou externes (activités de la concurrence). Comme la société ne peut pas influencer la concurrence, les modèles se focalisent sur les facteurs internes.

L'élasticité est également impactée par la nécessité du produit et les possibilités de substitution ainsi que par le revenu disponible et la durée de l'impact de la hausse. Le marketing peut également impacter l'élasticité. Cependant, les effets de ces facteurs sont difficilement quantifiables et sont donc en dehors de l'objet de cette étude.

La demande, et donc l'élasticité, peuvent être évaluées à travers le taux de rétention, pour les renouvellements (% des polices renouvelées), et le taux de conversion pour les affaires nouvelles (% des offres acceptées).

Afin de définir le niveau optimal de tarification, l'assureur doit commencer par définir une fonction de bénéfice, dont les composantes principales sont les suivantes :

- Modèle de la prime pure
- Frais généraux (ensemble de frais qui doivent être considérés) – frais d'acquisition, d'administration et de gestion, éventuellement à niveau différent en fonction des profils des assurés.
- Une latitude pour des gestes commerciaux.
- Taux de rétention ou de conversion modélisés en utilisant la modélisation de la demande
- Prime proposée à l'assuré potentiel
- Enrichissement du modèle pour tenir compte de la valeur ajoutée de chaque client potentiel
- Fonction de bénéfice elle-même, construite en fonction de l'horizon temporel de l'étude.

A l'issue de procédures d'optimisation, plusieurs scenarii de combinaisons de nombre de polices et de prix sont possibles, mais une seule combinaison résultera en maximum de la profitabilité pour un modèle de la demande globale.

Le maximum de la profitabilité est atteint si l'ensemble des polices se trouve sur la courbe des primes optimisées (Guven et McPhail, 2013). Cette ligne correspond à la frontière d'efficience par rapport au maintien de la demande et de la rentabilité attendue par un assureur. En suivant sa montée, la rentabilité diminue alors que la demande augmente, mais chaque point représente le niveau de rentabilité optimal. La Figure 24 présente un exemple d'une société dont la position n'est pas optimale. Cette société a plusieurs moyens d'augmenter sa rentabilité : 1) augmenter la demande en gardant le S/P au même niveau, 2) diminuer le S/P à demande stable ou 3) une combinaison des deux.

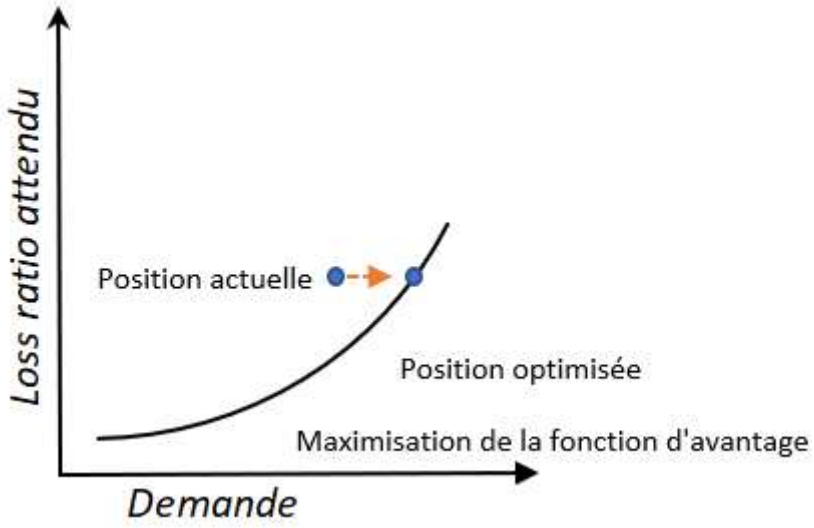


Figure 24 - Frontière d'efficience d'un portefeuille optimisé

Les auteurs proposent deux méthodes d'optimisation :

- Agir sur la grille de tarification en optimisant les facteurs des variables tarifantes. Facilement implémentée avec GLM, cette méthode ne permettra pas pour autant d'identifier les lacunes de tarification sur des segments où le potentiel de croissance n'est pas atteint. Autrement dit, l'assureur aurait pu augmenter son résultat net en baissant le prix sur certains segments et en attirant ainsi des assurés supplémentaires.
- Optimiser les prix sans tenir compte de la tarification actuelle, en déterminant un niveau de prime optimal pour chaque assuré dans le portefeuille et déterminant ensuite la grille tarifaire en appliquant l'ingénierie inverse. Cette méthode permettra à l'assureur de se positionner sur la frontière d'efficience, mais elle est beaucoup plus consommatrice de temps.

Il est à noter que le modèle d'optimisation ne sera pas le même entre :

- Une affaire nouvelle dont le prix optimal sera davantage impacté par les prix de la concurrence et l'image de la marque, mais aussi par le caractère immédiat de la décision tarifaire, et
- Un renouvellement où l'assureur proposera le prix qui sera fonction de sa stratégie de maintenir les segments des assurés profitables, de l'historique et de la valeur attendus de l'assuré individuel en question dans le futur (par exemple à travers la

vente d'autres produits) mais également davantage de temps pour tenir compte de toutes les variables.

Une analyse des scénarii agrégés des deux modèles d'optimisation permettrait de définir les mesures cibles et de vérifier que l'impact estimé soit positif. Nous avons simulé deux scénarii possibles, un avec une hausse tarifaire et un sans hausse, pour estimer l'impact sur le bénéfice (Tableau 16). Afin de maintenir la confidentialité, les montants affichés ne correspondent pas aux portefeuilles réels.

Dans l'exemple présenté l'assureur a effectivement intérêt à augmenter les primes car l'impact sur le bénéfice est positif. Cependant, le résultat est très sensible aux hypothèses de taux de transformation, de taux de renouvellement et du prix commercial appliqué.

Total (Affaires nouvelles + Renouvellement)

	Annee	Policies quote	Polices souscrites	Polices renouvelles	Primes Emises	Marge	Benefice
Scenario 1	1	2 000	500	800	780 000	2,5%	19 500
Base	2	2 400	550	1 040	982 620	2,5%	24 566
	3	2 800	605	1 272	1 194 786	2,5%	29 870
	4	3 208	666	1 502	1 421 157	2,5%	35 529
	5	3 633	733	1 734	1 666 032	2,5%	41 651
	TOTAL						151 115

Total (Affaires nouvelles + Renouvellement)

	Annee	Policies quote	Polices souscrites	Polices renouvelles	Primes Emises	Marge	Benefice
Scenario 2	1	2 000	400	750	759 000	3,0%	22 770
Hausse des prix	2	2 250	440	863	885 440	3,0%	26 563
	3	2 513	484	977	1 022 896	3,0%	30 687
	4	2 792	533	1 096	1 174 587	3,0%	35 238
	5	3 094	586	1 221	1 342 670	3,0%	40 280
	TOTAL						155 538

Comparaison Scenario 2 - Scenario 1	4 423

Tableau 16 - Exemple d'analyse de scenarii de hausse tarifaire

Une analyse de sensibilité par paramètre est présentée dans le tableau 17. Dans le cas étudié, nous avons testé l'impact d'une hausse de prix moins forte que dans le scenario central de 10% (de 3% à 7%) et une sensibilité des taux de transformation et de renouvellement moins importants que les scénarii de base de -10% et -5%. Il en ressort de cette analyse que le résultat du portefeuille est plus sensible au taux de

renouvellement qu'au taux de transformation. Dans ce cas de figure, l'assureur doit se focaliser sur l'étude des mécanismes de modélisation du taux de renouvellement.

Changement du taux de transformation

Changement du Prix	4 423	-10%	-8%	-5%	-3%	-1%
	7%	181	5 511	13 357	18 617	23 877
	6%	- 1 233	4 048	11 820	17 031	22 242
	5%	- 2 647	2 584	10 282	15 444	20 606
	4%	- 4 061	1 120	8 745	13 858	18 971
	3%	- 5 475	- 344	7 208	12 272	17 336

Changement du taux de renouvellent

Changement du Prix	4 423	-10%	-8%	-5%	-3%	-1%
	7%	- 12 975	- 7 896	181	5 889	11 870
	6%	- 14 266	- 9 235	- 1 233	4 422	10 347
	5%	- 15 557	- 10 573	- 2 647	2 955	8 824
	4%	- 16 848	- 11 912	- 4 061	1 487	7 301
	3%	- 18 139	- 13 250	- 5 475	20	5 777

Tableau 17 - Sensibilité du résultat aux hypothèses

1.4. Contraintes métier

L'application du modèle « optimal » dans la pratique se heurte à plusieurs contraintes métier :

- Les modèles sont conditionnés à la fonction de l'élasticité et la moindre erreur d'estimation, en particulier lorsqu'on diminue la tarification pour attirer la clientèle, pourra engendrer une perte pour l'assureur.
- Les effets, dans un contexte de vente indirecte, mettent plus de temps à se manifester.
- L'ajout d'un critère tarifant, par exemple l'âge de l'assuré, ou lieux de stationnement peut engendrer des coûts d'implémentation à cause des contraintes informatiques ou opérationnelles (les courtiers ont un système d'information établi qui collecte telle ou telle information, un changement et une revue des éditiques sont nécessaires)

- L'assureur ne peut pas changer le produit seul (contexte de partenariats avec des courtiers, qui sont impactés commercialement par les décisions aussi bien tarifaires mais aussi relatives aux produits)
- Dans la pratique, la démarche peut s'étaler sur plusieurs années.
- Impact sur le mix : lorsque l'assureur change le tarif, sa compétitivité sur chaque client évolue, ce qui a un impact sur la structure entière du portefeuille qui se trouve déformée. En particulier, une erreur de tarification qui impliquerait un tarif moins élevé que le marché sur les profils à risque élevé risque de générer l'effet d'antisélection, en attirant ce type de profils vers l'assureur.
- L'effet d'ajustements ponctuels est difficilement différentié de l'évolution naturelle de la sinistralité.
- Certaines contraintes règlementaires peuvent limiter l'utilisation de tel ou tel modèle de tarification. Par exemple, il est interdit en Europe de tarifier en fonction du sexe (Gender Directive du 21 décembre 2012). Il y a des pays où il est interdit de proposer un prix plus faible en affaire nouvelle au détriment des assurés à des profils similaires en renouvellement.

2. Autres leviers d'amélioration de la rentabilité

A part l'optimisation de la tarification, l'assureur dispose d'autres leviers pour améliorer la rentabilité technique de son portefeuille. Ces techniques visent à réduire la charge.

2.1. Sélection des risques

La première approche consiste à mieux sélectionner les assurés à travers les critères de souscription, observables ou non par l'assuré.

Dans le cas des critères de souscription observables, l'assureur refusera de proposer un produit à la catégorie de population qu'il ne souhaite pas avoir dans son portefeuille (par exemple les conducteurs de moins de 26 ans). Ce type de levier peut être restreint pour l'utilisation dans certains pays où l'assureur est obligé de vendre à toutes les catégories d'assurés. Ce n'est pas le cas de la France actuellement.

Dans le cas des critères de souscriptions non observables, l'assureur qui souhaite limiter son exposition à un segment de population en particulier, proposera aux autres segments un tarif significativement plus élevé que celui proposé sur le marché, de façon à dissuader l'achat.

2.2. Réduction des garanties ou des plafonds de garanties

Un autre levier, bien que limité, consiste à la réduction des garanties proposées, ou la baisse des plafonds de garanties. Ces changements sont généralement mal perçus par les assurés et, par conséquent, peuvent avoir un impact négatif sur le taux de renouvellement. Dans un contexte de co-construction de produit entre l'assureur et le distributeur, ce levier est assez limité.

L'action inverse, de proposer des options profitables, peut s'avérer plus productive.

2.3. Hausse des franchises

La franchise est la somme qui reste à charge de l'assuré dans le cas de la survenance d'un sinistre. Elle peut être appliquée lorsque l'assuré est responsable. En cas de co-responsabilité (RC 50%), le montant de franchise est divisé par deux.

Plusieurs types de franchises existent :

- La franchise absolue correspond à un montant qui reste toujours à charge de l'assuré. Lorsque le coût du sinistre dépasse le niveau de la franchise absolue, l'assuré se voit rembourser la différence entre le montant du sinistre et le montant de la franchise.
- La franchise relative sert à définir le montant à partir duquel l'assureur prendra en charge l'intégralité du sinistre.
- La franchise proportionnelle s'évalue en pourcentage du montant des réparations, avec un montant minimum et un montant maximum.
- La franchise « prêt de volant », généralement plus élevée qu'une franchise standard, peut être utilisée si le conducteur principal prête son véhicule à une

autre personne. L'objectif est de dissuader l'assuré de prendre des risques supplémentaires en prêtant son véhicule à un conducteur dont les critères sont inconnus de l'assureur. Il est également possible d'exclure totalement la possibilité de prêt du véhicule à une autre personne.

- La franchise « jeune conducteur » correspond à une variation de la franchise « prêt de volant » lorsque le conducteur principal prête son véhicule à un conducteur moins expérimenté (notamment moins de 3 ans de permis) alors que le conducteur principal bénéficie de CRM plus avantageux (notamment coefficient 50).

Le montant de franchise est librement établi par l'assureur. Ce moyen peut être utilisé jusqu'à une certaine limite car de plus en plus d'assurés y prêtent attention et la franchise fait désormais partie des critères de comparaison des produits. L'assureur ne devrait pas s'éloigner de l'offre de la concurrence au risque de provoquer une mauvaise image de marque.

2.4. Mesures de prévention

Un autre moyen d'agir sur la rentabilité concerne l'alignement des intérêts entre l'assuré et l'assureur. L'apparition récente des applications de suivi de conduite connectées permet aux assureurs de favoriser la conduite prudente en proposant un tarif moins élevé aux bons conducteurs. Le mécanisme est le suivant : l'application analyse le comportement du conducteur sur la route – la vitesse, l'accélération et le freinage, la vitesse des manœuvres pendant un certain temps. Passé un délai d'observation, le tarif du conducteur est ajusté en fonction des données de conduite obtenues. Une conduite prudente est gratifiée par une remise, une conduite moins prudente ne donne pas lieu au changement du tarif.

2.5. Proposition d'un produit innovant

L'idée du dernier levier de redressement de la rentabilité technique consiste en proposition d'un produit innovant. Citons les produits à l'usage comme une des voies

d’innovation. La tarification de ce type de produit présente un intérêt car il s’agit de trouver un équilibre entre forfait fixe et la partie variable du tarif.

En conclusion, le changement de tarification en l’absence d’information sur le positionnement compétitif de la politique tarifaire de l’assureur, présente de nombreux risques – changement du mix, évolution des taux de transformation et de renouvellement, antisélection.

Dans ce dernier chapitre, nous avons considéré les mesures d’optimisation de rentabilité dans leur ensemble. Plusieurs de ces mesures peuvent présenter un objet d’étude à part entière, comme l’optimisation du coût de la réassurance ou l’étude de sensibilité des assurés au prix lors du premier achat de police ou de renouvellement.

CONCLUSION

Nous avons exploré l'introduction des profils-type dans le modèle de tarification afin d'obtenir une meilleure adéquation de la tarification au risque. L'étude a été réalisée sur les données de six courtiers français et la garantie responsabilité civile. Il s'est avéré que l'impact de la variable profil-type n'apporte pas d'amélioration significative et n'a donc pas d'intérêt compte tenu de la complexité supplémentaire ajoutée à la modélisation.

En utilisant un outil de modélisation plus performant, notamment du point de vue de la possibilité d'analyse comparative de plusieurs paramètres de modèles, nous avons pu nous focaliser sur le choix d'un modèle le plus performant et l'application pratique.

Les profils obtenus initialement se sont relevés trop restrictifs pour permettre une tarification (de 8 à 15% de la population totale par profil). En revanche, ces profils montrent des écarts significatifs de fréquence et coût moyen. L'assouplissement des critères d'affectation au profil type engendre davantage de bruit dans la modélisation et diminue la significativité des résultats. Une validation du résultat sur une base de données plus large et donc en application de critères plus stricts apporterait probablement des résultats plus probants.

En revanche, nous avons démontré l'apport important des données géographiques sur la performance du modèle, notamment sur la modélisation de la fréquence.

Nous avons cherché à confirmer si le seuil de sinistralité attritionnelle actuellement utilisé par l'assureur est justifié en utilisant une approche graphique de Hill. Bien qu'acceptable, ce seuil reste assez élevé par rapport au profil de la sinistralité observé. Nous n'avons cependant pas testé l'impact de sa baisse sur le résultat de l'étude tarifaire pour rester dans le même référentiel que les équipes et pouvoir bénéficier de leur avis métier sans introduire de biais par le choix d'un seuil différent.

La modélisation traditionnelle fréquence x coût a été comparée à une modélisation directe de la prime pure en utilisant la distribution Tweedie. Les résultats démontrent que cette dernière constitue une alternative suffisamment performante par rapport à la modélisation fréquence x coût.

Nous avons ensuite cherché à déterminer les critères pour justifier la décision d'un assureur de changer de modèle de tarification et dresser la liste des impacts potentiels.

En l'absence des données de sensibilité au prix du taux de transformation et de renouvellement, nous avons proposé une méthode simpliste pour évaluer l'impact sur la rentabilité de la hausse tarifaire. Une étude plus approfondie de la sensibilité et du comportement des prospects pourrait et devrait affiner les conclusions avant la prise d'une décision finale du choix de la politique tarifaire.

Plusieurs composantes du prix commercial ont été considérées fixes, comme le prix de la réassurance et les frais généraux. L'analyse des moyens de leur optimisation constitue un objet d'étude à part entière mais reste en dehors du sujet de ce mémoire.

Enfin, nous avons dressé une liste des leviers alternatifs d'amélioration de la rentabilité d'un portefeuille. L'impact de chacun de ces leviers peut constituer un objet d'étude utile et intéressant.

BIBLIOGRAPHIE

Ouvrages

[1] LEBART L., MORINEAU A., PIRON M. (1995). *Statistique exploratoire multidimensionnelle*. Dunod.

Mémoires

[2] ENSAE ParisTech (2019) A. BARME. Tarification automobile dans un contexte concurrentiel. Mémoire d'actuariat.

[3] ISUP (2018) D. IMBERT. Création d'un modèle d'estimation des tempêtes en France métropolitaine.

[4] FCAS MAAA (2013) S. GUVEN, M. McPhail. Beyond the cost model : understanding price elasticity and its applications.

[5] CEA (20??) Y. NJOMO NANA. Segmentation et tarification de la garantie assistance automobile. Mémoire d'actuariat.

[6] CEA (20??) M. PLISSON, B. ROSSARD. Arbitrage entre mutualisation et segmentation : le cas des assurances dommages de collectivités. Mémoire d'actuariat.

[7] CEA (2013) E. RAIN, L. JACQUES. Du modèle GLM à une approche darwinienne : Nouvelle génération de concepts et d'indicateurs pour l'optimisation du renouvellement Auto.

[8] EURIA (2018) F.-Z. ZOUGGAGH. Tarification automobile à l'aide de modèles de machine learning et apport des données télématiques. Mémoire d'actuariat.

Publications

[9] H. BUHLMANN & A. GISLER (2005). A Course in Credibility Theory and its Applications. Springer.

[10] A. CHARPENTIER (2010). Statistique de l'assurance. 3rd cycle. Université de Rennes 1 et Université de Montré.

[11] J.-P. DINTILHAC (2005). Rapport du groupe de travail chargé d'élaborer une nomenclature des préjudices corporels.

[12] R. LITTLE and D. RUBIN (1987). Statistical analysis with missing data. New York: John Wiley & Sons.

[13] J. A. NELDER and R. W. M. WEDDERBURN (1972). Generalized linear models. Journal of the Royal Statistical Society, 135, 370–384.al, pp.133..

[14] J.L SCHAFER and J.W GRAHAM (2002). Missing Data: Our view of the state of the art. Psychological Methods.

[15] C. SCARROTT and A. MACDONALD (2012). A review of extreme value threshold estimation and uncertainty quantification. REVSTAT – Statistical Journal Volume 10, Number 1, March 2012, 33–60.

Cours

[16] IRMAR, Université de Rennes 1 (2016) V. MONBET Modèles linéaires généralisés.

Sites internet

[17] Institut National de la statistique et des études économiques (2021) Indice des prix à la consommation.

[18] Fédération Française de l'Assurance (2020) Assurances de biens et de responsabilité : données clés par année 2019.

[19] Fédération Française de l'Assurance (2021) L'assurance des motos, scooters et autres deux-roues à moteur.

[20] Fédération Française de l'Assurance (2021) Assurance auto, moto : le bonus-malus.

[21] Jason BROWNLEE, Python Machine Learning (2020) How to Configure k-Fold Cross-Validation

Autres

[22] J.Y. BAUDOT - concepts et techniques organisationnelles, descriptives, prédictives et prévisionnelles pour l'entreprise, la finance et l'économie (et fondements mathématiques).

ABREVIATIONS

Garanties :

ACCESS - Accessoires

ATTEN - Attentat

BDG – Bris de glace

CASQ - Casque

CATECH – Catastrophe technologique

CATNAT – Catastrophe naturelle

CIR - Circuit

DCX – Dommage collision

DRX – Défense recours

DTA – Dommages tous accidents

EQUIMOT – Equipement moto

INC - Incendies

OPPROPIL – Protection du pilote en option

PFX – Perte financière

PJ – Protection juridique

PROPIL – Protection du pilote

PTX – Perte totale

RCX – Responsabilité civile

RFRA – Rachat de franchise

SRA – L'association Sécurité et Réparation Automobile

VLX - Vol

VNEUFXX – Valeur à neuf, pour le nombre des mois XX

LISTE DES TABLEAUX

LISTE DES FIGURES

ANNEXES

Annexe 1 : Les projections techniques affichées sont-elles cohérentes par rapport au marché ?

% des primes émises brutes de réassurance	2015	2016	2017	2018	2019
Primes cédées	7,7 %	10,8 %	11,8 %	11,6 %	12,4 %
Charge des prestations	83,0 %	85,9 %	81,8 %	79,3 %	81,9 %
dont sinistres (nets de recours) payés	63,3 %	66,0 %	66,5 %	67,8 %	67,4 %
dont frais de gestion des sinistres	8,5 %	8,2 %	8,1 %	8,3 %	8,0 %
dont variations des provisions	11,2 %	11,7 %	7,2 %	3,3 %	6,5 %
Charges d'acquisition et de gestion	19,5 %	19,8 %	19,8 %	19,7 %	19,7 %
Solde financier	6,6 %	6,0 %	6,2 %	6,1 %	4,9 %
Résultat de la réassurance	- 0,3 %	0,4 %	- 1,4 %	-1,0 %	- 0,4 %
Résultat technique	3,2 %	0,2 %	2,5 %	5,4 %	2,1 %
Provisions techniques	220,2 %	227,2 %	232,1 %	231,9 %	232,6 %

Tableau 18 - Assurance automobile - Principaux ratios comptables

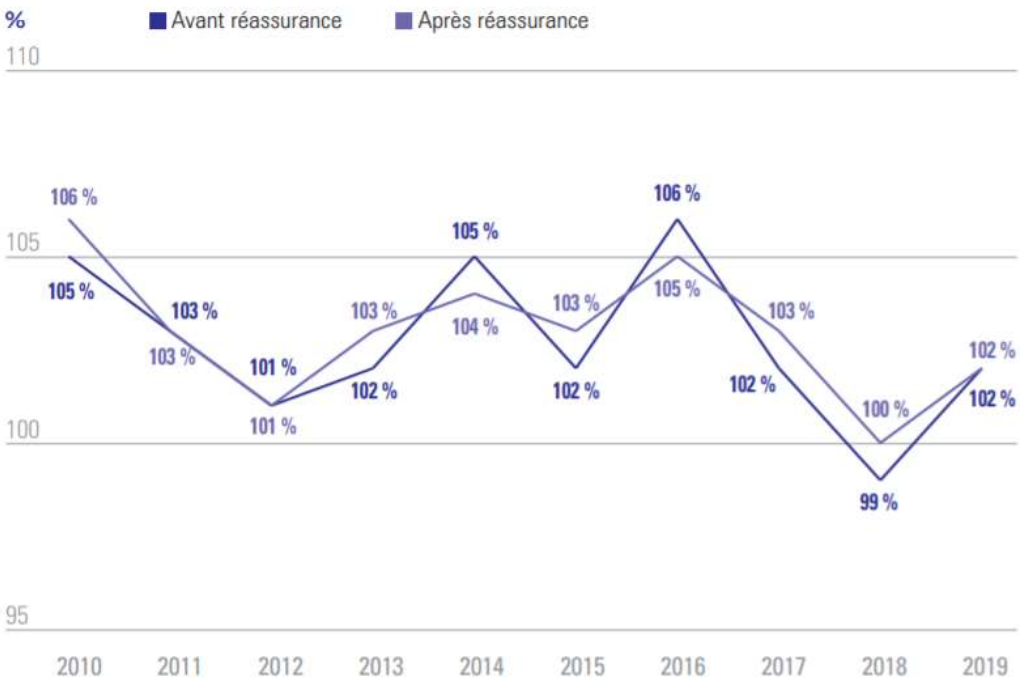
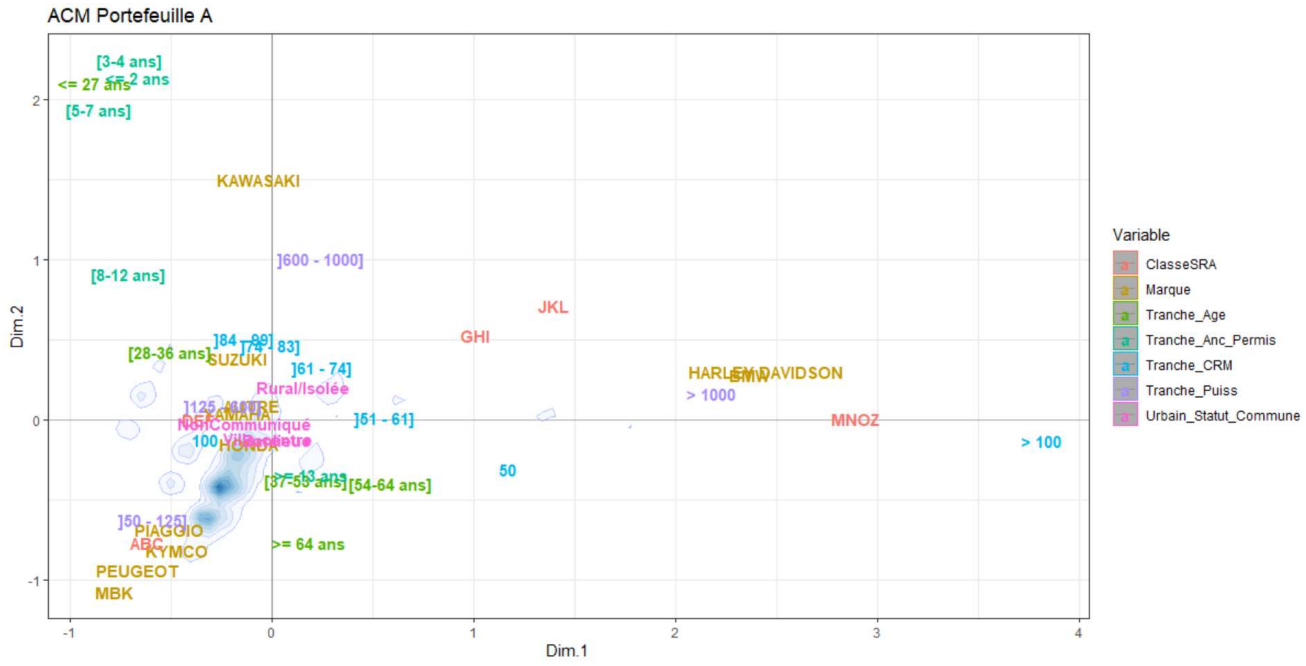
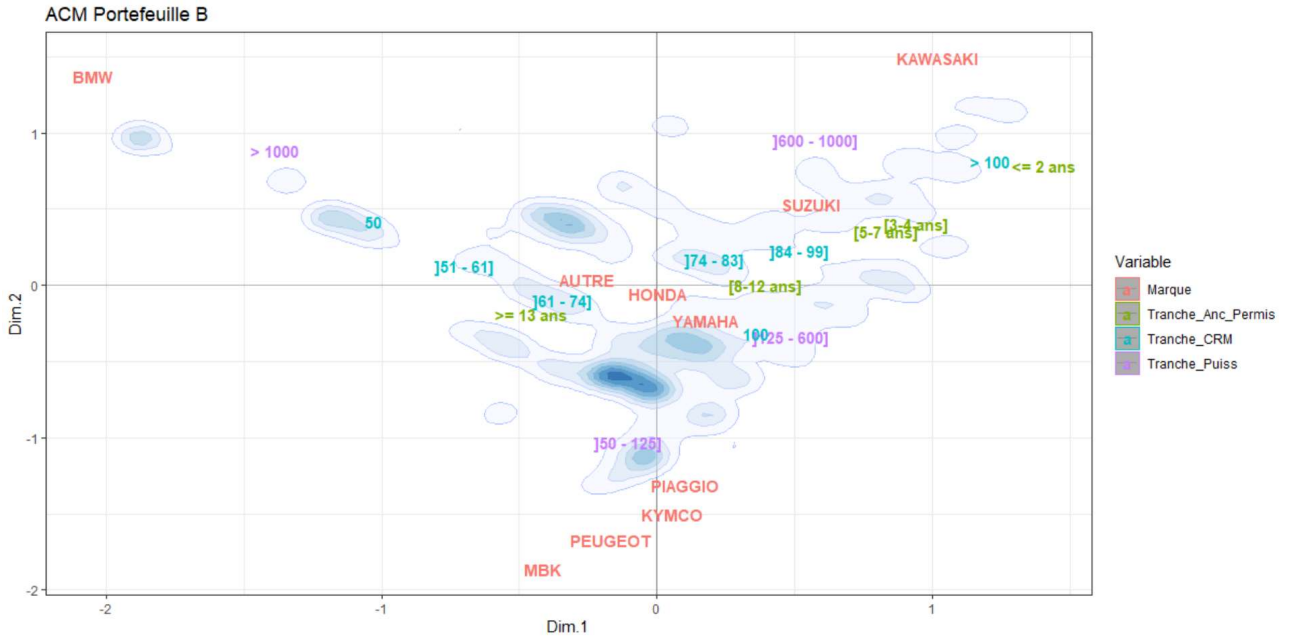


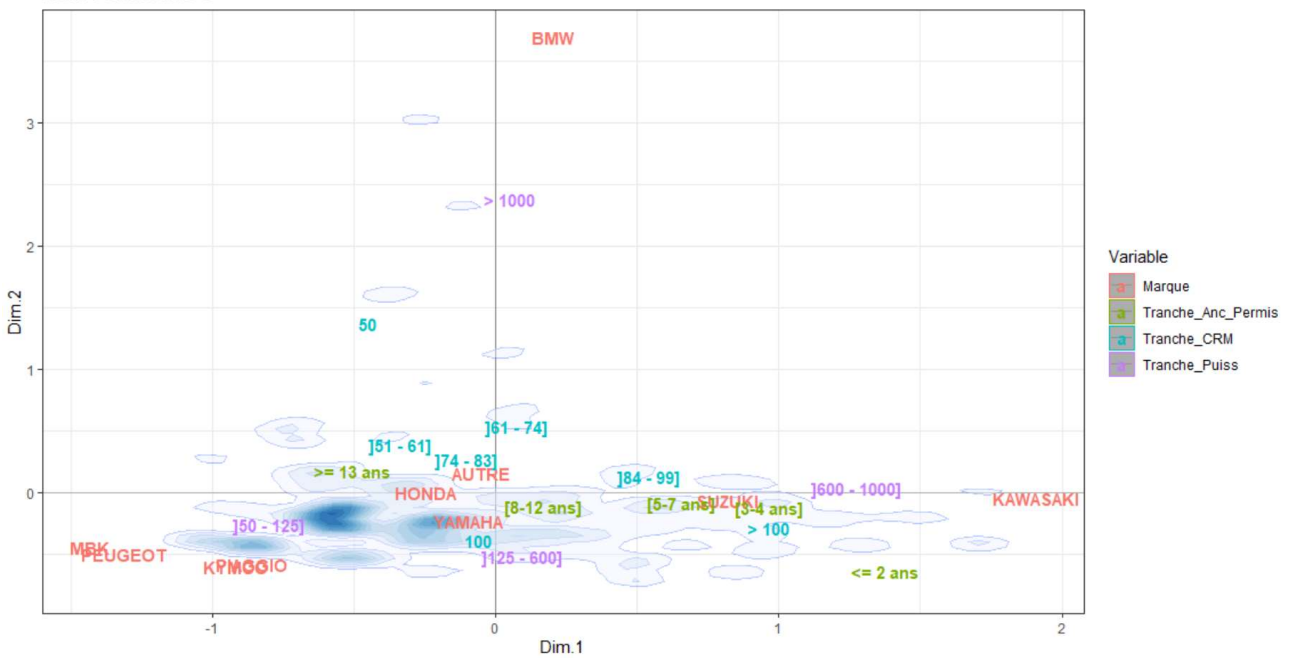
Figure 25 - Assurance automobile – Ratio combiné comptable

Annexe 2 : Répartition par profil-type choisie est-elle similaire et cohérente pour chacun des portefeuilles ?

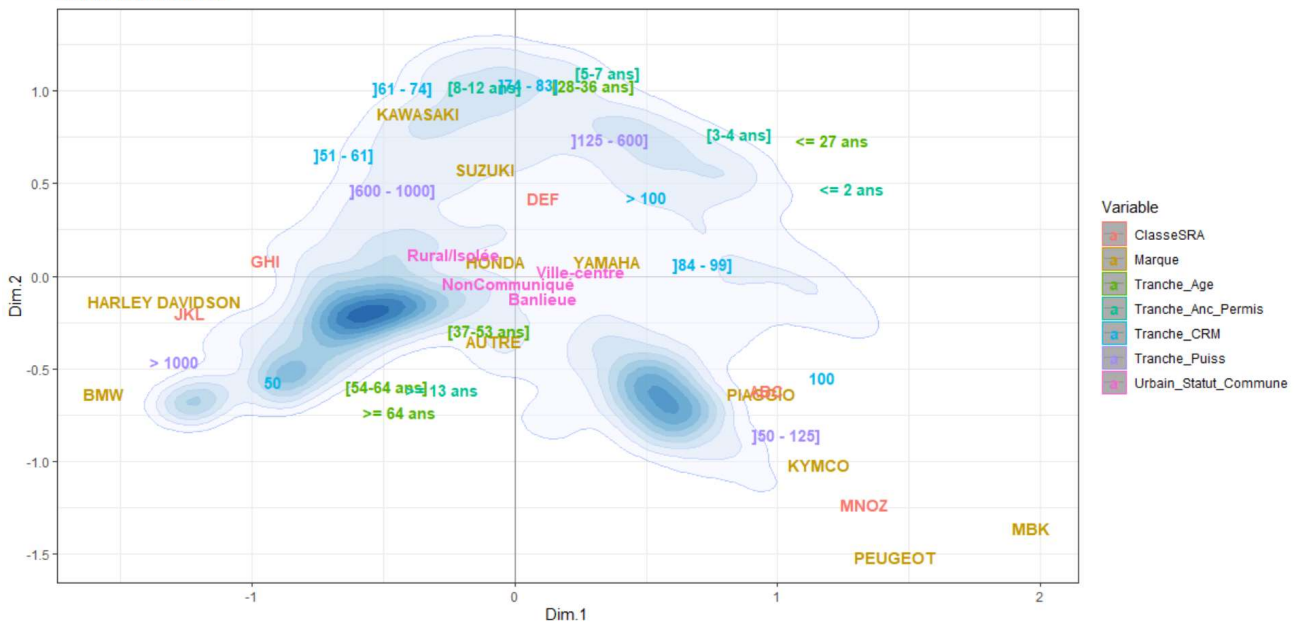




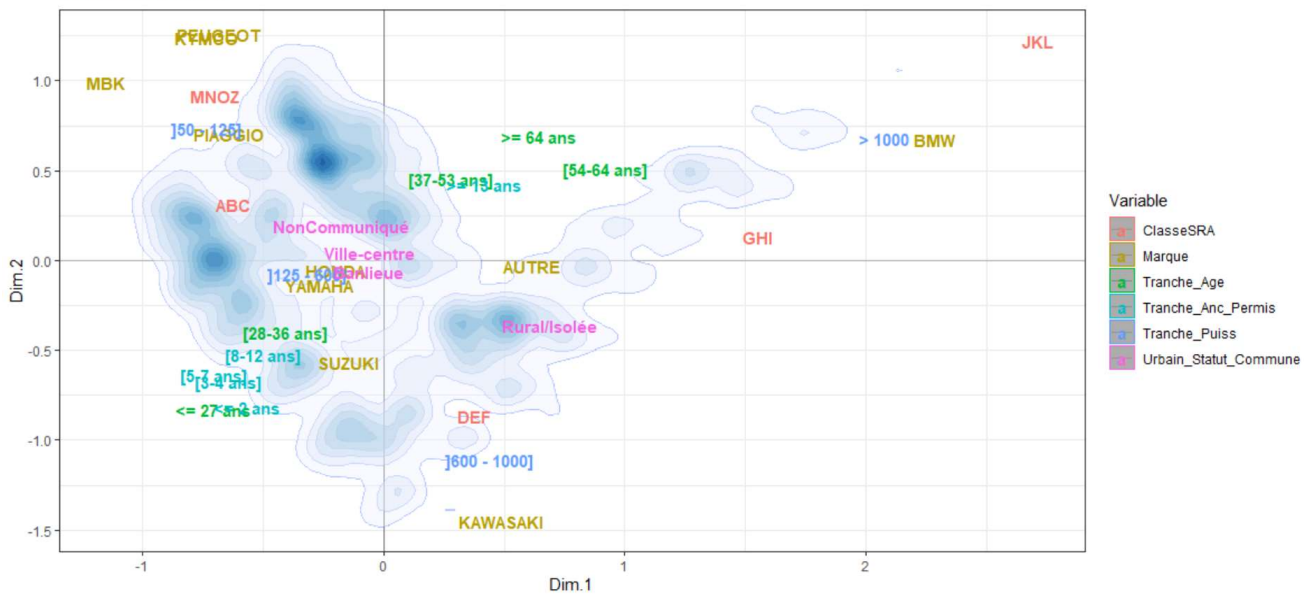
ACM Portefeuille C



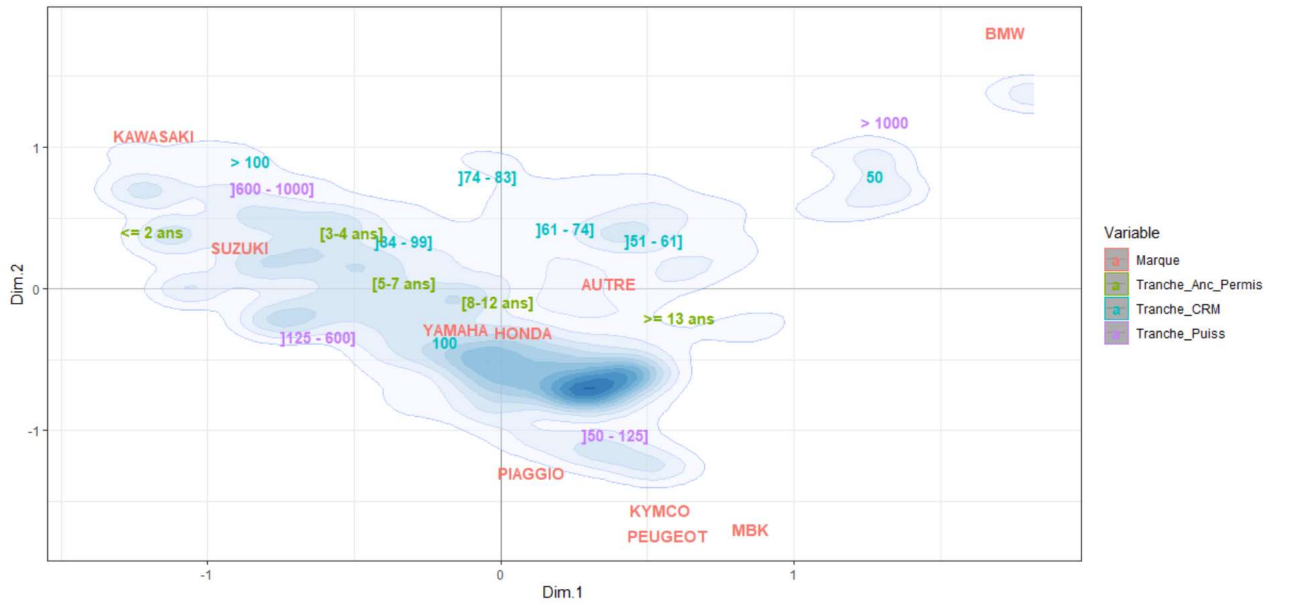
ACM Portefeuille D



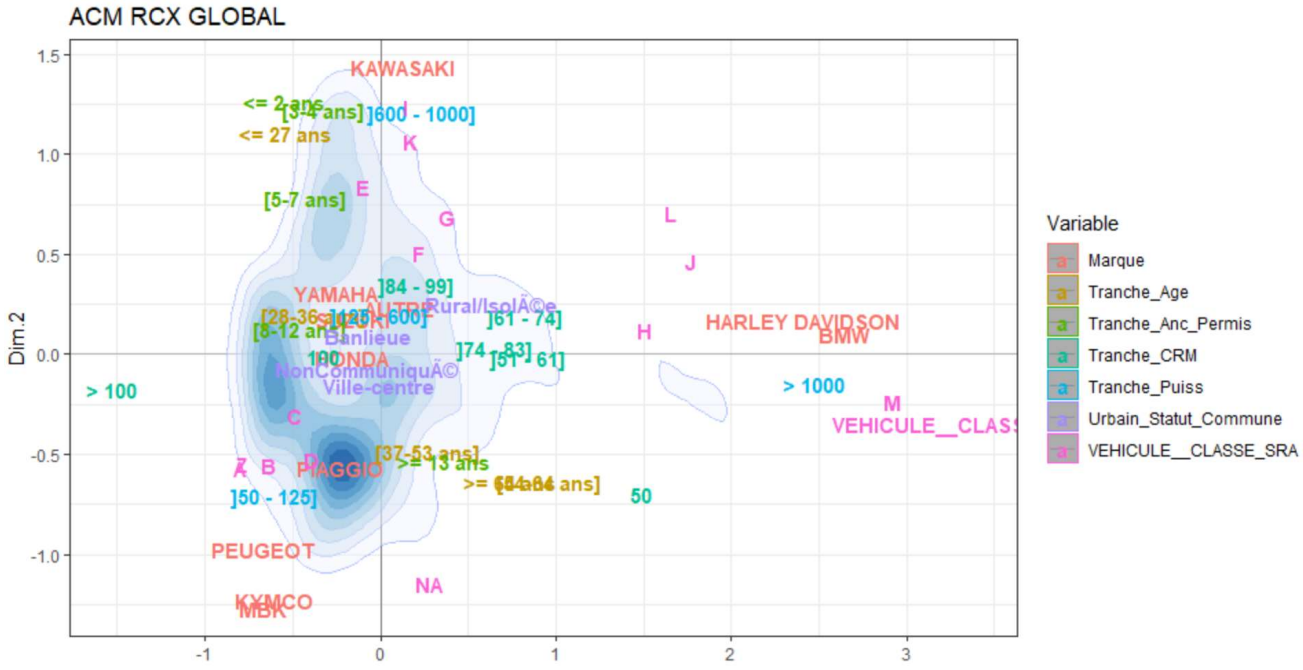
ACM Portefeuille E

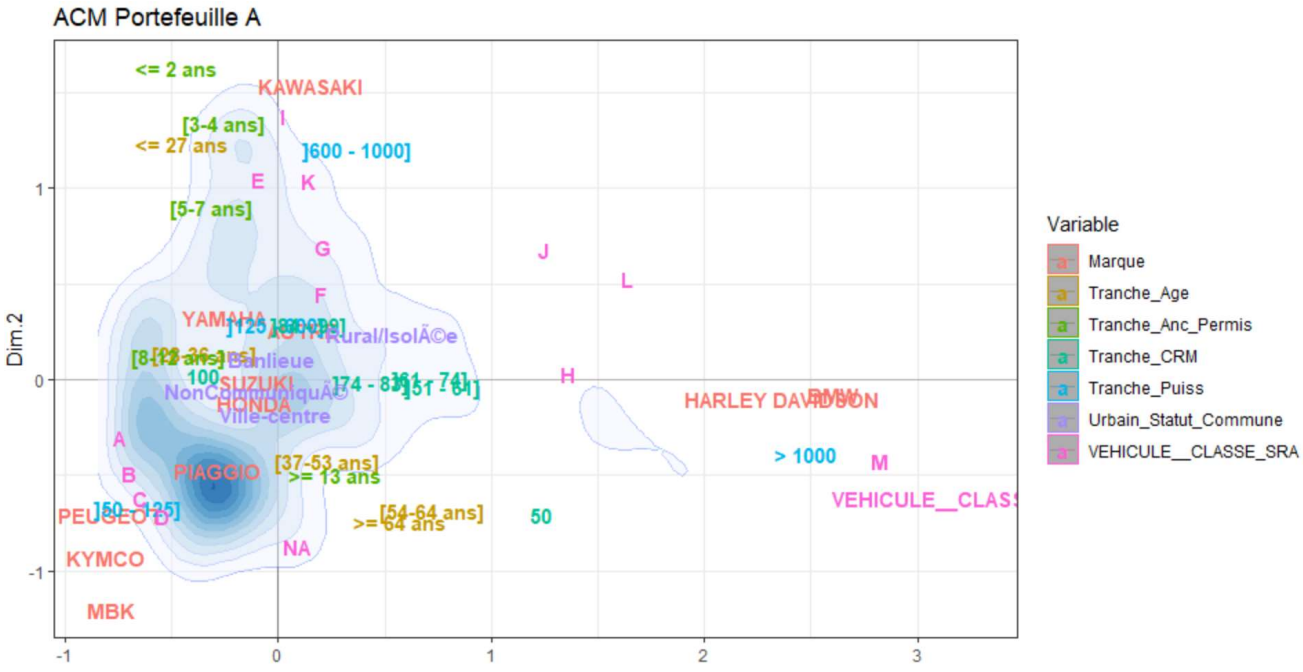


ACM Portefeuille F

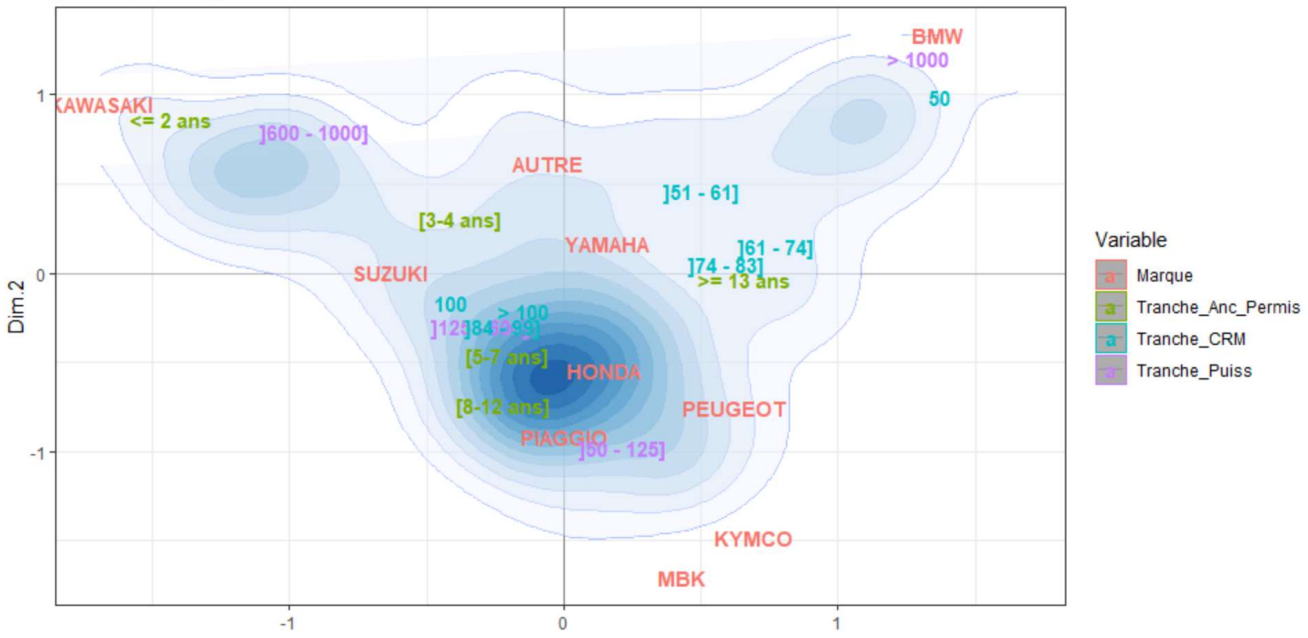


Annexe 3 : Analyse ACM sur les polices sinistrées

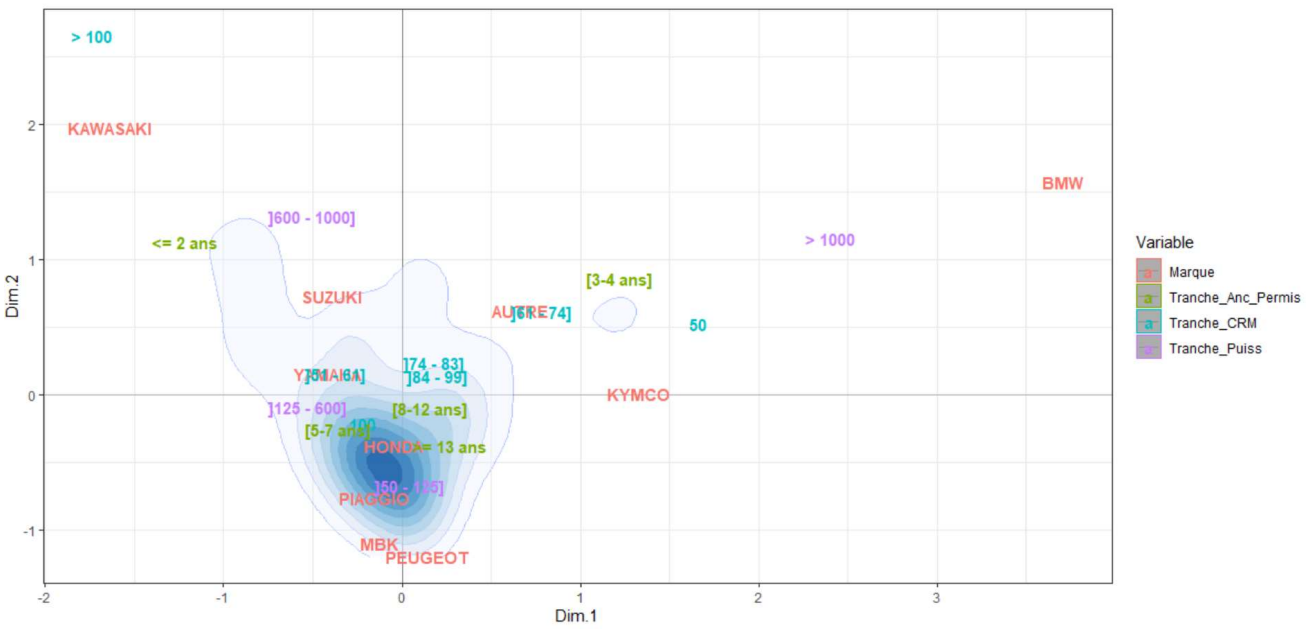




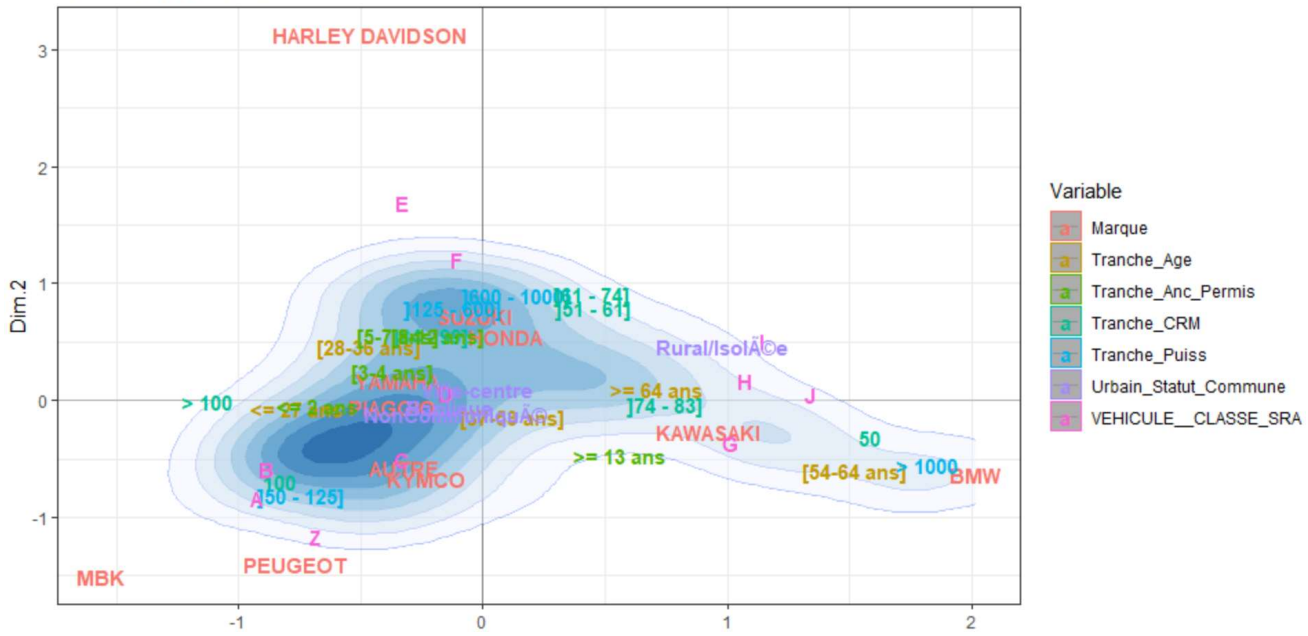
ACM Portefeuille B



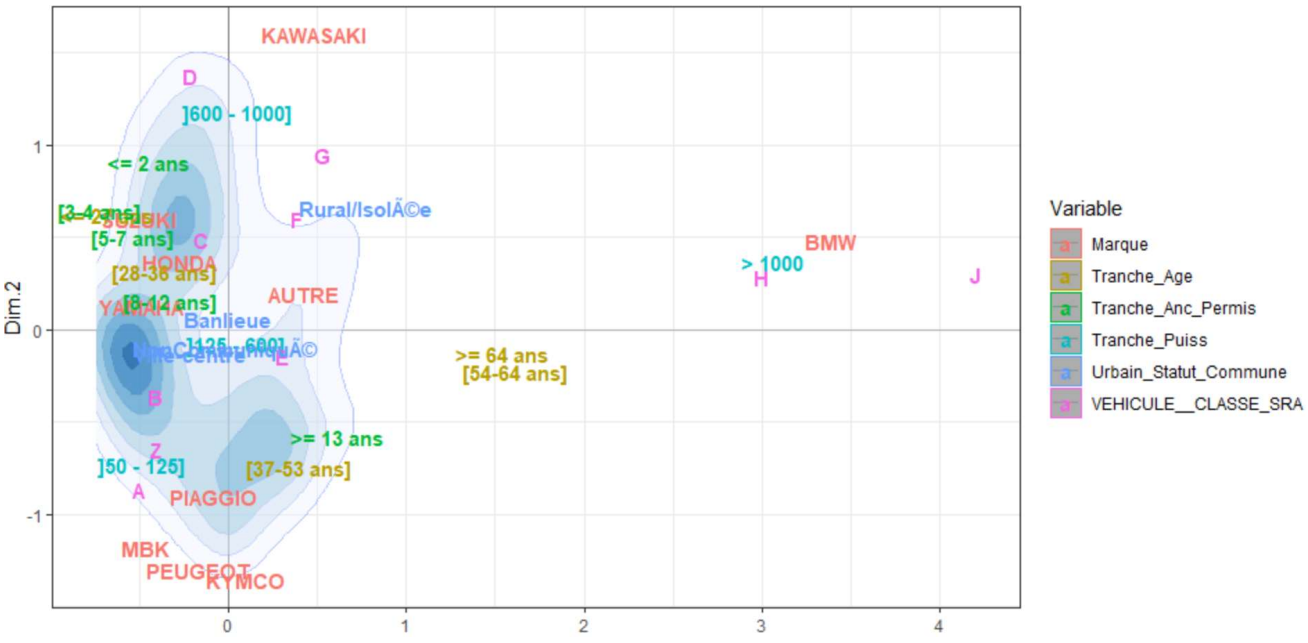
ACM Portefeuille C



ACM Portefeuille D



ACM Portefeuille E



Annexe 4 : Principales formes de la réassurance

La réassurance proportionnelle

Cette forme de réassurance permet de transférer au réassureur une proportion définie en avance des primes et des sinistres.

Il s'agit des contrats en quote-part (*quota share*) et excédent de plein (*surplus share*).

Le contrat en quote-part est la forme la plus simple où la cédante donne un pourcentage fixe de l'ensemble des primes et des sinistres dans le périmètre cédé (produit/risque/année). Les sorts de la cédante et du réassureur étant complétement alignés, ce type de contrat ne génère pas de coût supplémentaire pour l'assureur.

L'excédent de plein est une forme plus sophistiquée où la cédante transfère uniquement les risques dont les montants de garantie dépassent le niveau établi appelé plein de rétention. Pour chaque risque, le réassureur perçoit une prime correspondant à la proportion du dépassement. Le même taux est utilisé pour le partage des sinistres. Cette structure de partage permet à l'assureur de garder tous les petits risques et céder uniquement les risques qui dépassent son appétit. Du fait de cette possibilité du choix des risques par l'assureur, le contrat en excédent de plein peut théoriquement engendrer l'antisélection vis-à-vis du réassureur. L'incertitude dans le déroulement, l'exposition et le moment de règlement des sinistres corporels font en sorte que cette forme de réassurance est très rarement utilisée pour ses risques. La lourdeur des calculs est également un frein à son utilisation.

Le tableau 19 donne un exemple d'application d'un contrat en excèdent de plein selon lequel la cédante conserve les risques jusqu'à 100 000 et le réassureur accepte une tranche dans la limite de 400 000 :

Risque	Valeur totale du risque	Rétention par la cédante	Cession	Pourcentage de prime et de sinistres cédé
1	80 000	80 000	0	0%
2	120 000	100 000	20 000	20 000/120 000 = 16,7%
3	500 000	100 000	400 000	400 000/500 000 = 80%
4	700 000	100 000	400 000	400 000/700 000 = 57%

Tableau 19 - Application d'un contrat de réassurance en excédent de plein

Dans la pratique, le plan de réassurance est souvent composé de plusieurs tranches qui se complètent.

La réassurance non proportionnelle

Cette forme permet à la cédante, moyennant une prime annuelle unique, de transférer au réassureur le risque d'un ou plusieurs sinistres dépassant un certain seuil par sinistre ou en cumulé. Ce type de réassurance est souvent utilisé en assurance moto et engendre des frais importants.

Il s'agit des contrats en excédent de sinistres (*excess*) et cumul de rétention (*stop loss*).

Dans le cadre d'un contrat en excédent de sinistres, le réassureur prendra en charge la partie du sinistre limitée à un certain montant (appelé portée) au-delà de la partie conservée par la cédante (appelée priorité). Plusieurs tranches de cession peuvent exister pour répartir le risque entre un ou plusieurs réassureurs. La figure 26 présente un exemple d'application de la première tranche d'un contrat en excédent de sinistre avec une priorité à 100 000 et une portée à 400 000 sur un cas d'un sinistre à 700 000. La notation communément retenue pour cette tranche est 400XS100.

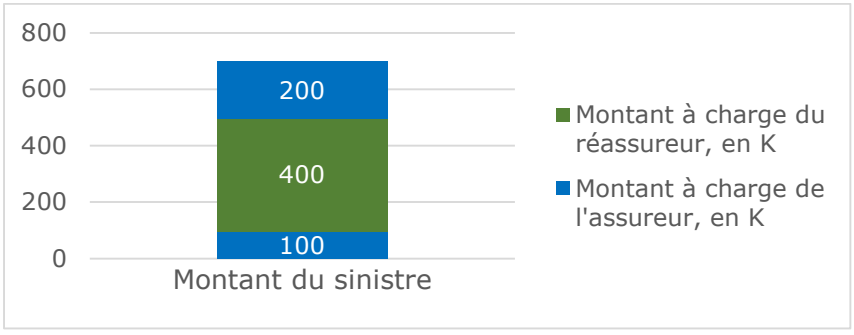


Figure 26 - Schéma d'un contrat 400XS100 en excédent de sinistre

Le tableau 20 présente la différence dans le calcul du montant de cession entre un contrat en excédent de plein et un contrat en excédent de sinistre. Il est à noter que la rétention dans le contrat en excédent de plein intervient uniquement lors de la définition de la proportion de cession, et c'est cette dernière qui est utilisée lors du calcul du montant de sinistre cédé.

Montant d'un sinistre	Calcul de cession en excédent de plein	Calcul de cession en excédent de sinistre
450 000	450 000 x 80% = 360 000	450 000 – 100 000 = 350 000

Tableau 20 - Calcul de cession en excédent de plein et en excédent de sinistre

Le contrat en cumul de rétention, quant à lui, permet de limiter la charge annuelle totale de l'assureur, souvent exprimée sous forme d'un ratio S/P maximum. Dans le cas d'une priorité de réassurance à un niveau de S/P à 100% et un ratio de S/P réalisé à 104%, le réassureur prendra en charge le montant de sinistres à la hauteur de 4% de primes de l'année d'affectation.