Mémoire présenté devant l'ENSAE Paris
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires
le 10/11/2021

Par : **Alexandre PAMBIANCHI**

Titre : **Tarification des traités en réassurance XS et comparaison des réassureurs en vision cédante**

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

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

Membres présents du jury de la filière

Nom : Wissal Sabbagh

Membres présents du jury de l'Institut des Actuaires

Entreprise : AXA Global Re

Signature :

Directeur du mémoire en entreprise :

Nom : Anne Suchel
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

Secrétariat :

Signature du candidat

Bibliothèque :

Ecole Nationale de la Statistique et de l'Administration Economique (ENSAE)
5, avenue Henry Le Chatelier - 91120 PALAISEAU, FRANCE

Résumé

L'objectif de ce mémoire est l'analyse et la tarification de traités de réassurance en excédent de sinistre avec prise en compte des cotations des réassureurs. Cette analyse s'effectue sur deux échelles de temps : une pré-cotation et une post-cotation.

Une cotation est une prime commerciale estimée par un réassureur pour un traité en XS. Celle-ci est envoyée à la cédante lorsqu'un traité est mis sur le marché de la réassurance. Cette information lui permet d'optimiser ses structures de réassurances et de confronter la vision interne du traité à celle externe. En effet, une cotation d'un traité reflète la vision du risque que le réassureur se fait du traité. Dans ce mémoire, ces cotations sont utilisées en premier lieu pour élaborer un outil de tarification.

Basé sur le principe de prime par écart type, cet outil est composé de plusieurs sous-modèles présentant chacun leurs avantages et leurs inconvénients. L'objectif principal est dans un premier temps d'estimer le coefficient de chargement β de l'écart type utilisé dans la formule de tarification. Pour ce faire, nous utilisons la cotation moyenne par traité. Cette information nous permet de construire deux grands types de modèles. L'un donnant une formule générale pour des groupes de traités et l'autre estimant directement le coefficient β de chaque traité. Ces différents modèles permettent donc à la cédante d'obtenir plus d'informations sur sa réassurance notamment en estimant la prime commerciale moyenne cotée d'un nouveau traité. Ainsi, cette tarification se place naturellement en pré-cotation.

La méthode par chargement constant utilise un algorithme de minimisation qui estime un $\hat{\beta}^*$ optimal pour un nombre de traités défini. Ceci permet d'obtenir une formule unique pour un nombre important de traités. Plusieurs façons de grouper les traités sont utilisées. Elles sont soit basées sur une analyse de la distribution de β soit sur des algorithmes de clustering. Elles offrent ainsi l'avantage d'être facilement interprétables mais se révèlent, dans certains cas, sous-performantes dans leurs estimations.

La méthode par prédiction directe utilise quant à elle des méthodes de prédiction usuelles comme le GLM et les forêts aléatoires. Le modèle des forêts aléatoires est celui offrant les meilleurs résultats avec une estimation satisfaisante de la cotation moyenne. Cependant, son interprétabilité est moins aisée que dans la méthode par β constant.

Enfin, nous utilisons les cotations pour établir des comparatifs entre réassureurs. Cette analyse est donc effectuée sur la période post-cotation. Pour ce faire, nous comparons les réassureurs à l'aide de plusieurs critères comme les ordres de grandeur de leurs cotations, leur nombre de traités ou encore leur solvabilité. Cela nous permet donc de déceler certains comportements de nos partenaires de réassurance en fonction des traités cotés.

Mots-clés : *cotation, réassurance en XS, machine learning, tarification, comparaison de réassureurs*

Abstract

The objective of this paper is the analysis and pricing of excess of loss reinsurance treaties, taking into account the reinsurers' quotations. This analysis is realized on two time scales : a pre-quotation and a post-quotation.

A quotation is a commercial premium estimated by a reinsurer for a XoL treaty. It is sent to the ceding company when a treaty is placed on the reinsurance market. This information allows it to optimize its reinsurance structures and to compare the internal vision of the treaty with the external one. Indeed, a quotation reflects the reinsurer's view of the treaty's risk. In this paper, these quotations are used primarily to develop a pricing tool.

Based on the formula of premium per standard deviation, this tool is composed of several sub-models, each with its own advantages and disadvantages. The main objective is first to estimate the loading factor β of the standard deviation used in the pricing formula. To do so, we use the average quotation per treaty. This information allows us to build two main types of models. One gives a unique formula for groups of treaties and the other directly estimates the β coefficient for each treaty. These different models allow the ceding company to obtain more information on its reinsurance, especially by estimating the average commercial premium quoted for a new treaty. Thus, this pricing is naturally placed in the pre-quotation phase.

The constant loading method uses a minimization algorithm that estimates an optimal $\hat{\beta}^*$ for a defined number of treaties. This allows to obtain a closed formula for many treaties. Several ways of grouping the treaties are used. They are either based on an analysis of the β distribution or on clustering algorithms. They offer the advantage of being easily interpretable but, in some cases, they are underperforming in their estimations.

The direct prediction method uses usual prediction methods such as GLM and random forests. The random forest model offers the best results with a correct estimation of the average quotation. However, its interpretability is less easy than in the constant β method.

Finally, we use the quotations to establish comparisons between reinsurers. This analysis is carried out over the post-quotation period. To do so, we compare reinsurers using several criteria such as the size of their quotes, their number of treaties or their solvency. This allows us to detect some behaviors of our reinsurance partners in accordance with the quoted treaties.

Keywords : *quotations, XoL reinsurance, machine learning, pricing, comparing reinsurers*

Remerciements

Je tiens tout d'abord à remercier le groupe AXA Global Re et particulièrement le service Analytics and Pricing pour m'avoir accueilli en stage pendant ces six mois.

Je remercie également et surtout ma tutrice en entreprise, Madame Anne SUCHEL, actuaire, pour son suivi, sa disponibilité et ses conseils de très grande qualité tout au long de mon stage.

Mes remerciements s'adressent aussi à Monsieur Jérôme CRETIEN, manager actuaire, pour son aide, ses enseignements et son expertise au cours de mon stage et plus globalement durant mon parcours universitaire.

Je tiens également à remercier le corps enseignant de l'ENSAE pour m'avoir permis d'acquérir l'ensemble des connaissances nécessaires à la réussite de ma formation et d'avoir partager leurs savoirs de manière approfondie avec une grande pédagogie. Un merci particulier à ma réfèrente pédagogique Madame Wissal SABBAGH pour son suivi.

Enfin, un grand merci à ma famille et à tous mes proches et connaissances professionnelles pour avoir su me soutenir et me motiver tout au long de mes années d'études supérieures.

Grâce à vous tous, j'ai pu concrétiser mon projet professionnel et parvenir à la rédaction de ce mémoire qui marque la fin de mon parcours scolaire et le début de ma vie active.

Table des matières

Introduction

La réassurance est un outil permettant à un assureur, nommé la cédante, d'être plus robuste dans la gestion de ses risques. Ce domaine d'activité est aujourd'hui un acteur très important du marché de l'assurance avec une influence croissante. Plusieurs facteurs comme la garantie de faire face à des sinistres très élevés, une aide à la cédante pour souscrire davantage d'affaires en lui offrant plus de capacité ou encore le potentiel lissage des portefeuilles expliquent en partie cet intérêt pour la réassurance. Le partenariat entre un réassureur et une cédante se matérialise par un traité ou un contrat.

Pour un traité (ou contrat) de réassurance il n'existe en général qu'une seule cédante mais plusieurs réassureurs. Ainsi, lorsque l'assureur émet un traité de réassurance, celui-ci est coté par les réassureurs. La cotation est un procédé où un réassureur estime le risque du traité. Ce risque se matérialise par une prime dite commerciale car elle correspond à la prime commerciale proposée à la cédante. Ainsi, lorsque le réassureur envoie un tarif à la cédante nous parlons d'une cotation. Lors de ce processus, la cédante reçoit donc un nombre important de cotations pour plusieurs traités de la part de ses partenaires de réassurance.

Cette information permet à la cédante de visualiser le risque que les réassureurs envisagent sur ses traités. Ce mémoire vise ainsi à utiliser ces cotations pour la réalisation d'études permettant à la cédante d'obtenir plus de transparence sur sa réassurance. Les cotations permettent de compléter les données internes de la cédante par l'apport d'une vision externe, celle des réassureurs. Plusieurs axes d'études sont alors possibles pour exploiter au mieux les cotations. Nous nous concentrons ici sur deux axes principaux : une tarification des traités XS et une comparaison des réassureurs sur ces mêmes traités.

En pratique, lorsque qu'une cédante souhaite tarifier elle-même ses traités, dans le but de mieux connaître le risque sous-jacent de ses portefeuilles, elle peut opter pour une modélisation de ses sinistres historiques afin de simuler un grand nombre d'années de sinistralité. Par la suite, elle peut appliquer les simulations réalisées à ses traités. Ceci lui permettra donc d'estimer en moyenne le montant de récupération d'un traité. Une récupération est le montant total annuel de sinistre pris en compte dans le traité c'est à dire le montant que devra payer le réassureur, en fonction des éventuelles reconstitutions ou AAL par exemple. Aussi, il est possible de calculer plusieurs indicateurs comme la volatilité de ces récupérations ou encore la fréquence de reconstitution. Grâce à cela, la cédante est dans la capacité d'estimer une prime pure de ses traités. En formulant plusieurs hypothèses, elle peut même être en mesure de calculer une prime commerciale. Cependant, cette vision du risque est uniquement interne et ne prend pas en compte un facteur essentiel lorsque les réassureurs cotent : la diversification.

En effet, un réassureur possède en réalité un nombre important de traités lui permettant de diversifier ses risques. De plus, en excédant de sinistre, il est confronté à une probabilité forte d'être dans l'obligation d'intervenir dans de très grands montants de sinistre. Ainsi, il éprouve donc le besoin primordial de diversifier ses traités. Cette diversification peut s'effectuer par une souscription de traités dans différents pays afin de ne pas limiter son exposition à une seule localisation géographique mais également par une sélection de diverses branches avec des traités disposant de capacités variées. Tous ces facteurs expliquent pourquoi la vision d'un réassureur d'un traité XS peut être en réalité moins risquée que celle de la cédante. L'assureur modélisant uniquement son traité, il ne dispose pas du facteur de diversification du réassureur pour tarifer. Ainsi, lorsque la cédante tente d'estimer la prime commerciale qui sera cotée par les réassureurs sur le marché de la réassurance, cette information lui est manquante. Ce mémoire propose donc

d'apporter l'information de ce facteur de diversification dans la tarification interne de la prime commerciale cotée par les réassureurs. Pour ce faire, nous utilisons la cotation moyenne par traité afin de comparer la vision externe moyenne des réassureurs avec notre vision interne en tant que cédante.

Ce modèle de tarification est basé sur le principe de prime par écart type. Le coefficient de chargement de l'écart type, noté β, est la variable d'intérêt des modèles développés par la suite. Le but est donc d'estimer β afin d'obtenir une formule de tarification pour chaque traité. Pour se faire, nous définissons deux grandes méthodes différentes : une basée sur une formule générale pour un grand nombre de traités et la seconde sur une estimation traité par traité de β.

La méthode permettant d'obtenir une formule générale est nommée méthode par chargement constant puisque β est fixé pour un ensemble de traités définis. Ce procédé permet ainsi à la cédante d'obtenir une formule de tarification commune à plusieurs traités. De plus, une façon de créer ces groupes de traités est d'utiliser des algorithmes de *clustering* qui offrent l'avantage d'identifier des ensembles de traités homogènes. Ainsi, la cédante peut à la fois regrouper ses traités et observer leurs caractéristiques communes et en même temps connaître la vision externe de ses traités grâce à sa formule de tarification de la cotation moyenne. La seconde méthode de tarification estime directement le coefficient de chargement β par traité. Elle permet ainsi une tarification plus précise et plus performante mais elle souffre cependant d'un manque de transparence induit par la nature même des algorithmes utilisés.

Ainsi, la cédante est donc en capacité de tarifier ses traités en estimant directement la cotation des réassureurs. De plus, l'information de la cotation permet aussi de fixer des critères de comparaison entre réassureurs. Ce mémoire propose donc, en dernière partie, une autre manière d'analyser les cotations par un comparatif entre réassureurs basé sur leurs cotations moyennes, leur tendance à proposer des primes plus ou moins importantes, leur notation Standard & Poor's ou encore leur part moyenne proposée par traité. Tous ces indicateurs nous permettent ainsi d'analyser l'existence de certaines tendances entre nos partenaires de réassurance.

1 Fonctionnement et objectifs de la réassurance

Table des matières

1.1 Définition de la réassurance

La réassurance est couramment nommée *l'assurance de l'assureur*. Plus techniquement, la réassurance est une opération par laquelle l'assureur (la cédante) transmet tout ou une partie d'un ou plusieurs risques à un réassureur moyennant une prime. Cette opération est matérialisée par un traité de réassurance ou par un contrat dans le cas des réassurances facultatives. Nous pouvons voir la réassurance comme un partage de sort entre l'assureur et le réassureur. Cependant, celui-ci se réalise *verticalement* et non *horizontalement* (comme la co-assurance). Ainsi, la réassurance permet la division des risques sans pour autant diviser le contrat de l'assuré. Par ailleurs, le réassureur n'est pas responsable des engagements que la cédante a envers ses assurés.

1.2 Histoire de la réassurance

Historiquement, la création de l'assurance en 1666 après le grand incendie de Londres, marqua un changement dans la façon d'aborder les divers risques de la vie. En s'assurant, il est devenu possible de *transformer* un risque individuel en risque collectif. Par ailleurs, en 1681, la *grande ordonnance de la marine* de Colbert permit aux assureurs de se faire assurer par d'autres assureurs à travers des syndicats. La réassurance moderne fit son apparition en Allemagne en 1842. Pays fortement industrialisé et très touché par les incendies de plus en plus fréquents et onéreux, les assureurs eurent de plus en plus de mal à honorer leurs engagements. Ainsi, le 22 décembre 1842, suite à l'incendie de Hambourg, l'idée de créer une société de réassurance fut émise. Finalement, en 1846, Cologne Re (Kölnische Rückversicherungs-Gesellschaft) fut fondée (aujourd'hui Gen Re). Son premier traité fut souscrit en 1852. Les techniques de réassurance s'améliorèrent alors au fur et à mesure du temps.

1.3 Rôles de la réassurance

En se réassurant, la cédante peut bénéficier de plusieurs types de soutiens en fonction de ses besoins. Tout d'abord, dans le but de répondre aux exigences de la réglementation Solvabilité 2, la réassurance permet à la cédante de mieux maîtriser sa probabilité de ruine, et ce, par plusieurs mécanismes, comme par exemple, en transférant des portefeuilles risqués (volatiles) et/ou mal connus. Aussi lors du lancement d'un produit d'assurance, le besoin en fonds propres peut être important. Néanmoins, la compagnie d'assurance ne dispose pas nécessairement de

fonds propres suffisants (ou ne veut pas allouer une partie de ses fonds propres) pour le démarrage de son produit. Dans ce cas, la réassurance permet de compenser cette absence de fonds propres (par exemple avec une réassurance de type quote-part, avec un taux de cession élevé) à travers une forme d'assistance technique et de partenariat. Également, si la cédante possède un portefeuille de petite taille, la volatilité de la fréquence de sinistre peut être élevée. Parallèlement, si le portefeuille est grand, la fréquence est bien estimée mais la probabilité de survenance d'un sinistre de sévérité forte peut être importante. Dans ce cadre, le réassureur peut intervenir pour lisser les résultats de ces portefeuilles. Finalement, dans un contexte de sinistres élevés demandant une forte capacité, le réassureur peut proposer d'en prendre en charge une grande partie. Dans le cas de catastrophes naturelles (tremblements de terre, tempêtes, tsunamis, ...), il est aussi possible d'utiliser la réassurance pour mutualiser ces risques à l'échelle mondiale. Il existe deux types de réassurance pouvant chacune prendre quatre formes.

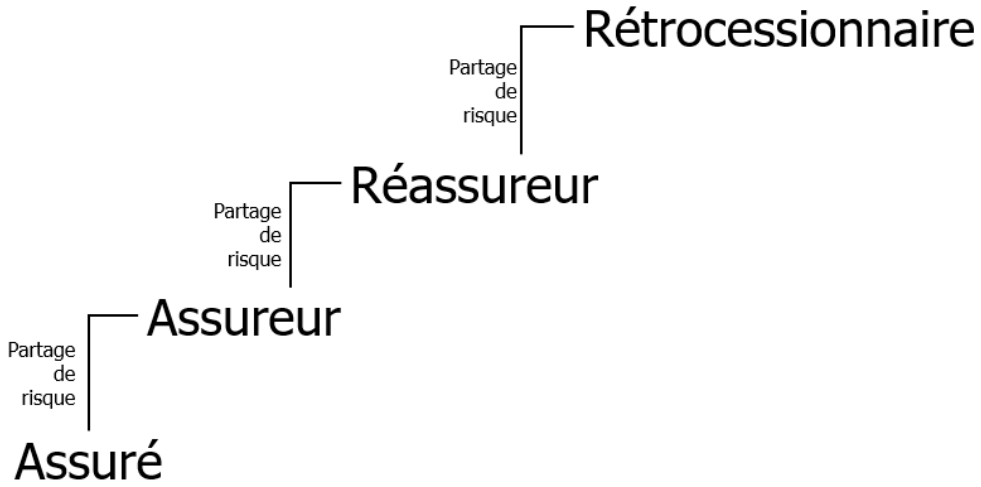


FIGURE 1 – Hiérarchie des transferts de risque

Finalement, le réassureur peut lui aussi céder ses risques à un rétrocessionnaire.

1.4 Types et formes de réassurance

Juridiquement, une réassurance facultative est matérialisée par un contrat et se fait risque par risque. Elle est dite facultative car la cédante et le réassureur ne sont pas dans l'obligation de faire intervenir le contrat de réassurance. Plus précisément, la cédante n'a pas pour obligation de céder et le réassureur d'accepter les risques. D'un autre côté, la réassurance obligatoire se matérialise par un traité et se fait sur un portefeuille. Celle-ci porte sur une période déterminée avec cession obligatoire des risques concernés par le traité pour la cédante et acceptation obligatoire des risques pour le réassureur.

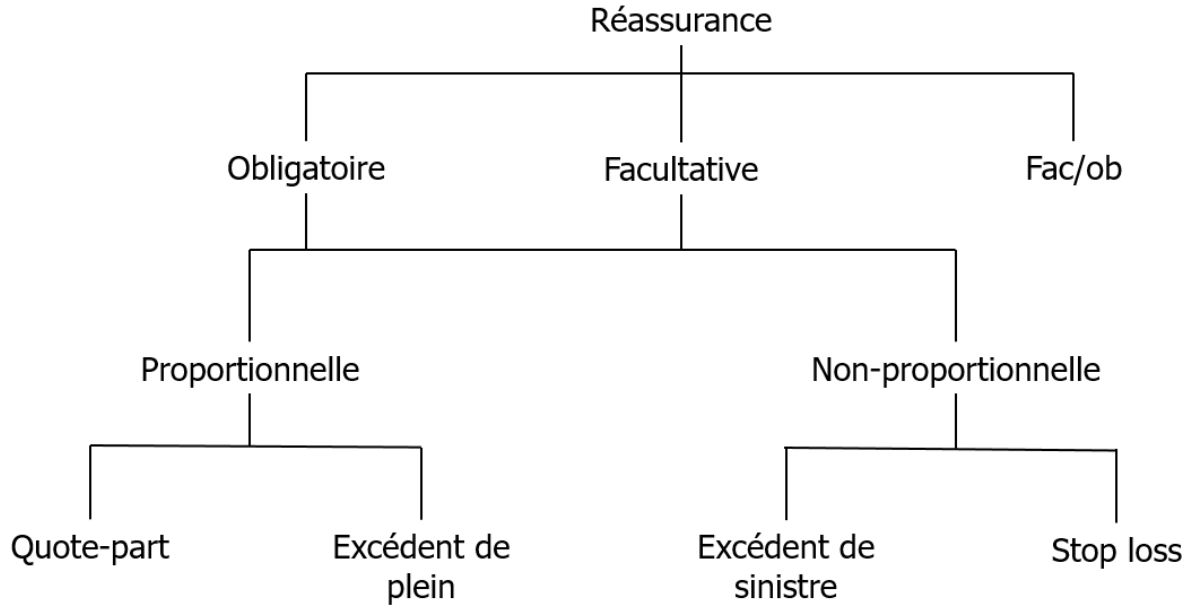


FIGURE 2 – Types et formes de réassurance

Plusieurs formes de réassurance sont possibles, elles se distinguent par leur caractère proportionnel (ou non). Ainsi, elles sont logiquement nommées réassurances proportionnelles (basées sur le risque) et non proportionnelles (basées sur le sinistre).

1.4.1 Réassurance proportionnelle

Pour ce type de réassurance, cédant et réassureur partagent :

— Les sommes assurées
— Les primes
— Les sinistres

Le pourcentage de partage définit la part cédée par la cédante. De plus, un taux de commission permettant à la cédante de payer ses divers frais est réglé par le réassureur.

1.4.1.1 Quote-part

La réassurance en quote-part (ou *quota-share* en anglais), notée QP, est définie par un taux de cession qui indique quel pourcentage de prime et sinistre l'assureur cède au réassureur. La réassurance en quote-part nécessite un portefeuille homogène où les risques sont du même type et les sommes assurées d'un même ordre de grandeur. Notons c le taux de cession (où $0 \leq c \leq 1$), P la prime annuelle d'assurance et S le montant total annuel de sinistre alors :

— $P \times c$ est la prime cédée par l'assureur au réassureur.
— $S \times c$ est le montant de sinistre cédé par l'assureur au réassureur.
— $P \times (1 - c)$ est la prime conservée par l'assureur.
— $S \times (1 - c)$ est le montant de sinistre conservé par l'assureur.
— $(P - S) \times (1 - c)$ est le résultat de l'assureur.
— $(P - S) \times c$ est le résultat du réassureur.

Par exemple, considérons un traité de réassurance en QP où le taux de cession est de 40 %.

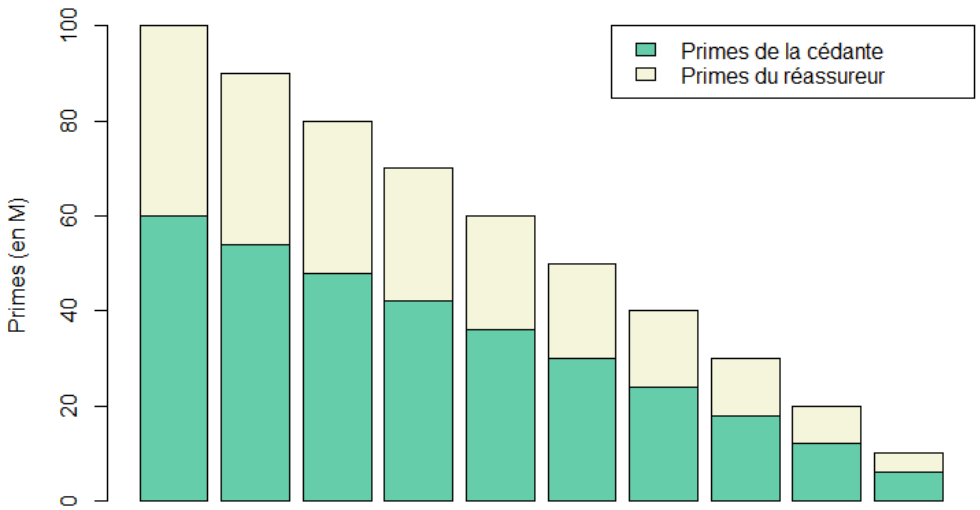


FIGURE 3 – Exemple de réassurance en QP avec taux de cession à 40 %

Prenons un cas fictif où $P = 200$ et $S = 100$. Sans intervention de la réassurance (communément appelée *à 100 %*), le résultat annuel brut est de 100. En appliquant la cession de 40 % au réassureur (80 de sinistres et 40 de primes), l'assureur conserve alors 60 % de son portefeuille, soit un résultat total de 60 après cession au réassureur.

	Brut	**Réassurance**	**Net de réassurance**
Primes	200	80	120
Sinistres	100	40	60
Résultat N	100	40	60

TABLE 1 – Exemple de réassurance en QP avec un taux de cession à 40 %

Cette forme de réassurance peut permettre de protéger les fonds propres de l'assureur ou simplement d'effectuer un partage d'affaires dans un cadre plus commercial. Elle peut aussi aider l'assureur lorsque la variabilité des sinistres est trop aléatoire (mal estimée). Dans ces cas, la cession peut avoir un impact fort sur le résultat de l'assureur. Considérons le même portefeuille que précédemment dont la prime annuelle s'élève à 200. Cependant, l'année suivante est bien plus sinistrée avec un total de 228 en montant de sinistre (au lieu de 60 de l'an passé). Supposons que la quantité de fonds propres détenue est de 30. De ce fait, le résultat sera le suivant :

Fonds propres au 31/12/N : 30

	Brut
Primes	200
Sinistres	228
Résultat N+1	-28

Fonds propres au 31/12/N+1 : 2

TABLE 2 – Exemple d'impact sur les fonds propres sans réassurance en QP

Ainsi, nous remarquons que la quasi-intégralité des fonds propres a été consommée en une année. Cependant, en utilisant la réassurance en QP 40 % nous obtenons le résultat suivant :

Fonds propres au 31/12/N : 30

	Brut	**Réassurance**	**Net de réassurance**
Primes	200	80	120
Sinistres	228	91.2	136.8
Résultat N+1	-28	-11.2	-16.8

Fonds propres au 31/12/N+1 : 13.2

TABLE 3 – Exemple d'impact sur les fonds propres avec réassurance en QP 40%

De la sorte, avec cette réassurance, une partie non négligeable des fonds propres (13.2 au lieu de 2 sans quote-part) a été conservée améliorant ainsi la marge de solvabilité de la compagnie. Néanmoins, ce type de réassurance ne modifie pas le *Loss Ratio* et est *simple* à organiser avec un réel partage du sort entre réassureur et cédante, qui suivent tous deux (*modulo* le taux de cession souscrit) le même résultat en fonction du portefeuille. Malgré cela, même après application de la réassurance, le portefeuille réassuré reste hétérogène s'il l'était avant la souscription de la quote-part. De plus, il n'y a pas de lissage du résultat dans le temps et parfois le cumul de primes transférées au réassureur peut s'avérer (trop) élevé.

1.4.1.2 Excédent de plein

La réassurance en excédent de plein (ou *surplus share*) est la deuxième forme de réassurance proportionnelle. Celle-ci est similaire à de la réassurance en quote-part car elle se base aussi sur un taux de cession. Cependant, celui-ci n'est pas identique pour tous les risques réassurés. En effet, pour chaque risque, un montant réassuré est calculé en fonction du plein de rétention et de la capacité.

Définitions

- **Plein de rétention/conservation (ou *line*)** : montant fixe du risque retenu par l'assureur. Le réassureur n'intervient qu'au-dessus de ce montant.
- **Capacité (ou *limit*)** : montant maximal assurable par le réassureur, souvent exprimé en nombre de plein de rétention.
- **Plafond (ou *ceiling*)** : montant au-dessus duquel l'assureur prend tout en charge, égal au plein de rétention plus la capacité.

Ainsi, chaque montant réassuré est exprimé en pourcentage de la somme assurée du risque. Plus techniquement, en posant i le i$^{\text{ème}}$ risque, c_i et SA_i respectivement le taux de cession et la somme assurée du risque i, R le plein de rétention et C la capacité :

$$c_i = \frac{\min(C, \max(SA_i - R, 0))}{SA_i}$$

Graphiquement, nous pouvons visualiser cette réassurance comme ceci :

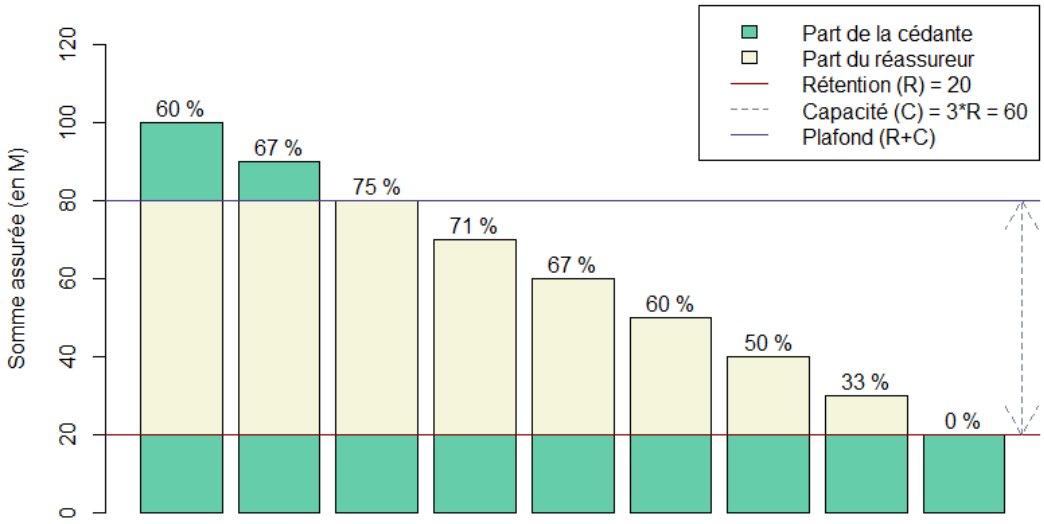


FIGURE 4 – Exemple de réassurance en excédent de plein

Interprétation de la figure : dans le cas où la somme assurée est de 100, le réassureur intervient sur le montant au-dessus de R (soit 20) dans la limite de la capacité C (3 pleins donc 60). Ainsi, le montant réassuré est égal à C car la somme assurée est supérieure au plafond (égal à 80). Le taux de cession associé à ce risque est donc de $60/100 = 60$ *%. Concernant le risque où la somme assurée est de 70, le réassureur intervient sur 70-20=50. En effet, le plafond n'étant pas atteint, le réassureur prend en charge la totalité du risque au-dessus du plein de rétention, le taux de cession est donc de* $50/70$ *soit environ 71 %. Enfin, dans le dernier cas où la somme assurée est de 20, la cédante conserve l'intégralité du risque (car R = 20 = somme assurée).*

Plus généralement, plusieurs cas sont possibles. Si R est grand, la majorité des petits portefeuilles ne seront pas cédés tandis que dans le cas contraire, la majorité des portefeuilles seront réassurés (si la capacité est élevée). Cette forme de réassurance a donc pour avantage d'être plus *flexible* que la quote-part. La cédante peut conserver les risques les plus faibles, généralement mieux connus et donc moins risqués, puis céder une plus ou moins grande partie des risques élevés, en fonction de son appétit au risque (et celui du réassureur) qui sont, eux, moins bien maîtrisés, car généralement plus volatiles et moins fréquents.

1.4.2 Réassurance non proportionnelle

Cette forme de réassurance, basée sur le sinistre ou l'évènement, ne repose pas sur un partage du sort complètement proportionnel entre réassureur et assureur. Elle est dite non proportionnelle, car le réassureur définit un montant maximum (portée ou limite) qu'il souhaite réassurer et un minimum avant son intervention (priorité) moyennant une prime annuelle fixe, peu importe le(s) sinistre(s) survenu(s).

Définitions

— **Portée/limite (ou *limit*)** : montant maximal assurable par le réassureur.
— **Priorité (ou *priority/deductible*)** : seuil à partir duquel intervient la réassurance.

Ainsi, la cédante peut verser une prime de réassurance sans pour autant qu'un sinistre ne soit couvert par le traité souscrit. Parallèlement, le réassureur peut se voir contraint de payer plus de sinistres que ne couvre sa prime. L'enjeu de ce type de réassurance repose donc sur les modélisations et négociations autour de la prime. La prime annuelle est calculée grâce au taux de prime de réassurance. De ce fait, elle est le produit du taux de prime et de l'EPI (pour *Estimated Premium Income*, la prime annuelle estimée).

1.4.2.1 Excédent de sinistre

La réassurance en excédent de sinistre, définit par évènement ou par sinistre (ou tête), aussi appelée XS (*XL* en anglais), est déterminée par sa limite et sa priorité. Elle est notée *limite* XS *priorité*.

Cette forme de réassurance permet à la cédante d'écrêter ses risques de pointe (les hauts sinistres) et d'obtenir une sinistralité plus homogène, plus lisse. Lorsqu'elle est définie par sinistre, seul le montant du sinistre importe pour le calcul de la charge du réassureur. Cependant, dans les évènements *catastrophes*, la cédante n'est pas nécessairement protégée si elle utilise un XS par sinistre. En effet, il est possible qu'un sinistre catastrophique comme la grêle ou une tempête touche plusieurs lieux sans pour autant que chaque sinistre survenu soit assez grand pour atteindre la priorité. Alors, en cumulant l'ensemble des sinistres de la catastrophe, la cédante peut se trouver en grande difficulté. Dans ce cas, elle peut définir un traité par événement, permettant de définir la somme des sinistres dus à l'évènement comme un seul montant de sinistre. Ceci permet de faire intervenir la réassurance bien plus efficacement. Pour calculer la charge du réassureur, par sinistre, nous pouvons utiliser la formule suivante :

$$S_R = \min(L, \max(S_A - P, 0))$$

où L est la limite (ou portée), P la priorité, S_A le montant de sinistre à charge de l'assureur (aussi appelé montant FGU pour *From the Ground Up*) et S_R le montant de sinistre réassuré. Ainsi, il ne reste qu'à la charge de l'assureur le montant $S_A - S_R$.

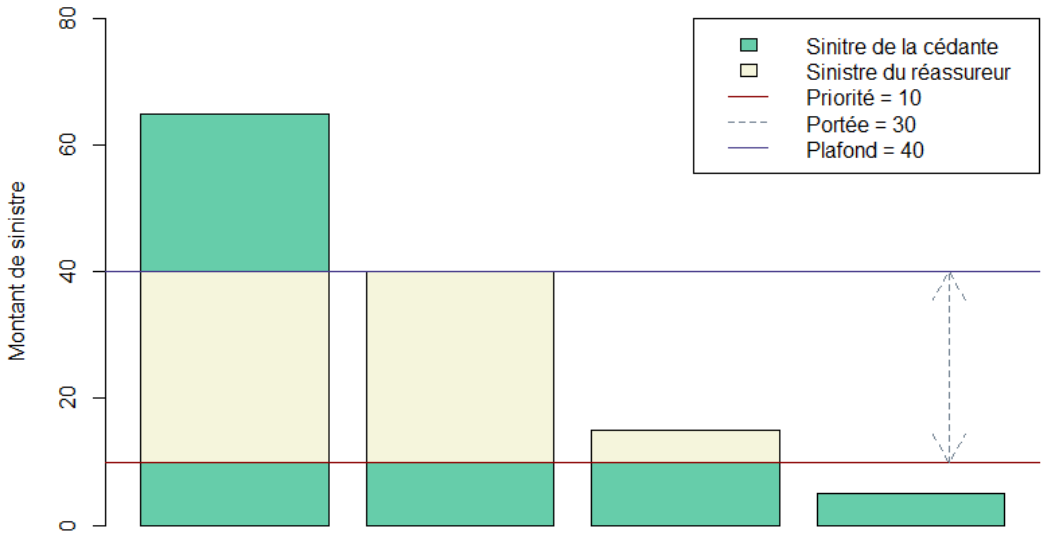


FIGURE 5 – Exemple de réassurance en excédent de sinistre 30 XS 10

Interprétation de la figure : Dans cet exemple, le traité est de la forme 30 XS 10. Pour le premier sinistre égal à 65, le réassureur prend en charge 30 tandis que la cédante doit payer les 35 restants à charge. Le deuxième sinistre atteint le plafond, le réassureur paie donc 30 tandis que la cédante paie 10. Le troisième sinistre de 15 ne dépasse la priorité que de 5, somme à régler par le réassureur. Finalement, le dernier sinistre est trop faible pour être réassuré.

Par ailleurs, il est possible qu'un traité cumule des tranches pour couvrir plusieurs ordres de grandeur de sinistre. En pratique ce cas est même fréquent car il permet une optimisation du tarif de réassurance. Parfois, sont spécifiés un AAL (*Annual Aggregate Limit*) et/ou un AAD (*Annual Aggregate Deductible*). L'AAL est le montant total de sinistre réassurable (même fonctionnement qu'un plafond) tandis que l'AAD est le montant agrégé minimum de sinistre réassuré à atteindre avant intervention du réassureur (même fonctionnement qu'une franchise).

Aussi, lorsque que le total réassuré (charge totale du réassureur) dépasse la portée (lorsque la portée est dite *consommée*), il est possible de reconstituer la portée. Le nombre de reconstitutions est défini contractuellement et est noté *nombre de reconstitution@tarif*. Par exemple, si le taux de prime de réassurance est de 2 % et que les reconstitutions sont de la forme 1@100%, 2@25%, 1@0% alors la première reconstitution coûte 100 % du taux de prime soit 2 % de la prime d'assurance, les deuxième et troisième coûtent 0.5 % (25 % × 2 %), tandis que la dernière est gratuite. La prime de reconstitution est calculée dès qu'un sinistre consomme tout ou une partie de celle-ci, selon plusieurs méthodes :

— **Prorata capita** : calculée sur la base du sinistre survenu, elle est égale au montant de sinistre après application de la réassurance en proportion de la portée multiplié par la prime de reconstitution :

$$\frac{\text{sinistre réassuré}}{\text{portée}} \times \text{prime de réassurance} \times \text{taux de prime de la reconstitution}$$

— **Double prorata (prorata temporis et prorata capita)** : calculée sur la base du temps écoulé au moment de la survenance du sinistre et sur son montant, elle est égale au montant de sinistre après application de la réassurance en proportion de la portée multiplié par la prime de reconstitution et le temps restant avant la fin du contrat :

$$\frac{\text{sinistre réassuré}}{\text{portée}} \times \text{prime de réassurance} \times \text{taux de prime de la reconstitution} \times \frac{365 - i}{365}$$

où i est le nombre de jours écoulés depuis le début de l'année au moment de la survenance du sinistre.

Ainsi, en posant n le nombre de sinistres et m le nombre de reconstitutions, le total à charge du réassureur est

$$\begin{aligned} \text{Charge totale du réassureur } &= \min(\sum_{i=1}^{n} \min(L, \max(S_{A_i} - P, 0)), m \times L) \\ &= \min(\sum_{i=1}^{n} S_{R_i}, m \times L) \end{aligned}$$

où S_{A_i} est le i$^{\text{ème}}$ sinistre de la cédante et S_{R_i} le i$^{\text{ème}}$ sinistre du réassureur. La prime totale de réassurance dépend donc fortement des reconstitutions.

1.4.2.2 Stop Loss

La réassurance en Stop Loss est similaire à celle en XS à la différence qu'elle porte sur le résultat (le S/P = sinistre/prime) du portefeuille (et non sur le sinistre). Ainsi, priorité et portée sont toutes deux définies en pourcentage.

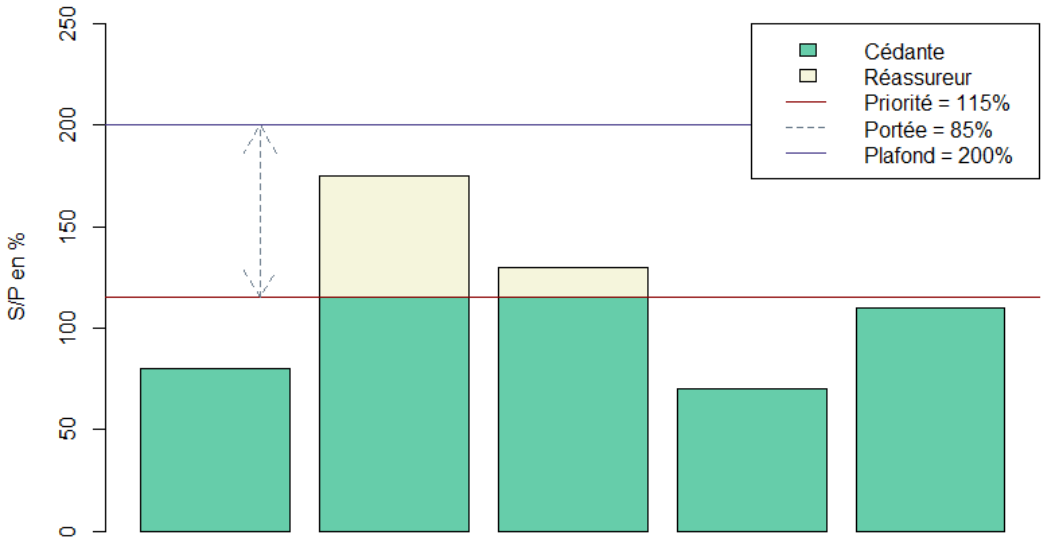


FIGURE 6 – Exemple de réassurance en Stop Loss 0.85 XS 1.15

Interprétation de la figure : Dans cet exemple, le traité est de la forme 85 % XS 115 %. Seules les deuxième et troisième années sont concernées, le réassureur intervient donc au-dessus de 115 %.

Le réassureur se doit de régler à la cédante le pourcentage de résultat réassuré à hauteur de la prime d'assurance, soit :

$$C_R = P \times \min(\text{portée}, \max(S/P - \text{priorité}, 0))$$

où C_R est la charge du réassureur, P la prime d'assurance annuelle et S le montant total de sinistre annuel.

Cette forme de réassurance permet, sur plusieurs années, de lisser le résultat d'un portefeuille. Elle est utilisée pour les risques ayant une fréquence connue et élevée, les risques dits cycliques, comme les tempêtes. En effet, la sévérité de ces sinistres peut être faible, ainsi un XS n'aura aucun impact, cependant si la fréquence est élevée alors la charge totale peut très fortement dégrader le résultat de la cédante.

Introduisons dès à présent les données utilisées dans le cadre de ce mémoire.

2 Description des données

Table des matières

2.1 AXA Global Re au sein du groupe AXA

La réassurance du groupe AXA s'organise intégralement autour d'une entité spécifique : AXA Global Re (AGRe). Ainsi, lorsqu'un traité de réassurance doit être souscrit ou renouvelé, celui-ci est cédé à AXA Global Re. Cette entité agît comme la plateforme des couvertures de réassurance du groupe. En pratique, lorsqu'une entité d'AXA, dans un pays en particulier, veut se couvrir à l'aide d'un traité en réassurance, AXA Global Re se place en tant que réassureur. Nous parlons alors de traités locaux. Ceux-ci sont parfois rétrocédés par AGRe (*inward treaties* dans le schéma ci-dessous). Dans le cas des traités dits groupes, toutes les entités d'AXA seront réassurées. Ils ont vocation à protéger le *pool* ou la *rétention* d'AXA Global Re (*outwards treaties* dans le schéma ci-dessous, partie basse *Reinsurance pooling*).

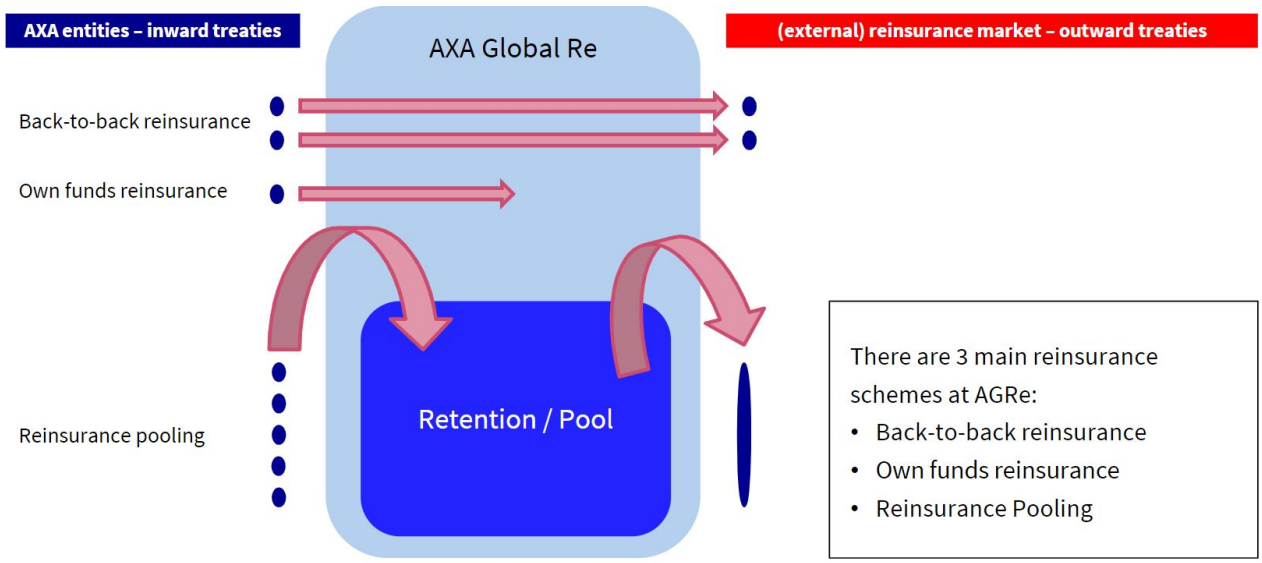


FIGURE 7 – Fonctionnement de la réassurance AXA

Ainsi, trois cas sont possibles :

1. **Back-to-back reinsurance :** le traité est rétrocédé à 100% en externe sur le marché de la réassurance. Il sera réassuré par un ou plusieurs réassureurs.
2. **Own funds reinsurance :** le traité est réassuré à 100% par AXA Global Re.
3. **Reinsurance pooling :** le traité est partiellement réassuré par AXA Global Re. Dans un *pool* de réassurance, le résultat des traités est redistribué aux entités participantes au pool, dont AGRe possède une part. Dans le cas inverse de la *rétention*, l'ensemble du résultat est conservé par AGRe.

Ce partage de sort entre entités et réassureurs externes permet d'optimiser les risques en fonction de la connaissance et des résultats des branches réassurées.

Scheme	Technical risk / result	Service fees	Counterparty risk
Back-to-back	No	Yes	Yes
Own-funds	Yes (higher risk compared to Retention/Pool)	No	No
Retention/Pool	Yes	Partially (only for the Group Cover placement)	Reduced (compared to back-to-back scheme)

FIGURE 8 – Implication de chaque schéma de réassurance

Par conséquent, chaque schéma de réassurance présente des avantages et des inconvénients. Le résultat technique est définit comme suit :

(+)	Primes entrantes
(+)	Récupérations sortantes
(-)	Primes sortantes
(-)	Récupérations entrantes

TABLE 4 – Résultat technique

Les primes incluent les commissions (en réassurance proportionnelle) ou les primes de reconstitutions (en réassurance non proportionnelle). Les récupérations sont définies comme la sinistralité associée au traité. Ainsi, la somme des sinistres éligibles à la réassurance est appelée récupération du traité. Les *service fees* sont principalement les frais de courtage et administratifs. Comme indiqué en [8], le risque de contrepartie peut être contrôlé grâce à ces schémas. Dans le contexte *back-to-back*, le traité étant rétrocédé à 100%, le risque de contrepartie existe et repose sur les réassureurs externes. A contrario, dans le cas où AXA Global Re réassure intégralement (*own-funds*), ce risque n'existe pas. Finalement, le schéma *retention/pool* permet d'ajuster ce risque en fonction de la proportion du traité réassuré par AXA Global Re.

2.2 Données et cadre de l'étude

Dans le cadre de la réalisation des travaux constituants ce mémoire, une base de données de traités en excédent de sinistre sur 2021 et 2020 est utilisée. Par soucis de confidentialité, les données nécessitant d'être affichées seront normalisées ou cachées quand cela se révélera nécessaire.

2.2.1 Contexte

Les données utilisées sont des cotations envoyées par les réassureurs. Lors d'une souscription ou du renouvellement d'un traité de réassurance rétrocédé, les réassureurs cotent celui-ci. Ils envoient alors une cotation (un tarif, une prime commerciale) correspondant au traité. Plusieurs réassureurs peuvent participer à ce processus de cotations. Finalement, une cotation finale et identique pour tous les réassureurs sera adoptée. Parfois, après réalisation d'un benchmark [1], les courtiers envoient directement une cotation vision marché. Toutes ces estimations de prix de réassurance permettent d'évaluer la perception externe des réassureurs des traités mis sur le marché par AXA Global Re. À cela, des données internes de modélisations de réassurance comme les pertes moyennes annuelles par traités sont ajoutées.

Lorsqu'un nombre suffisant de cotations nous est donné, plusieurs études sont réalisables. Dans ce mémoire, nous nous intéressons tout particulièrement à l'analyse des cotations des réassureurs. Celles-ci peuvent nous permettre d'estimer le tarif d'un traité que nous souhaitons rétrocéder. En effet, pour la réassurance en XS, il est possible de créer un modèle de tarification interne estimant le prix marché du traité. Pour ce faire, plusieurs méthodes sont possibles comme des algorithmes de prédiction en machine learning supervisé ou des algorithmes de minimisation basés sur une segmentation. Nous parlons alors de tarification *interne-externe*, car la modélisation de la prime commerciale d'un traité se base sur des informations à la fois internes (calculs des modélisations internes de réassurance) et externes (cotations des réassureurs). Cependant, avant d'analyser les cotations, il est primordial de s'assurer de la cohérence et de la fiabilité des données utilisées pour nos calculs. En effet, l'émergence de la quantité massive de données permet aujourd'hui un grand nombre d'études qui étaient impossibles auparavant. Selon Statista, nous assistons à une évolution exponentielle des données.

1. Étude de marché

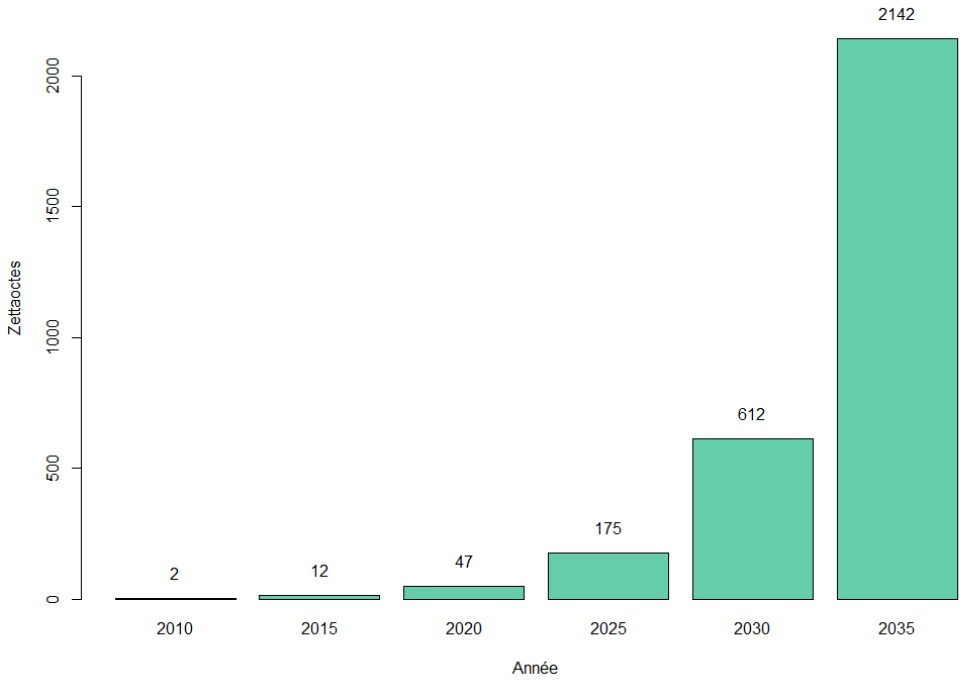


FIGURE 9 – Évolution de la quantité de données

Nous constatons une évolution très forte des zettaoctets [2] disponibles. À l'échelle mondiale, le volume de données a été multiplié par plus de vingt au cours de la dernière décennie. La démocratisation du *tout connecté* est l'un des principaux moteurs de cette augmentation. Pour faciliter la représentation, il faudrait aujourd'hui posséder 470 000 000 des plus gros disques durs actuels dans l'optique de stocker les données de 2020. L'évolution du volume entre 2020 et 2035 est une estimation.

Cette production de données est une chance pour les assureurs de mieux connaître le profil de leurs assurés. Ainsi, les modèles de tarification, la segmentation ou encore l'anti-sélection pourraient bénéficier d'une précision accrue. Cependant, tout ceci pose plusieurs problèmes potentiels :

— **Éthique :** mieux connaître ses assurés à travers des objets connectés ou l'achat de données pose un problème de protection de la vie privée. Pour remédier à cela, le 25 mai 2018, le Parlement européen a adopté un règlement général de protection des données (RGPD). Il permet de mieux protéger la vie privée des individus, à travers deux principaux axes. L'un sur le traitement des données et le second sur la responsabilisation des personnes (morales ou physiques) utilisant ces données. Il doit donc exister une certaine balance entre efficacité des modèles et précision/nombre de variables utilisées afin de rentrer dans le cadre du RGPD.
— **Technique :** le cumul du nombre de données rend la probabilité d'erreurs potentielles croissante. De plus, le travail de nettoyage s'avère être donc de plus en plus long et compliqué. La fiabilité est plus complexe à évaluer à mesure que la taille des bases de données augmente. L'actuaire doit donc être vigilant lorsqu'il exploite celles-ci pour ses travaux. Il doit toujours s'assurer de la cohérence et de la qualité des informations dont il

2. 1 zettaoctet = 1 000 milliards gigaoctets

dispose. Le rôle des services *data* est lui aussi extrêmement important dans la pérennité des travaux des assureurs.

De surcroît, la réassurance est un domaine où les données sont plus rares de par la nature même de ce type d'activité, par exemple lors de la tarification de traités en XS disposant de grandes priorités. En effet, plus la priorité est élevée plus la probabilité qu'un sinistre d'assurance soit éligible au traité est faible. Ainsi, plusieurs méthodes comme celle du burning cost deviennent inexploitables lorsque la quantité de sinistres historiques est trop faible. Il en va de même pour les méthodes d'estimations de lois sous-jacentes aux sinistres. Lorsque la quantité de données manque (en terme d'historique et de fréquence), la précision des paramètres des lois de probabilités se dégrade.

Cependant, dans notre cas, le nombre de cotations reçues ne suit pas une croissance exponentielle car il n'est pas lié à la quantité de données disponible. Néanmoins, nous pouvons supposer que les cotations sont de plus en plus précises car les sinistres d'assurance bénéficient d'une meilleure qualité [3] de modélisation avant envoi aux réassureurs. En clair, l'impact majeur de l'avènement du *Big Data* dans nos données est plutôt lié à la précision qu'à la quantité de celles-ci.

Finalement, le temps de collecte, de mise en forme et de nettoyage des données représente une très grande partie du travail de ce mémoire et consiste en une étape essentielle pour la suite des études. La rigueur lors de l'analyse de ces données est primordiale, sans cela la crédibilité de tous les travaux construits sur cette base en serait très fortement entachée.

2.2.2 Risques étudiés

Les études réalisées dans ce mémoire reposent uniquement sur de la réassurance non-vie.

2.2.2.1 Définition de l'assurance non-vie

L'assurance non-vie est une opération par laquelle l'assuré contracte, moyennant une prime, une prestation de l'assureur en cas de réalisation d'un risque. Les étapes de conception d'un produit d'assurance non-vie sont multiples.

Etape	**Activité du produit**	**Temps de l'activité**
1	Conception et tarification	Plusieurs mois
2	Vente et promotion	Plusieurs années
3	Encaissement des primes	Plusieurs années
4	Placements financiers	Plusieurs années

TABLE 5 – Conception d'un produit d'assurance non vie

L'aléa de ce type d'assurance porte à la fois sur le montant et sur la date de versement des flux. En effet, un sinistre, en fonction des types de branches (courtes ou longues), peut mettre plusieurs années avant de se clôturer et son coût ultime est inconnu. Ainsi, le risque principal pour l'assureur est la grande variabilité des sinistres que l'on qualifie de risque de variance. L'assurance non-vie concerne les activités ne dépendant pas de la vie des assurés comme les assurances de biens, de frais médicaux ou encore de responsabilité. Contrairement à l'assurance vie, où la probabilité de survenance du sinistre est certaine dans la majorité des cas (car le décès est certain), en non-vie, cette probabilité ne l'est pas. Il est possible qu'un risque ne se réalise

3. Car nous avons accès à une plus grande quantité d'information disponible sur les sinistres

jamais (exemple : un assuré n'ayant jamais eu d'accident de voiture). De plus, le montant du sinistre est rarement connu, ce qui est une autre particularité de cette assurance.

2.2.2.2 Description des risques de l'étude

Dans cette étude, les traités de réassurances sont concernés par cinq grands risques :

1. **Motor Third-Party Liability (MTPL) :** l'assurance responsabilité civile automobile garantit les dommages aux biens et à la santé causés par un accident automobile, lorsque le conducteur du véhicule est couvert par un contrat d'assurance. En France, elle est obligatoire pour se déplacer en voiture sur les routes publiques. En effet, la sévérité d'un sinistre pouvant être très importante, l'assuré n'est généralement pas assez solvable pour indemniser les personnes ou les biens auxquels il a causé des dommages. Cette branche fait partie de celles qui comptent le plus grand volume de primes avec au moins 30 % du total de primes en assurance non-vie [7]. La concurrence forte en MTPL a un tel impact à ce jour que, selon la FFA [8], les résultats de cette branche sont supérieurs à 100 % depuis 2016 [4].
2. **General Third-Party Liability (GTPL) :** l'assurance responsabilité générale garantie les dommages subis par un tiers ou une entreprise. Par exemple, si l'animal d'un tiers blesse une personne, celle-ci sera indemnisée par l'assurance GTPL de ce tiers. Aussi, elle peut intervenir dans les cas où un médecin cause des dommages à son patient (RC médicale). Généralement, ce type de garantie est compris dans le produit Multirisque Habitation (la MRH).
3. **Property (PTY) :** l'assurance habitation couvre les logements des particuliers et les bureaux d'entreprises. Les garanties généralement proposées protègent contre les inondations, incendies, bris de fenêtre ou encore le vol ayant lieux dans les locaux de l'assuré. En France, nous observons une augmentation de la sévérité et de la fréquence des dégâts des eaux (respectivement de 1.9 % et 0.9 % par an en moyenne) tandis que pour les incendies, la fréquence diminue de 2.6 % mais la sévérité augmente de 2.4 % [8]. Le résultat de cette branche est de 97.3 % en 2020.
4. **Property - Catastrophe (CAT) :** Ce risque est une extension du PTY, la différence se fait dans la cause des sinistres : ils sont d'origine catastrophique. Les catastrophes couvertes par ce type d'assurance sont de deux sortes : humaines ou naturelles. Un sinistre catastrophique causé par un humain peut être par exemple un attentat terroriste. Les sinistres d'origines naturelles sont des tempêtes, séismes, inondations, etc. Avec le changement climatique, ces derniers sont un vrai enjeu. Sur le territoire français, les risques climatiques ont générés plus de 3 milliards € de sinistres en 2020 :

4. L'année 2020 fait exception avec un résultat de 96.1 % principalement dût à la pandémie de Covid-19 ayant fortement diminué le trafic routier français.

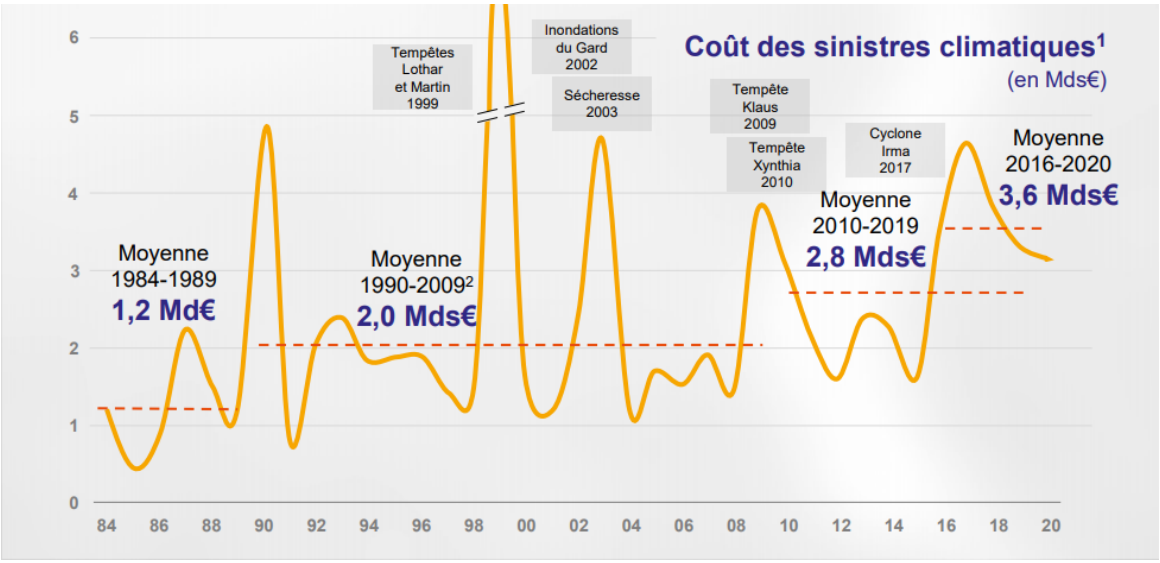


FIGURE 10 – Coûts des sinistres climatiques
[8]

La fréquence et la sévérité de ces sinistres allant en s'accroissant, la réassurance de ce type d'assurance est un pilier majeur de la stratégie de couverture des assureurs.

5. **Marine :** la branche Marine du groupe AXA couvre des risques très divers. Logiquement, le commerce maritime (navires, cargaisons, en mer ou sur un fleuve) est assuré. Aussi, nous retrouvons toutes les activités liées au pétrole (stations offshore et oneshore), à l'aviation, la guerre, les violences politiques ou encore les œuvres d'art.

2.2.3 Étude descriptive des données

La base de données des traités en excédent de sinistre regroupe un grand nombre de cotations issues de réassureurs ou de courtiers. Ces cotations sont des primes commerciales. Chaque ligne représente une prime commerciale estimée par un réassureur pour un traité (ou par traité et tranche si le traité a plusieurs tranches) en particulier.

2.2.3.1 Variables essentielles

Afin d'apporter le plus d'information possible, une grande variété de variables est utilisée. Elles sont de tous ordres (quantitatives, quantitatives) et reflètent les caractéristiques des traités. Nous pouvons cependant distinguer deux types d'informations :

— **Contractuelles :** issues des traités (limite/portée, priorité, reconstitution(s), cotation...).
— **Modélisées :** issues des modélisations internes AGRe (récupération moyenne du traité, facteur de reconstitution, probabilité d'épuisement des reconstitutions...).

En somme, cette double vision permet d'étudier la tendance des cotations externes par rapport au modèle interne de réassurance AXA. Plus précisément, nous pouvons par exemple analyser les primes cotées par les réassureurs en fonction de la récupération moyenne estimée. Ceci permet de challenger le modèle, étudier la vision des réassureurs sur nos traités et la comparer avec notre vision. Ci-dessous les variables de la base de données utilisée :

Variables contractuelles	Variables modélisées
Référence du traité	Récupérations moyennes annuelles du traité
Année de renouvellement	Écart type des récupérations
Risque [5]	Facteur de reconstitution
Entité réassurée	Probabilité d'attachement
Pays de l'entité	Probabilité d'épuisement
Région de l'entité	Prime pure du traité
Réassureur	
Courtier	
Part du réassureur dans le traité	
Limite	
Priorité	
AAD	
AAL	
Reconstitutions	
Taux de cession du traité sur le marché	
Rate on Line	
Prime commerciale cotée	

TABLE 6 – Variables essentielles de la base des traités en XS

Tous les montants sont en euros. Pour plus de clarté, définissons désormais les variables modélisées. Leur compréhension est essentielle pour la suite.

2.2.3.1.1 Définition des variables modélisées

Tout d'abord, introduisons les données sur lesquelles ces variables sont calculées. Avant chaque modélisation d'un traité de réassurance, une base de sinistres simulés associée à ce traité doit être utilisée comme support de calculs. Le calibrage de ces simulations peut s'effectuer sur des sinistres historiques ou par des générateurs de scénarios ou bien avec une méthode par exposition. Lorsque la simulation se réalise sur une base historique, nous parlons alors d'une approche fréquence-sévérité.

Rappel de l'approche par fréquence-sévérité

Cette approche très populaire repose sur deux principes autour de la charge de sinistre : une loi de fréquence et une loi de sévérité. Plaçons-nous dans un modèle collectif où S est la charge annuelle de sinistre de plusieurs polices d'assurance. Alors,

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

où X_i est le montant du sinistre i de la police et N le nombre annuel de sinistres. Par hypothèse :

— les X_i sont tous indépendants et identiquement distribués.
— $\forall i, N \perp\!\!\!\perp X_i$.

5. CAT/MTPL/GTPL/Marine/PTY

Cependant, l'indépendance des sinistres entre eux n'est pas toujours vérifiée. Ceci est le cas en assurance catastrophe naturelle lorsque les sinistres ont tous le même fait générateur (une tempête ou une inondation par exemple). L'hypothèse d'indépendance entre le nombre et la sévérité des sinistres n'est par exemple pas vérifiée en assurance automobile responsabilité civile : les zones rurales ont une fréquence de sinistre plus faible avec des coûts généralement plus importants tandis qu'au contraire, en zone urbaine, la fréquence est plus élevée mais la sévérité est plus faible. L'importance de la segmentation lors de la simulation des sinistres est alors très importante afin de regrouper les polices homogènes. Cependant, le risque de modèle est accru car la quantité de données par segment diminue. Posons X de même loi que les X_i, de moyenne m et de variance σ^2. Listons quelques propriétés de S, en supposant nos hypothèses valides :

Moyenne de S

$$\mathbb{E}[S] = \mathbb{E}[\mathbb{E}[S \mid N]] = \sum_{n=0}^{\infty} P(N=n) n \mathbb{E}[X] = \mathbb{E}[X]\mathbb{E}[N].$$

Variance de S

$$\mathrm{Var}(S) = \mathbb{E}[\mathrm{Var}(S \mid N)] + \mathrm{Var}(\mathbb{E}[S \mid N]) = \mathbb{E}[N]\sigma^2 + \mathrm{Var}(N)m^2.$$

Fonction génératrice de S

$$\begin{aligned} M_S(t) &= \mathbb{E}_N \mathbb{E}_{S|N}\left[e^{tS} \mid N\right] = \sum_{n=0}^{\infty} P(N=n)\mathbb{E}\left[e^{tS} \mid N=n\right] = \sum_{n=0}^{\infty} P(N=n)\left(M_X(t)\right)^n \\ &= \mathbb{E}\left[M_X(t)\right]^N = \mathbb{E}\left[e^{\ln M_X(t) N}\right] = M_N\left(\ln M_X(t)\right). \end{aligned}$$

Les lois usuelles pour la distribution de la fréquence sont généralement celles de Poisson ou Binomiale Négative. Pour la sévérité, un grand nombre de lois est possible comme celle de Gamma, Pareto ou encore Log Normale.

Méthode par exposition

Cette méthode est basée sur des courbes d'exposition qui représentent un taux de destruction, noté TD par la suite. Le principe de la méthode par exposition repose sur une séparation entre la somme assurée, notée SA, et le taux de destruction $TD = \text{sinistres}/SA$. Cette approche a été développée pour répondre aux deux principaux problèmes de la calibration de lois sur les sinistres historiques :

— L'hypothèse de stabilité du portefeuille modélisé (*le passé permet de prédire le futur*)
— Le manque d'historique de sinistres atypiques dans la calibration

L'hypothèse de l'approche par exposition est la suivante : la *quantité* de risque, ici estimée par SA, est inversement proportionelle au taux de destruction moyen. Nous pouvons illustrer cet exemple par le cas empirique du *feu de maison/château*. Une maison en feu a une grande probabilité d'être entièrement détruite par un incendie tandis qu'un château sera probablement partiellement détruit. Par conséquent, les polices d'une même branche avec une quantité similaire de risque ont le même taux de destruction. Ainsi, pour chaque police homogène, nous modélisons le taux de destruction et la somme assurée. Pour se faire, nous utilisons la fonction de répartition empirique de SA et TD qui nous permet de simuler la sévérité définie par $X = SA \times TD$ par la méthode de Monte-Carlo. La distribution de la fréquence est déduite du

loss ratio de chaque groupe homogène de polices.

Méthode par scénario

Cette approche permet de calibrer un modèle avec un scénario [6]. Elle nécessite cependant trois informations : une fréquence moyenne, les montants et les périodes de retour des scénarios. Elle est utilisée lorsque, par exemple, aucune courbe d'exposition ne peut être calculée ou lorsqu'il n'y a pas d'expérience sur les sinistres ou encore quand des sinistres extrêmes sont apparus sur le marché mais qu'ils ne sont pas dans les sinistres historiques. Un scénario peut être :

— Un sinistre extrême apparu mais non concerné par une entité.
— Une perte estimée suite à une attaque terroriste.
— Une perte bicentenaire estimée suite à un séisme.

La méthode est basée sur la formule de l'OEP [7] et sur un algorithme de minimisation entre l'OEP et les scénarios donnés en input. Le résultat est un modèle fréquence-sévérité.

Cas particulier du CAT

Désormais, nous savons comment nos sinistres d'assurances sont générés. Nous en distinguons deux formes :

— **Par évènement :** les sinistres sont tous causés par le même fait générateur (une tempête par exemple).
— **Par risque :** les sinistres sont indépendants entre eux.

Dans nos données, le risque Property CAT est le seul par évènement. Les sinistres sont tous considérés comme **atypiques** et sont simulés par un générateur atypique par évènement, contenant plusieurs sous-modèles. En effet, l'indépendance entre la fréquence et la sévérité n'étant plus vérifiée, les sinistres sont liés par l'intensité de l'évènement sous-jacent.

6. Une perte de marché ou une opinion d'expert
7. $\forall x > 0,\ \text{OEP}\ (x) = P\left[\max_{i=1}^{N}(X_i) > x\right]$

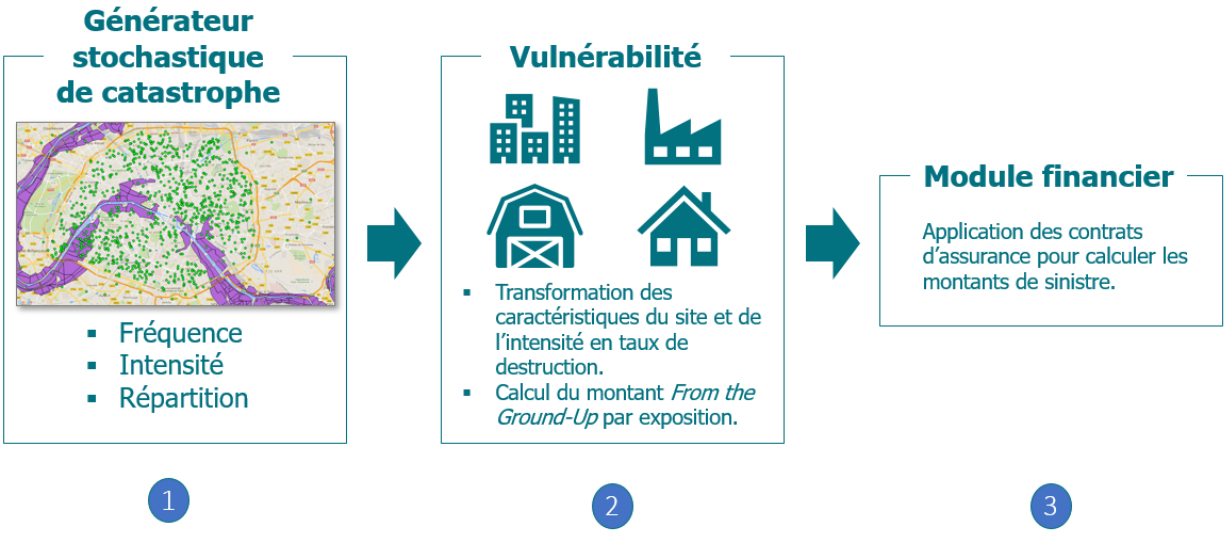


FIGURE 11 – Modélisation des sinistres CAT

Ainsi, nous générons des catastrophes en calculant leur intensité, répartition et fréquence. Une fois ces informations obtenues, nous pouvons impacter le site de la catastrophe en fonction des caractéristiques des biens présents sur celui-ci. L'impact est mesuré par le taux de destruction permettant d'obtenir le montant FGU (montant total de la catastrophe). Par après, nous étudions la façon dont nos polices d'assurance prennent en compte cet évènement. Ici, nous considérons les polices du risque Property. Nous obtenons alors, par police, un montant de sinistre catastrophe.

Modélisation des traités de réassurance

Ainsi, pour chaque branche et portefeuille homogènes, nous simulons des sinistres après utilisation des approches explicitées ci-dessus. Afin d'obtenir une bonne précision dans les calculs qui sont réalisés, nous simulons un grand nombre d'années de sinistres (50 000). Puis, si un portefeuille est réassuré, nous lui appliquons le(s) couverture(s) de réassurance dont il fait l'objet afin d'obtenir une base de sinistres *post-réassurance*. Par la suite, nous calculons alors les **variables modélisées** présentées auparavant. Ci-dessous une vision simplifiée du procédé de modélisation de réassurance :

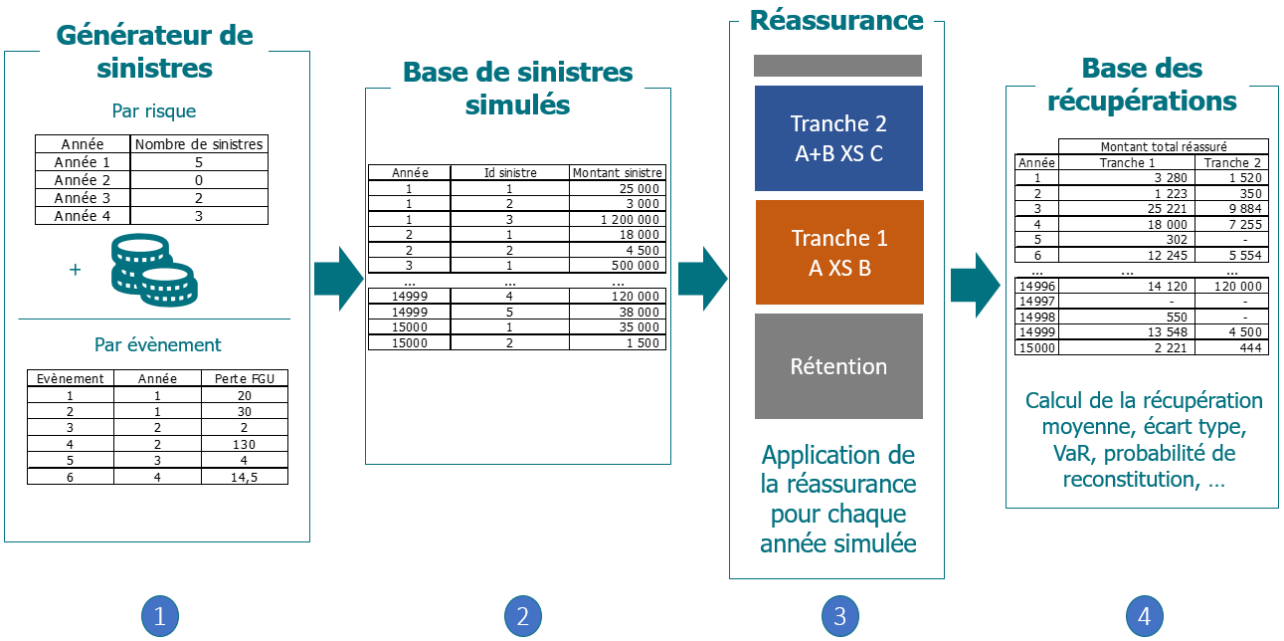


FIGURE 12 – Modélisation des récupérations de réassurance

Nous pouvons à présent introduire les variables modélisées. Posons n le nombre d'années de sinistres simulées (50 000), M la variable aléatoire de la fréquence annuelle de sinistre d'assurance avec M_i la $i^{\text{ème}}$ réalisation de celle-ci pour l'année simulée i et X_j^T le sinistre j après application du traité de réassurance T. La **récupération** d'un traité T pour l'année i se définit comme la somme des sinistres d'assurance de l'année i après application de la réassurance T. Alors, les variables modélisées sont calculées comme suit :

— **Récupération moyenne :** moyenne de la somme annuelle des sinistres après réassurance, égale à

$$\bar{R} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{M_i} X_j^T.$$

Par convention, si la réalisation i de M est égale à 0 ($M_i = 0$), $\sum_{j=1}^{M_i} X_j^T = 0$.

— **Écart type des récupérations :** écart type de la somme annuelle des sinistres après réassurance, égale à

$$\sigma_R = \sqrt{\frac{\sum_{i=1}^{n} (\sum_{j=1}^{M_i} X_j^T - \frac{1}{M_i} \sum_{j=1}^{M_i} X_j^T)^2}{n}}.$$

— **Facteur de reconstitution :** facteur indiquant la part des primes de reconstitutions dans la prime de réassurance. Il est calculé en divisant la prime totale (avec reconstitutions) moyenne sur la prime de réassurance. Il peut s'écrire de la forme

$$Rec_{factor} = \frac{f(P)}{P}, \quad P = \text{prime de réassurance}$$

Or, $f(P)$ est une fonction de la forme : $f(P) = P + \alpha \times P + \beta \times P + ... + \delta P$ où $\alpha, \beta, ..., \delta$ sont les tarifs des différentes reconstitutions par sinistre, calculés par pro-rata capita, définis dans le traité T. Ainsi, Rec_{factor} ne dépend pas de P et est de la forme $1 + \alpha + \beta + ... + \delta$. Si ce facteur vaut 1, alors il n'y a pas de reconstitution en moyenne pour le traité T.

— **Probabilité d'attachement :** probabilité que le traité T ait une récupération positive. En clair, c'est la probabilité que le traité T soit *touché* par au moins un sinistre :

$$Attachement_{prob} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{\sum_{j=1}^{M_i} X_j^T > 0\right\}.$$

— **Probabilité d'épuisement :** probabilité que la capacité totale du traité T soit entièrement consommée. La capacité totale est la perte maximale que le traité T peut couvrir, c'est l'AAL. Elle est soit fixée comme un montant lors de la souscription soit égale au nombre total de reconstitutions plus une fois la limite :

$$Perte_{max} = (\#\text{Reconstitutions} + 1) * \text{Limite} = AAL.$$

Si la limite du traité est infinie, la capacité totale est aussi infinie. Ainsi, la probabilité d'épuisement de la capacité totale est égale à

$$Epuisement_{prob} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{\sum_{j=1}^{M_i} X_j^T > Perte_{max}^T\right\},$$

où $Perte_{max}^T$ est la capacité totale du traité T.

— **Prime pure du traité :** estimation de la prime pure du traité (vision interne), égale à la récupération moyenne divisée par le facteur de reconstitution :

$$PP = \frac{\bar{R}}{Rec_{factor}}.$$

2.2.3.2 Nettoyage des données

La cohérence des données est essentielle pour pouvoir les exploiter. Ainsi, l'étape de nettoyage est primordiale avant toute étude. Nous distinguons deux types de procédés de nettoyage :

— Traitement des valeurs manquantes
— Traitement des valeurs extrêmes

2.2.3.2.1 Valeurs manquantes

Les valeurs manquantes ne sont pas, la plupart du temps, acceptées par les modèles de prédiction. De plus, elles nuisent à la qualité des statistiques qui vont être calculées car elles réduisent le nombre de valeurs et donc la précision des indicateurs que nous souhaitons mesurer. Cependant, il n'existe pas de technique universelle pour traiter ces cas. En fonction de nos besoins et de la connaissance des données nous distinguons plusieurs manières de corriger les valeurs manquantes :

— **Correction par valeur fixe :** remplacement de la valeur manquante par une valeur fixe, pouvant être la moyenne, la médiane, une constante, ... Cette méthode est rapide et simple à mettre en place, cependant, dans certains cas, il peut s'avérer impossible de la remplacer par une valeur fixe car ceci se révélerait incohérent. Par exemple, si l'AAD d'un traité est manquant, il est aberrant de le remplacer par une valeur fixe, car cette variable n'a pas de tendance, ni de valeurs usuelles. Nous ne pouvons donc pas la remplacer par l'AAD moyen, ni même par l'AAD moyen des traités *voisins* (qui se ressemblent).

— **Correction par prédiction :** utilisation d'un algorithme de prédiction entraîné sur les valeurs renseignées. Ainsi, les valeurs manquantes sont prédites sur la base des autres observations par l'algorithme. Cependant, le nombre de données est primordial pour s'assurer de la qualité de prédiction. De plus, il faut accepter que certaines valeurs manquantes soient remplacées par de fausses valeurs (en classification) ou qu'elles soient mal estimées (en régression). Ce choix peut fortement nuire aux futurs modèles qui seront entraînés après correction des valeurs manquantes.
— **Suppression :** suppression des individus (lignes) qui comportent des valeurs manquantes. Lorsqu'il est impossible de corriger ces valeurs, l'unique solution est de les supprimer. De plus, en ne corrigeant pas ces valeurs, il n'y a pas de risque de mauvais remplacement. Cependant, si leur nombre est grand le risque de sous-échantillonnage est élevé, rendant les études sur les données bien moins qualitatives. Ainsi, cette technique est réservée aux cas où la quantité de valeurs manquantes est faible.

Dans notre cas, étant donné le faible nombre de valeurs manquantes, nous utiliserons la méthode de suppression. En effet, les autres méthodes pouvant créer des corrections incohérentes, il n'est pas souhaitable ici de les utiliser. Pour illustrer ce propos, si le nombre de reconstitutions d'un traité est manquant, il est alors insensé de le remplacer par une moyenne des reconstitutions ou par une prédiction. Cette variable étant le fruit de choix commerciaux et financiers ainsi que de l'appétit au risque de l'assureur, il serait incohérent de choisir cette valeur par prédiction ou par valeur fixe. Il en va de même pour les autres variables informatives des traités.

2.2.3.2.2 Valeurs extrêmes

Aussi appelées *outliers*, elles peuvent être de deux sortes :

— **Aberrante :** valeur qui est manifestement fausse. Dans ce cas, il est nécessaire de la supprimer.
— **Atypique :** valeur qui se *détache* des autres mais correcte. Nous pouvons alors distinguer les valeurs atypiques et standard, pour les étudier séparément. Sinon, il est également possible de les réintégrer au jeu de données standard. Par exemple, si nous traitons des montants de sinistres, nous pouvons fixer un seuil extrême au dessus duquel les valeurs sont considérées comme atypiques. Par la suite, nous majorons uniformément chaque sinistre attritionnel (usuel) par la charge de sinistre atypique agrégée.

Graphiquement, la différence entre atypique et aberrant peut se faire de cette façon :

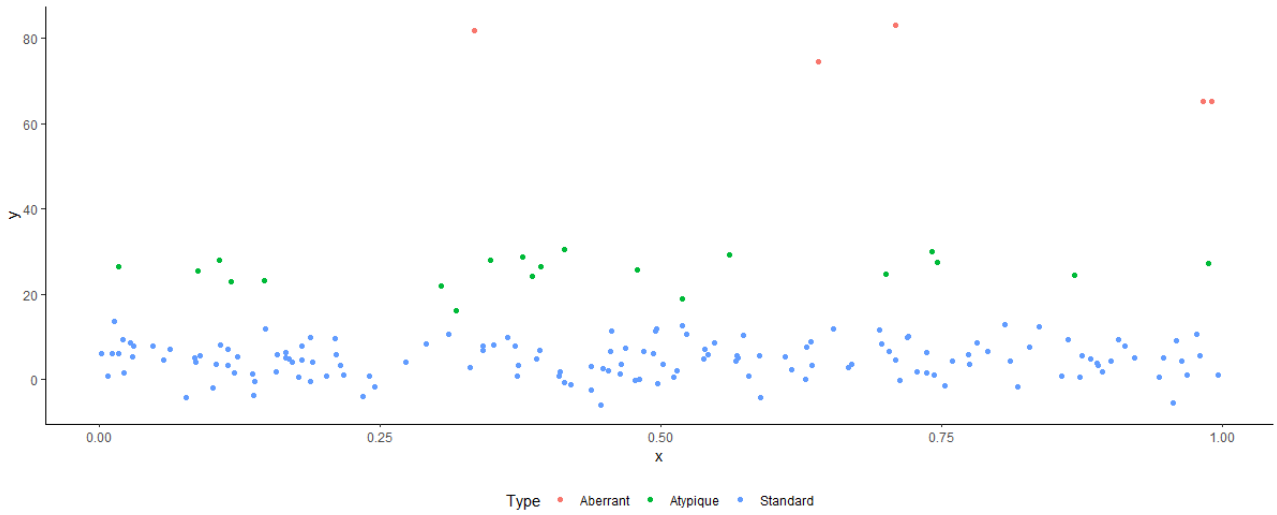


FIGURE 13 – Données atypiques et aberrantes

Les points atypiques se démarquent mais ne sont pas très éloignés des données standards tandis que celles aberrantes sont nettement au-dessus du reste. Il est alors possible de classer ces données par type pour les traiter. Dans notre cas, l'analyse des données extrêmes porte sur les variables modélisées. Une variable pertinente pour quantifier cet extrémisme est le **multiple** :

$$\begin{aligned}\text{Multiple} &= \frac{\text{Rate on line}}{\text{Loss on line}} \\ &= \frac{\text{Prime de réassurance}}{\text{Limite}} \times \frac{\text{Limite}}{\text{Prime pure}} \\ &= \frac{\text{Prime de réassurance}}{\frac{\bar{R}}{Rec_{factor}}}\end{aligned}$$

Ici, la prime de réassurance considérée est la cotation du traité calculée par le réassureur. Le multiple représente en somme la rentabilité du traité et étant la prime sur la perte espérée, il peut être résumé à l'indicateur répondant à la question : *Combien me coûte la réassurance d'un sinistre à 1 € ?* Supposons un traité avec un multiple de 200 %, cela signifie que 2 € de prime protège (en moyenne) contre 1 € de sinistre. Ceci permet par exemple de mesurer l'écart de vision entre cédant et réassureur, de mesurer leur appétit au risque ou de visualiser le chargement appliqué sur la prime pure (en vision interne) pour estimer la cotation du réassureur.

Par ailleurs, nous pouvons dès à présent avoir une idée du comportement de cet indicateur. En effet, s'il prend des valeurs très élevées (de l'ordre de 20 ou plus) ou très faibles (proches de 0) celles-ci sont considérées automatiquement comme aberrantes. Effectivement, cela signifie que la vision du modèle interne est trop éloignée de celle des réassureurs, conduisant à des résultats incohérents si ces observations sont maintenues dans notre base. Un multiple très grand implique que les récupérations sont sous-évaluées : les lois sous-jacentes des sinistres ne sont peut-être pas bien estimées. Il est également possible que le réassureur n'ait pas assez de données pour tarifier correctement. Sa cotation est donc imprécise. Il peut donc la majorer afin de réduire le risque de prime[8], impliquant une cotation élevée. Inversement, un multiple strictement inférieur à 1 signifie que la vision interne de modélisation des traités de réassurance est pessimiste comparée à celle du réassureur. La vision de la cédante est donc plus risquée que celle du réassureur. Aussi, le réassureur disposant d'une capacité de diversification géographique

8. Risque de mal estimer la prime et donc de subir des pertes réelles plus élevées que celles estimées.

forte, les pertes pouvant être compensées par d'autres traités, la prime proposée peut se révéler plus *faible* que celle estimée en interne, calculée sur un traité uniquement. Nous procédons donc à un pré-nettoyage en supprimant les données aberrantes par nature. Analysons désormais le comportement de cet indicateur sur le graphique ci-dessous où chaque point est une cotation d'un traité.

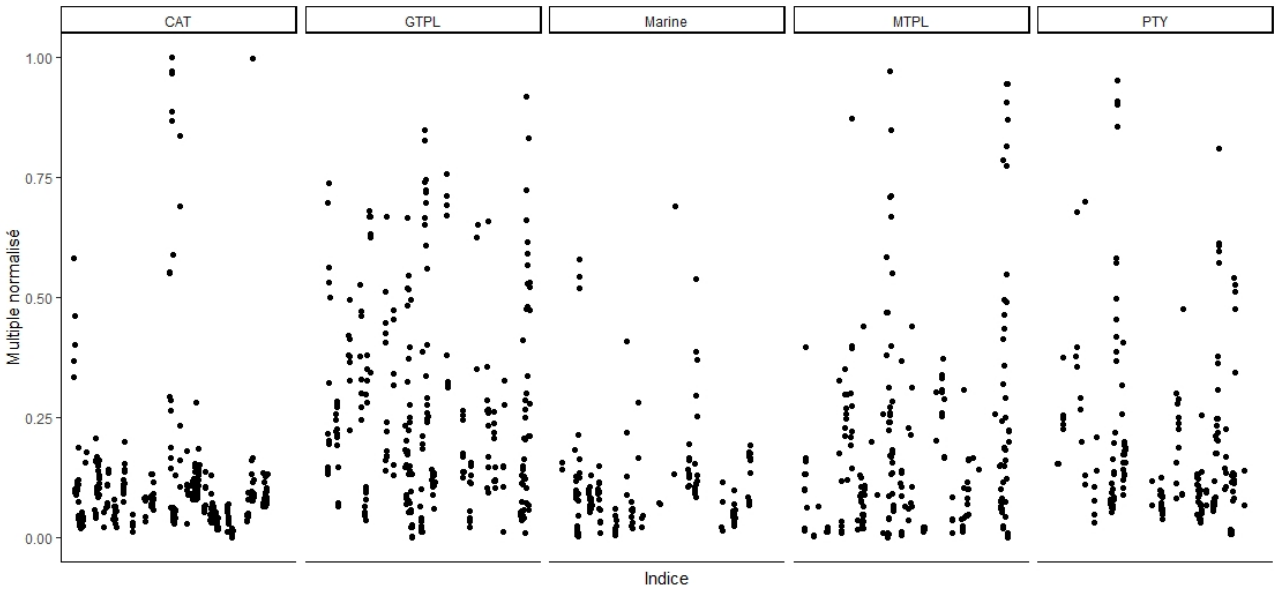


FIGURE 14 – Multiple par risque avant nettoyage

Premièrement, nous remarquons une tendance différente par risque. Certains étant moins bien connus que d'autres, le résultat était attendu. Ainsi, pour le CAT nous observons une nette séparation des points dès que le multiple (normalisé par mesure de confidentialité) dépasse 0.25. Cependant, tous ces points ne sont pas aberrants, certains sont atypiques et doivent donc être conservés. Le Marine possède aussi un seuil proche de 0.25 mais les valeurs aberrantes sont plus faibles. Également, le MTPL suit approximativement le même comportement sans démarcation nette. Nous remarquons tout de même des points proches de 1 semblant être aberrants, le seuil atypique se situe autour de 0.4. En PTY, plusieurs points sont isolés dès que le multiple normalisé est supérieur à 0.7. Enfin, le GTPL est le risque le plus homogène, il n'y a pas de groupes atypiques et aberrants visibles. Après étude de chaque risque, nous sommes en capacité de déterminer les points atypiques ou aberrants. Concentrons-nous désormais sur les cas où le multiple est inférieur à 1.

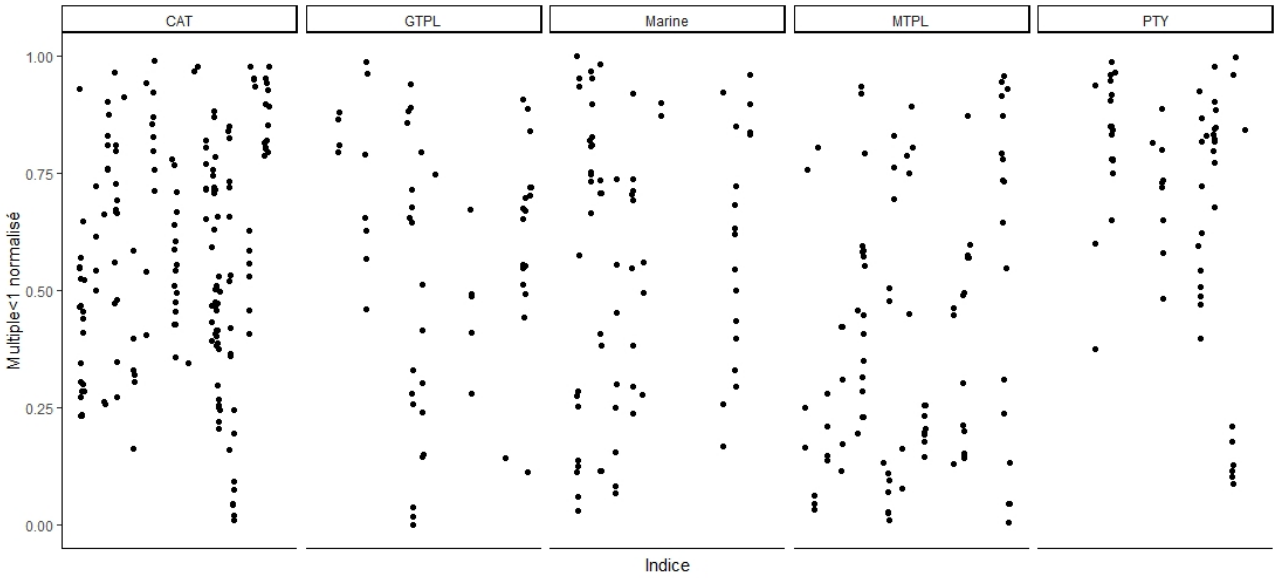


FIGURE 15 – Multiple inférieur à 1 par risque avant nettoyage

Un nombre de points non négligeable apparaît sur ce graphique. Cependant, comme attendu une nouvelle fois, les risques ne sont pas tous concernés de la même manière. Le risque CAT est celui comptant le plus de points où le multiple est inférieur à 1. Par ailleurs, il n'y a pas de tendance nette se dégageant des nuages de points. Nous percevons tout de même quelques points proches de 0, qui sont aberrants au sens du modèle interne, où la prime cotée par le réassureur est bien trop faible par rapport à la prime pure estimée. Ces données sont considérées comme aberrantes et sont alors supprimées. Maintenant, il est nécessaire de fixer un seuil en dessous duquel les points sont tous considérés comme aberrants. Malheureusement, aucune démarcation visible ne permet de trancher. Dans ce type de cas, nous pouvons faire appel à l'expertise interne des collaborateurs pour nous éclairer sur les raisons de l'existence de ces multiples. Avec l'expérience, ils peuvent nous indiquer, par branche, par traité, quels sont les résultats habituels attendus. En effet, parfois, les méthodes statistiques et/ou visuelles ne sont pas suffisantes pour analyser des données. Il est alors important de pouvoir se référer à des experts qui, grâce à leurs analyses et leur expérience, nous aiguillent sur les choix à effectuer. Finalement, nous déterminons le seuil par expertise interne. Les autres variables modélisées étant calculées sur la base des récupérations, le nettoyage du multiple permet de traiter l'ensemble des variables issues du modèle de réassurance.

2.2.3.3 Analyse descriptive des traités en XS

Le jeu de données étant désormais considéré comme exploitable, nous pouvons l'explorer afin de mieux le comprendre. Cette étape est nécessaire afin de d'ores et déjà avoir une idée sur la façon de réaliser nos futurs études : faire un modèle par risque, utiliser cette variable pour imager un besoin donné, segmenter selon cette variable, déceler des corrélations et des tendances, etc.

2.2.3.3.1 Exposition géographique

Analysons l'exposition géographique des traités locaux. Pour rappel, un traité est dit local s'il est souscrit par une entité d'AXA (AXA France, AXA Belgique, AXA Japon, ...) et qu'il couvre uniquement celle-ci. Il existe des cas particuliers où l'exposition peut être régionale [9].

9. Régional signifie dans notre cas une partie du monde comme l'Asie ou le Moyen-Orient.

Dans ce cas, le traité couvre plusieurs entités à la fois. Considérons l'exposition d'un traité comme sa cotation moyenne issue des réassureurs. Ainsi, plus la prime cotée moyenne est élevée plus l'entité est couverte par des traités. Effectivement, il paraît raisonnable de choisir une variable de prime comme exposition. En assurance, l'assiette de prime est souvent considérée comme l'exposition d'un portefeuille. Ainsi, pour chaque pays, nous sommes en mesure de calculer la cotation moyenne totale, c'est-à-dire la somme des cotations moyennes des traités. L'exposition au risque d'une entité peut être considérée comme inversement proportionnel à sa prime totale de réassurance. En effet, nous pouvons considérer que plus celle-ci paie de réassurance plus elle est protégée. Ci-dessous la carte des expositions de réassurance 2021.

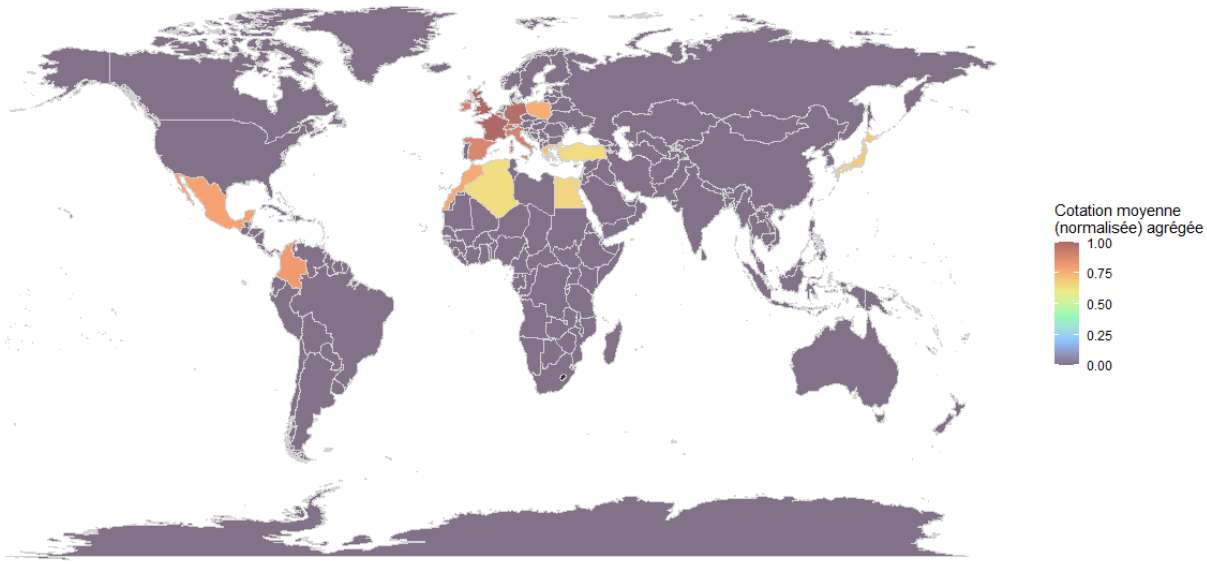


FIGURE 16 – Carte des cotations des traités locaux

D'emblée nous remarquons une disparité entre pays et régions qui peut s'expliquer de plusieurs façons. Tout d'abord, certains pays possèdent des entités AXA depuis plus longtemps que les autres. Nous les appelons les entités *matures*. Nécessairement, le nombre de traités et de risques souscrits est potentiellement plus élevé que dans un pays où une entité n'est implantée que depuis peu de temps. De plus, certains risques ne sont pas équivalents pour tous les pays. Par exemple, le risque catastrophe naturelle diffère entre le Japon et la France. L'un est touché par des séismes ou des tsunamis (Japon) tandis que le second est plutôt concerné par le risque de grêle ou d'inondations (France). Ainsi, pour une *même* branche, les garanties peuvent nettement changer, impliquant des primes de réassurance différentes. De plus, certains pays ayant une meilleure connaissance de leurs assurés que d'autres, notamment en terme d'historique de données, la précision de la prime de réassurance peut varier. Supposons qu'AXA Maroc possède plus d'historique qu'AXA Algérie, les réassureurs seront plus concernés par le risque de prime en Algérie plutôt qu'au Maroc.

Finalement, notre carte montre que l'Europe de l'Ouest est la région avec le plus de primes de réassurances. Au sein de cette région, la France ainsi que le Royaume-Uni et l'Allemagne se démarquent des autres pays (couleur la plus foncée). L'Asie du Sud ainsi qu'une partie du Maghreb, de l'Europe de l'Est et du Moyen-Orient sont celles où la cotation moyenne totale est la plus faible.

2.2.3.3.2 Courbe Rate on Line Loss on Line

La courbe Rate on Line Loss on Line (Courbe RoL LoL) permet de comparer la prime de réassurance et la prime pure de réassurance. Cette courbe est un bon indicateur du comportement du RoL et donc de la cotation des réassureurs.

Rappels

$$\text{Rate on line} = \frac{\text{Prime de réassurance}}{\text{Limite}}, \quad \text{Loss on line} = \frac{\text{Prime pure de réassurance}}{\text{Limite}}$$

Théoriquement, si la prime est estimée par le même acteur, le RoL est supérieur au LoL. En effet, la prime de réassurance étant une moyenne majorée par définition, elle doit être au-dessus de la prime pure. Cependant en pratique, deux acteurs entrent en jeu : le réassureur calcule sa cotation avec son propre modèle tandis que la cédante modélise sa prime pure de son côté. Ces deux acteurs n'ayant pas la même vision (différents portefeuilles, historique de données, diversification géographique et temporelle, etc...), le RoL n'est pas nécessairement supérieur au LoL. En effet, comme remarqué précédemment lors de l'étude du multiple [2.2.3.2.2], ce cas est fréquent et diffère selon les risques. Généralement, l'interprétation de ce type de courbe se réalise ainsi :

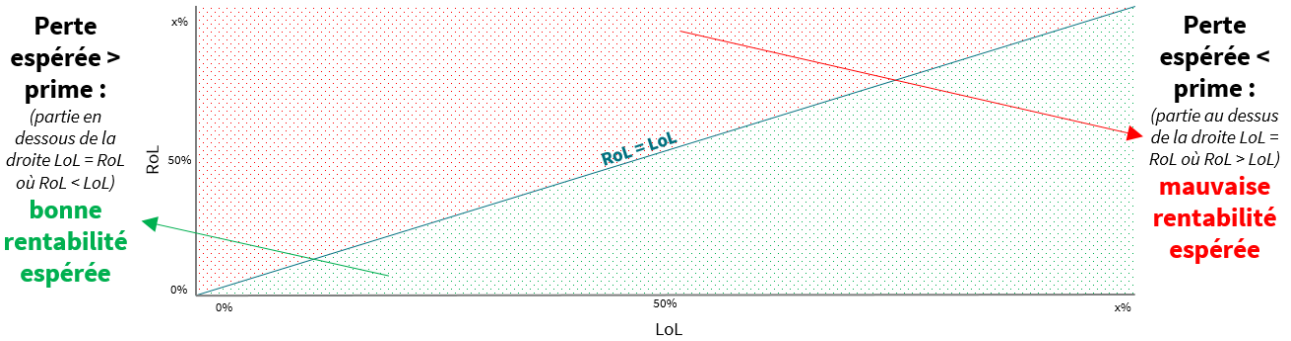


FIGURE 17 – Interprétation de la courbe Rate on Line Loss on Line

Comme indiqué sur ce graphique, la partie où le RoL est supérieur au LoL correspond aux cas où, en moyenne, le résultat du traité (prime de réassurance - sinistres) sera négatif. Inversement, la partie verte, celle où le LoL est inférieur au RoL, le résultat espéré (en moyenne) du traité est positif. Ainsi, tout repose sur les modèles sous-jacents de ces deux indicateurs, leur façon d'être calculés. Ceci aura un impact fort sur les résultats potentiels des deux parties. Penchons-nous dès à présent sur nos données.

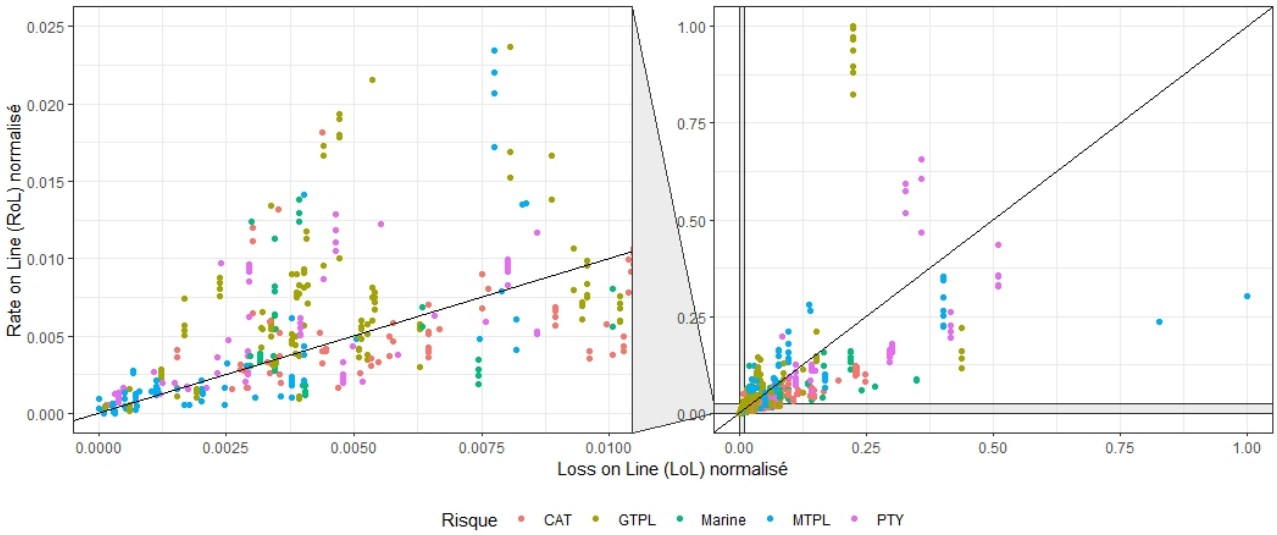


FIGURE 18 – Rate on Line - Loss on Line

Interprétation des deux graphiques : le graphique de droite représente tous les points de la base de données. Le graphique de gauche est un zoom sur la partie où le LoL normalisé est entre 0 et 0.1 et le RoL normalisé entre 0 et 0.2.

Analysons chaque graphique un à un :

- **Graphique de droite :** les points où le LoL est le plus important représentent les traités ayant des tranches [10] basses [11] tandis que ceux de gauche concernent les traités possédants les tranches les plus hautes. Par définition, les tranches basses étant les plus travaillantes [12], le RoL et le LoL sont plus élevés. La volatilité autour de la tarification du RoL est croissant avec la hauteur de la tranche. En effet, les tranches les plus hautes sont les moins travaillantes conduisant donc à une estimation moins précise des sinistres potentiels qui pourraient toucher le traité. En GTPL, deux traités possèdent tout particulièrement des RoL bien plus élevés que leur LoL associé.
- **Graphique de gauche :** la différence d'ordre de grandeur des RoL/LoL nécessite de *zoomer* sur les tranches faibles. Nous remarquons une volatilité assez élevée autour de la ligne médiane (celle où RoL = LoL). Aucune tendance par risque ne se dégage. Plus généralement, nous observons un nombre de points au-dessus de la ligne médiane croissant au fur et à mesure de la diminution du LoL.

Finalement, les LoL des tranches les plus basses sont en majorité plus élevés que leurs RoL. Une des raisons potentielles principales réside dans le fait que les réassureurs bénéficient probablement d'une mutualisation forte sur les tranches basses, leur permettant d'être compétitifs dans leurs cotations par rapport à la récupération moyenne du traité modélisé en interne. De plus, la quantité de sinistre étant plus forte, leur tarification est plus fiable et donc le risque de prime est minoré ce qui leur permet de proposer ces tarifs. Inversement, les tranches élevées sont peu travaillantes : le nombre de sinistres touchant la réassurance est faible (voir inexistant dans

10. Une tranche d'un traité XS est un terme désignant sa portée et sa priorité, équivalent à son exposition au risque.
11. Une tranche d'un traité XS est basse si la priorité est basse.
12. Une tranche d'un traité XS est travaillante si sa probabilité d'être touchée par un sinistre est élevée.

certains cas) rendant la tarification compliquée et obligeant le réassureur à se couvrir contre le risque de prime en proposant des tarifs importants. La hauteur de la tranche n'est cependant pas le seul critère influant sur la tendance des cotations.

2.2.3.3.3 Cotation par tranche

Une façon intéressante de visualiser la tendance des cotations est d'analyser leur évolution en fonction de la tranche. Celle-ci étant représentée par sa priorité et sa limite, nous pouvons l'afficher l'observer ainsi :

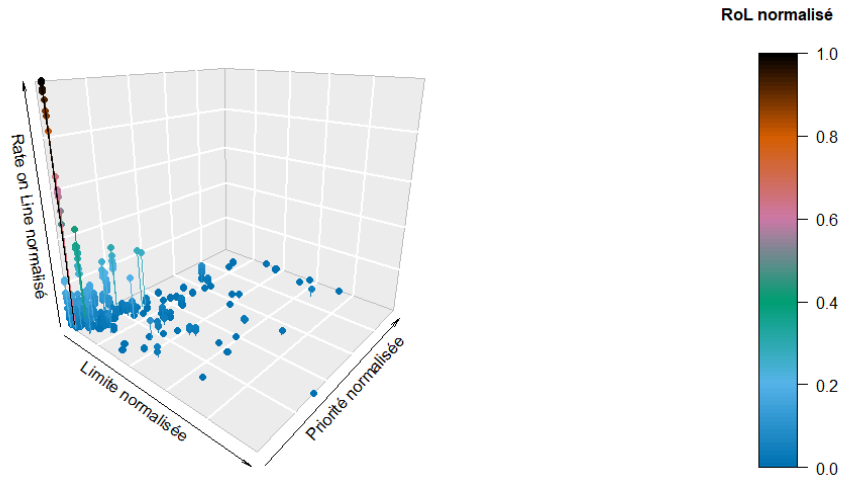


FIGURE 19 – RoL en fonction des tranches

Comme attendu, le RoL diminue en fonction de la hauteur de la tranche. Plus elle est faible (respectivement haute), plus le RoL est important (bas). Logiquement, la tranche est donc un bon indicateur sur l'ordre de grandeur d'une cotation d'un réassureur. Néanmoins, il existe une tendance différente entre risques. Celle-ci est difficilement modélisable sur un graphique en trois dimensions. C'est pourquoi pour illustrer ce phénomène, nous définissons un nouvel indicateur : le **RMP**. Le Relative Median Point (le point médian relatif) se calcule comme le point milieu de la tranche et est égal à Priorité $+ \frac{1}{2} \times$ Limite. Il permet de représenter la tranche d'un traité par une seule valeur. Analysons les cotations par risque :

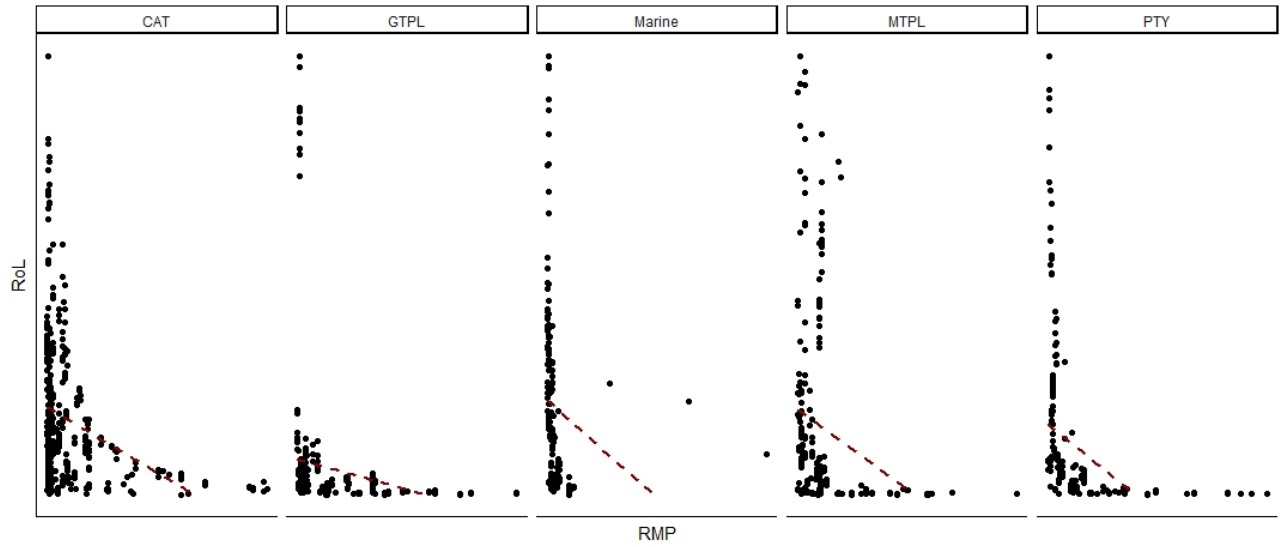


FIGURE 20 – RoL en fonction des tranches et par risque

Interprétation de la figure : La ligne rouge représente la régression linéaire du RoL en fonction du RMP, par risque.

La tendance de décroissance linéaire du RoL est identique pour tous les risques. Chaque risque possède cependant sa propre *vitesse* de décroissance. En Marine, trois traités possédant des RMP bien plus grands que les autres de ce risque se distinguent. De plus, pour un même RMP, deux branches ne sont pas nécessairement touchées de la même manière. Tout dépend des lois de sinistres sous-jacentes aux portefeuilles réassurés. Un traité CAT 1M [13] XS 10M n'est pas équivalent en termes de sinistralité à un traité 1M XS 10M en GTPL. Tout ceci nous permet de conclure que le RMP, par risque, est potentiellement une variable explicative de la cotation des réassureurs. Utilisons désormais ces cotations pour élaborer des modèles de tarification.

13. M = Million

3 Tarification et clustering des traités en XS

Table des matières

Dans ce chapitre, nous tenterons de tarifer nos traités de réassurance. Un tarif, une cotisation ou une prime est calculé par une mesure de risque. Celle-ci vérifie plusieurs propriétés permettant à l'actuaire, en fonction du contexte, de choisir sa façon de calculer la prime.

3.1 Notions autour de la prime

3.1.1 La prime d'assurance et de réassurance

La prime d'assurance correspond au montant payé par l'assuré pour que le bénéficiaire (qui peut être l'assuré, par exemple en RC automobile) puisse être indemnisé en cas de réalisation du risque assuré par le contrat d'assurance souscrit. Il est possible que plusieurs bénéficiaires soient concernés, comme en prévoyance avec des rentes éducation que les enfants peuvent percevoir pour financer leurs études en cas de décès d'un parent. De plus, le souscripteur du contrat peut aussi être différent de l'assuré notamment dans les contrats collectifs où l'entreprise souscrit pour ses salariés.

La prime d'assurance se décompose en plusieurs parties : la prime pure, les chargements et les frais. Plus précisément, si X représente le risque, alors $\mathbb{E}[X] = \int_{\mathbb{R}} x \, \mathrm{d}F(x)$ représente le risque moyen, c'est-à-dire la prime pure qui permet à l'assureur de faire face en moyenne au risque couvert par le contrat. Ajoutons les chargements, définis de plusieurs façons, notés ici $C(X)$. Aussi appelés la *marge pour risque*, ils sont utilisés pour prendre en compte la volatilité et ainsi créer une marge de sécurité permettant à l'assureur de se couvrir contre les sinistres dans $\alpha\%$ des cas (α à déterminer en fonction de l'appétit au risque de l'assureur). Le tarif n'est pas proposé tel quel sur le marché, il doit encore être majoré des frais de gestion, de courtage (s'il fait appel à un courtier/agent), diverses taxes et une marge de marché en fonction de la concurrence. La somme de ces éléments constitue la prime commerciale.

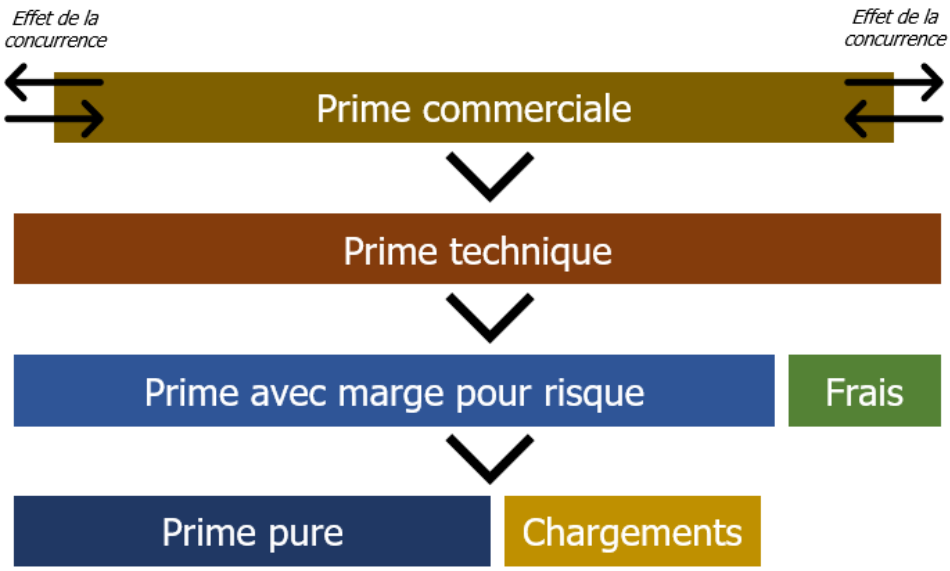


FIGURE 21 – Décomposition de la prime

La prime de réassurance est similaire à la prime d'assurance, elle est payée par la cédante au(x) réassureur(s). C'est cette prime qui nous intéresse ici en tant que cédante et donc acheteur de réassurance. Cependant, cette prime ne se calcule par nécessairement comme la prime d'assurance. Du fait de la spécificité de l'activité de réassurance, d'autres moyens comme le *burning cost* permettent au réassureur de proposer un tarif en pourcentage de l'assiette de prime des

portefeuilles réassurés. Ces méthodes permettent d'estimer le montant de sinistre à la charge du réassureur, en se basant sur les sinistres d'assurances de la cédante. Dans cette section, nous tenterons de tarifer nos traités de réassurance avec un principe de prime usuel tels que définis ci-après.

3.1.2 Les principes de calcul de prime

3.1.2.1 Définition

Le principe de prime est une fonction f qui à X (variable aléatoire) associe $f(X)$ telle que $f(X)$ soit un tarif satisfaisant pour l'assureur, c'est-à-dire qui ne le conduira probablement pas à la faillite s'il accepte d'assurer le risque X. Notons $f(X) = P(X)$. La valeur de $P(X)$ permet ainsi de mesurer le risque X. Si $P(X)$ est élevée, l'assureur demande un tarif élevé à l'assuré, car il considère que le risque est important. Inversement, si $P(X)$ est faible alors le risque est considéré comme faible et le tarif proposé est bas. Dans le cas extrême où $P(X) = \infty$, le risque est considéré comme inassurable.

3.1.2.2 Mesures de risque usuelles

Plusieurs fonctions permettent de calculer un tarif. Toutes ne respectent pas les mêmes propriétés. Énumérons en quelques-unes d'entre-elles :

— **Principe de l'espérance :**
$$P(X) = (1 + \beta)E[X]$$
— **Principe de la variance :**
$$P(X) = E[X] + \beta \operatorname{Var}[X]$$
— **Principe de l'écart-type :**
$$P(X) = E[X] + \beta\sqrt{\operatorname{Var}[X]}$$
— **Principe exponentiel :**
$$P(X) = \frac{1}{\beta} \ln E\left[e^{\beta X}\right], \quad \beta > 0$$
— **Principe de la Value at Risk :**
$$P(X) = VaR[X; \beta] = F_X^{-1}(\beta)$$
— **Principe de la Tail Value at Risk :**
$$P(X) = TVaR[X; \beta] = \frac{1}{1-\beta} \int_\beta^1 VaR[X; \alpha] d\alpha$$

Les propriétés principales de ces mesures de risque sont les suivantes :

Principe \ Propriété	Profitabilité	Pas de chargements injustifiés	Translation	Sous-additivité	Homogénéité
Espérance ($\beta \neq 0$)	✓			✓	✓
Variance	✓	✓	✓		
Ecart type	✓	✓	✓	✓	✓
Exponentiel	✓	✓	✓		
VaR		✓	✓		✓
TVaR	✓	✓	✓	✓	✓

TABLE 7 – Propriétés vérifiées par les mesures de risque

La définition exacte de ces propriétés est disponible en annexe 4.6. Désormais, intéressons-nous au choix de la mesure de risque la plus adéquate à notre problématique.

3.2 Objectifs

Dans la section 3, nous souhaitons estimer la prime commerciale des traités. Usuellement en assurance, la tarification est réalisée par celui qui crée le produit c'est-à-dire l'assureur lui-même. De sa conception à sa mise sur le marché, le produit d'assurance est le fruit de calculs actuariels ainsi que de choix financiers [14] et commerciaux [15]. Cet ensemble est mesuré par un seul et même acteur : l'assureur. Or, en réassurance, les rôles sont différents. La cédante émet un produit sans savoir à l'avance le tarif qui lui sera proposé. En fonction de l'appétit au risque ou encore du besoin en capital, la structure d'un traité en excédent de sinistre est déterminée par la cédante. Cependant, le tarif est lui estimé par le(s) réassureur(s) lors d'appels d'offres où chaque participant propose sa prime commerciale. Ainsi, cette fois le produit d'assurance (le traité) n'est pas uniquement le fruit des choix de la cédante mais de plusieurs parties qui entrent en jeu.

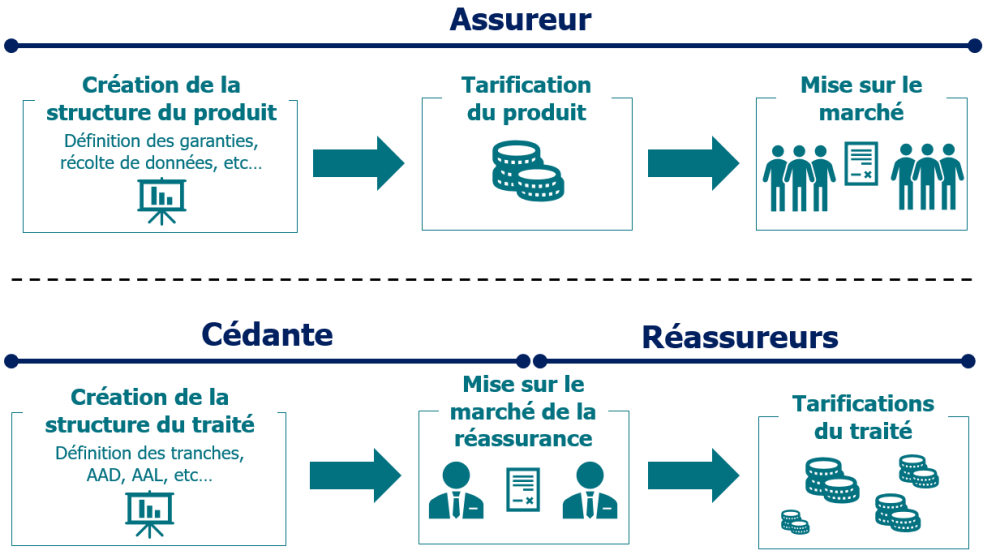


FIGURE 22 – Différences entre un produit de réassurance et d'assurance

Généralement en assurance, l'entreprise est seule du début à la fin de la chaîne de construction de son produit. En réassurance, cette chaîne est inversée : le traité est d'abord mis sur le marché avant d'être finalement tarifé et ce sont les réassureurs qui émettent le prix du contrat. La cédante se trouve dans la même position qu'un client potentiel à la différence près que c'est elle qui émet les garanties qu'elle souhaite avoir pour sa couverture.

Il peut donc être difficile pour un assureur (ou un réassureur en rétrocession) de savoir si la prime commerciale de réassurance cotée est réaliste, avantageuse ou au contraire trop onéreuse car il ne possède aucune vision sur le modèle de tarification interne des réassureurs. Pour y remédier, il peut tenter d'estimer lui même une prime de réassurance mais uniquement basée sur sa vision du risque. En pratique, l'assureur peut tout de même chercher à faire comme s'il était le seul acteur en tarifant le traité avec une mesure de risque usuelle qu'il choisit. Le tarif sera certainement différent des cotations mais cela permet tout de même d'estimer grossièrement la prime commerciale potentielle. Néanmoins, il est possible d'améliorer l'estimation des cotations. En effet, nous disposons d'une variable essentielle pour ajuster notre modèle : les cotations effectuées par les réassureurs. En d'autres termes, nous avons l'information de

14. Placements des cotisations sur les marchés
15. Rémunérations des salariés, agents et courtiers

la **prime commerciale de réassurance** pour plusieurs traités très différents. Cette variable permet à la cédante d'ajuster un modèle de tarification bien plus précis pour estimer sa prime de réassurance.

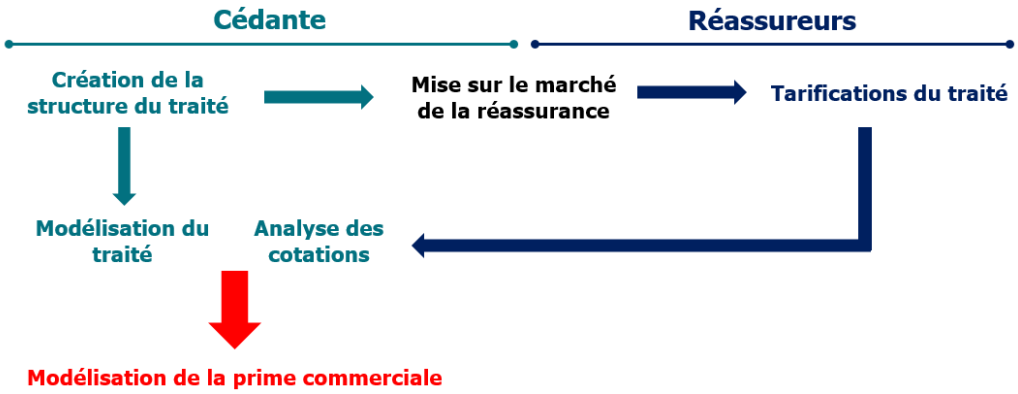


FIGURE 23 – Étapes de la modélisation de la prime commerciale de réassurance en vision interne

La modélisation interne de la prime commerciale peut apporter beaucoup d'informations à la cédante, comme :

1. **Évaluer les futures cotations :** lors de chaque année de renouvellement, la cédante peut modéliser en interne la prime commerciale du traité avant sa mise sur le marché de la réassurance. Une fois les cotations des réassureurs reçues, elle peut évaluer si elles paraissent, du point de vue de son modèle, chères, équilibrées ou bien faibles. Ainsi, elle peut mieux négocier la prime définitive avec les réassureurs puisqu'elle a une idée de l'ordre de grandeur que doit avoir ce tarif.
2. **Analyser la vision des réassureurs :** la cédante effectuant des modélisations internes, elle peut, par l'analyse des cotations, comparer celles-ci à ses récupérations modélisées. En effet, la différence entre les données internes et les cotations permet de mieux connaître l'évaluation que les réassureurs font du traité et plus généralement des portefeuilles réassurés de la cédante. Si les primes commerciales sont faibles comparées aux récupérations moyennes, il se peut que sa vision du risque soit plus forte que celle des réassureurs et inversement. Enfin, l'écart entre prime commerciale cotée et récupération moyenne permet d'estimer le chargement potentiel qu'appliquent les réassureurs à la prime pure.
3. **Tarifer un nouveau traité :** un des principaux avantages de la modélisation de la prime commerciale est que le produit de réassurance peut suivre les mêmes étapes que le produit d'assurance. Étant donné que la cédante dispose de toutes les informations nécessaires au calcul d'une prime commerciale, elle peut directement estimer son tarif avant l'intervention des réassureurs et donc faire *comme si* elle était le seul acteur dans la conception de son produit (le traité de réassurance), pareillement à l'assurance. Cette estimation n'est évidemment pas le tarif définitif. Néanmoins, elle apporte tout de même une part d'information non négligeable.
4. **Optimiser la réassurance :** la cédante peut hésiter entre plusieurs structures possibles pour un traité [16]. Elle peut estimer la prime commerciale par option en faisant varier les différents paramètres de son traité (AAD, ALL, reconstitutions, portée, priorité) et calculer une prime commerciale pour chacune d'entre-elles. Ainsi, elle dispose d'une idée des tarifs probables de ces options et il lui est possible de sélectionner seulement celles qui lui paraissent les plus optimales et les proposer sur le marché de la réassurance.

16. Les différentes structures possibles pour une même tranche sont appelées des options.

Pour évaluer ces différents points, nous utilisons principalement deux méthodes de tarification. Une dite par **chargements constants** et l'autre par **prédiction directe**. Le choix de la mesure de risque s'effectue en fonction des données disponibles et des propriétés que nous souhaitons vérifier. Nous disposons des récupérations moyennes ainsi que de l'écart type de celles-ci. Ainsi, naturellement, nous choisissons le **principe de l'écart type** comme mesure de risque. Afin d'obtenir une prime commerciale, il faut ajouter différents frais. En premier lieu, il convient d'estimer les frais de gestion du réassureur car il doit les ajouter à sa prime commerciale. Ces frais permettent de payer la gestion de l'entreprise, rémunérer les salariés, payer les locaux, etc. Généralement, ils sont exprimés en pourcentage de la prime. Pour les estimer, nous réalisons un *benchmark*[17] des différents rapports annuels publiés par nos partenaires de réassurance. Nous calculons le taux de frais de gestion comme les frais administratifs divisés par le chiffre d'affaires. Les résultats sont les suivants :

Réassureur	**Taux de frais de gestion en 2020**
1	4.91 %
2	1.93 %
3	5.73 %
4	4.90 %
5	1.90 %
⋮	⋮
n	2.60 %

TABLE 8 – Benchmark des frais de gestion des réassureurs

Le taux final choisit est le taux moyen soit **3.94%**. Les frais de courtage représentent le coût du courtier lorsqu'il vend un traité de réassurance. Ce taux est choisi sur la base des tarifs de courtage proposés à AXA Global Re. Nous pouvons finalement introduire notre formule de tarification de la prime commerciale :

$$\boxed{PC = \frac{\bar{R} + \beta \times \sigma_R}{Rec_{factor} \times (1 - \alpha)}}$$

où PC est la prime commerciale (ou la cotation des réassureurs en fonction de la vision), $\bar{R}$ la récupération moyenne, σ_R l'écart type des récupérations, β le coefficient de chargement, Rec_{factor} le facteur de reconstitution et α le taux de frais total.

L'objectif des parties suivantes est de trouver un lien entre modélisations internes et cotations externes. Un principe de prime par écart type basé sur l'estimation de β peut nous permettre d'y parvenir. Avant de commencer une quelconque modélisation, il convient toujours d'afficher un aperçu des données. Calculons en premier lieu le coefficient de chargement estimé, pour chaque ligne de notre base. Pour ce faire, nous inversons la formule de tarification pour isoler β :

$$\beta = \frac{PC \times Rec_{factor} \times (1 - \alpha) - \bar{R}}{\sigma_R}$$

Pour une même tranche d'un traité, il existe donc plusieurs taux de chargements dans le cas où plusieurs cotations sont proposées par différents réassureurs. Afin de ne pas donner trop d'importance aux traités ayant reçu de nombreuses cotations, nous décidons de travailler sur la cotation moyenne par tranche :

17. Étude de marché

Tranche	Réassureur	Cotation	Cotation moyenne
1	1	100	125
1	2	150	
2	4	200	200
3	2	100	137,5
3	3	150	
3	4	120	
3	5	180	

TABLE 9 – Exemple de calcul de cotation moyenne par tranche

Ce choix permet de ne pas pondérer l'estimation de β par le nombre de cotations. Cependant, l'information de volatilité de la cotation n'est plus présente et le nombre de points est diminué. Ci-dessous les nuages de points des β estimés [18], par risque.

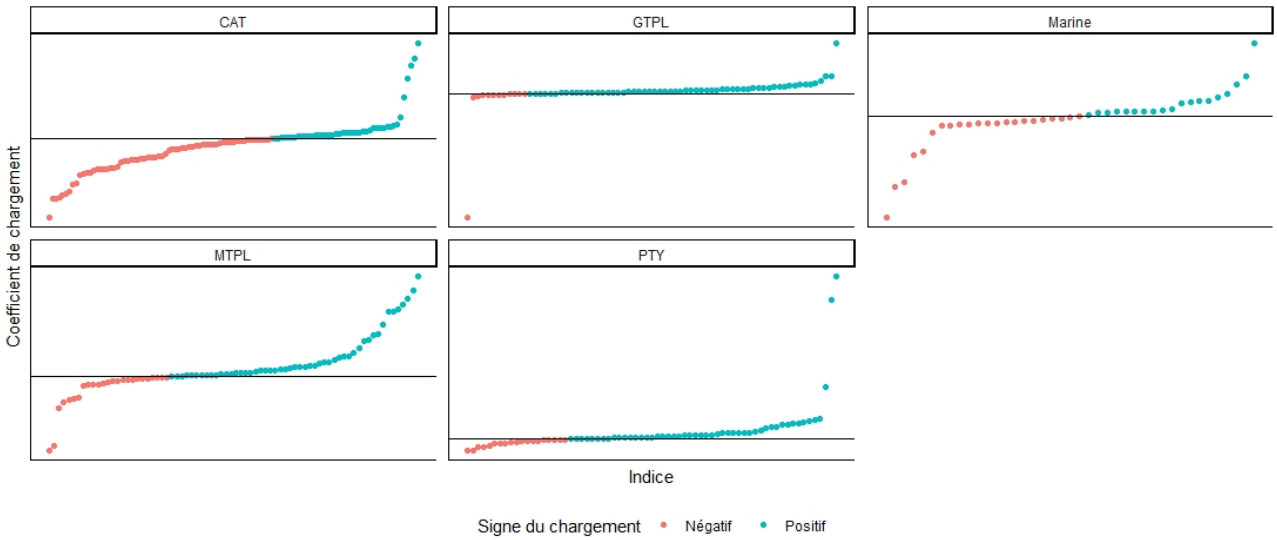


FIGURE 24 – β estimé pour chaque cotation par risque

Interprétation de la figure : les points rouges représentent les cas où le coefficient de chargement estimé est négatif ($\beta < 0$). Les points bleus sont ceux où $\beta \geq 0$. Par souci de confidentialité, les valeurs sur l'axe du coefficient de chargement ne sont pas affichées. La tendance des points est suffisante pour comprendre le graphique. Cependant, il est important de noter que les ordres de grandeurs de β ne sont pas nécessairement les mêmes par risque. De plus, l'échelle est différente pour chaque risque.

Nous remarquons ici que la totalité des risques présentent des chargements négatifs, notamment le risque CAT se démarquant nettement des autres par sa forte proportion de $\beta < 0$ (de l'ordre de 50 % environ). Intuitivement, ce cas ne semble pas être possible. En effet, un $\beta < 0$ revient à *décharger* la sinistralité moyenne en la diminuant de $\beta \times \sigma_R$. La prime commerciale est donc inférieure à la prime pure soit probablement un résultat négatif pour ces traités. Néanmoins dans notre cas, un $\beta < 0$ n'est pas insensé. Il faut rappeler que nous confrontons ici deux visions : une externe (réassureur) et une interne (cédante, AGRe). Alors, un $\beta < 0$ peut nous informer de plusieurs choses :

18. $\hat{\beta} = (Cotation\ moyenne \times Rec_{factor} \times (1-\alpha) - \bar{R})/\sigma_R$

— **La prime pure modélisée en interne est supérieure à la prime de réassurance moyenne cotée.** La modélisation interne des récupérations moyenne du traité est donc pessimiste comparée à la vision des réassureurs. Ceci peut être dût à des modèles différents entre cédante et réassureur.
— **Les réassureurs disposent d'un pouvoir de mutualisation et de diversification.** En effet, il est primordial de noter que nos modélisations se font traité par traité. Ainsi, il n'existe aucune forme de compensation de sinistralité entre eux. Nous calculons des primes pures individuelles, pour chaque traité. Parallèlement à l'assurance, cela revient à modéliser chaque assuré individuellement et donc à lui proposer une prime uniquement basée sur son profil sans prise en compte des autres assurés. Ainsi, si celui-ci présente un risque élevé, sa prime peut être très importante si aucun facteur de mutualisation ou de diversification n'est pris en compte. A contrario, lorsque les réassureurs cotent un de nos traités ils disposent en réalité d'une multitude d'autres traités leur permettant de bénéficier de l'effet de mutualisation et de diversification. Ainsi, les tarifs proposés sont bien plus faibles que s'ils ne réassuraient que notre traité. Il paraît donc sensé que des cas où notre prime pure est inférieure à la cotation moyenne surviennent.
— Enfin, les deux explications ci-dessus peuvent tout à fait se produire pour un seul et même traité, conduisant à des β fortement négatifs.

De nombreux cas où $\beta < 0$ sont visibles, signifiant donc que la cotation est supérieure à la prime pure. Les risques GTPL, Marine et MTPL partagent tous des points extrêmes dans la partie négative. Ces points sont éloignés des autres, ce qui est dû à une trop grande différence de vision entre AGRe et les réassureurs sur la modélisation des récupérations. Aussi, et ce pour tous les risques, des coefficients atypiques positifs sont présents. Les $\beta \geq 0$ en PTY sont les plus volatils. Plus généralement, tout point élevé et se démarquant nettement des autres est considéré comme aberrant. Pour les identifier précisément et ainsi créer des bornes inférieures et supérieures de valeurs acceptables, nous pouvons tracer le nuage de points de β tous risques confondus (graphique en annexe 76). Les bornes sont choisies visuellement, lorsque les points élevés semblent *décrocher* de la distribution générale des valeurs de β. Tous les points en dehors de ces bornes sont donc supprimés. Cela permet de pas nuire à la qualité des futurs modèles de β.

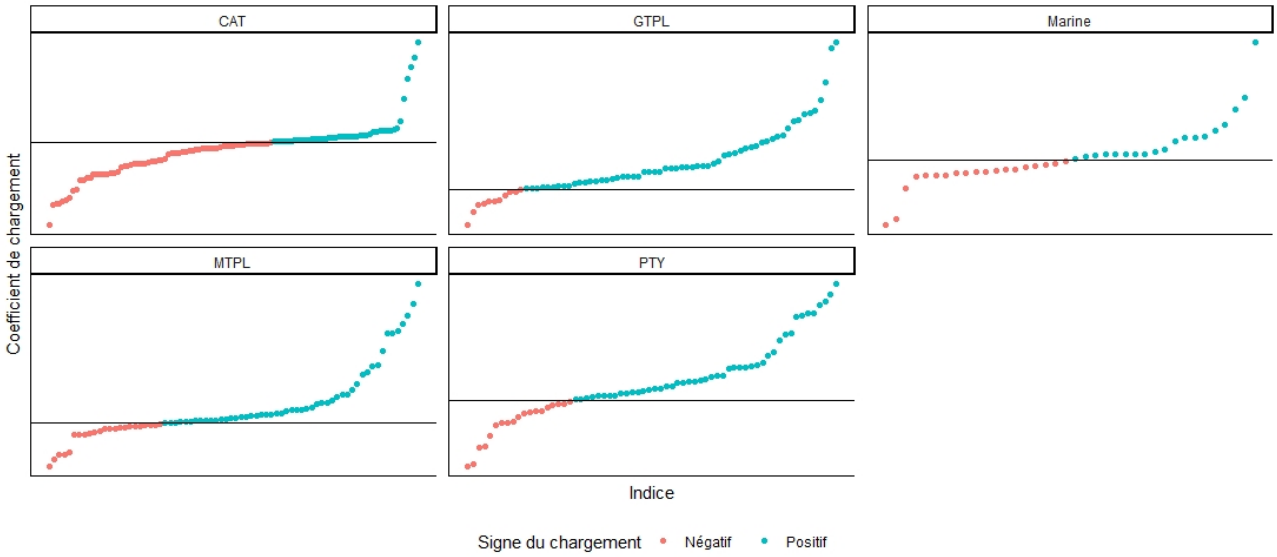


FIGURE 25 – β estimé pour chaque cotation par risque après nettoyage

Finalement, tous les risques ne présentent pas la même quantité de points. Le CAT est celui qui en contient le plus tandis que le Marine est celui en contenant le moins. Ceci nous rappelle l'importance d'un nettoyage approfondi et rigoureux de nos données. Le multiple, introduit précédemment, n'était pas un indicateur suffisant pour créer une base de données cohérente pour cette partie. Plus généralement, dès qu'une nouvelle variable est créée, un bref aperçu par nuage de points est un minimum obligatoire pour repérer les cas atypiques et aberrants.

Dans les parties qui suivent, nous aborderons des méthodes de prédiction paramétrique (GLM) et non paramétriques (machine learning supervisé). De plus, nous utiliserons des méthodes de groupage (machine learning non supervisé). Ainsi, avant d'expliciter la première méthode, il convient d'introduire les principes de ces algorithmes de prédiction et de groupage (aussi appelés algorithmes de *clustering*).

3.3 Le modèle linéaire généralisé (GLM)

Représentons nos données par la matrice suivante :

$$\begin{bmatrix} X_{1,1} & X_{1,2} & \dots & X_{1,n} \\ \vdots & \ddots & & \vdots \\ X_{n,1} & X_{n,2} & \dots & X_{n,n} \end{bmatrix} \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$$

Où X est la matrice des $(X_{i,j})_{1,\dots,n}$ appelés **descripteurs** et Y le vecteur des $(Y_i)_{1,\dots,n}$ appelés **labels**. Plus clairement, le label Y est le vecteur que nous cherchons à prédire en se basant sur les informations que sont les descripteurs X. Par exemple, Y peut être des montants d'accidents automobiles et X les caractéristiques des voitures et des conducteurs. Ces informations peuvent être qualitatives ou quantitatives. Le GLM est un modèle cherchant à estimer Y de la façon suivante :

$$\mathbb{E}\left(Y_i \mid X_i\right) = \mu_i = g^{-1}\left(X_i'\beta + \xi_i\right),$$

où g est la fonction de lien, ξ est l'*offset*[19] et $X'\beta$ est le prédicteur linéaire. En fonction du type du label, la loi potentielle de $Y|X$ est différente :

Type du label	Intervalle	Loi possible
Continu	$(-\infty, \infty)$	Normale (μ, σ^2)
Binaire	$\{0, 1\}$	Bernoulli (p)
Comptage	$\{0, 1, \dots, n\}$	Binomiale (n, p)
Comptage	$\{0, 1, \dots\}$	Poisson (λ)
Continu	$(0, \infty)$	Gamma (α, β)
Continu	$(0, 1)$	Beta (α, β)

TABLE 10 – Lois du label Y

Pour plus de clarté, posons $\xi = 0$ (les individus sont observés avec une durée égale). Alors, le GLM modélise l'espérance conditionnelle

$$\mu_i = \mathbb{E}\left(Y_i \mid X_i\right), \qquad \text{avec } g(\mu_i) = \left(X_i'\beta\right).$$

Habituellement, la notation $\eta_i = \left(X_i'\beta\right)$ est utilisée. En pratique, plusieurs fonctions de liens sont utilisées :

19. L'offset est un terme de décalage souvent utilisé pour corriger l'exposition de temps différente entre les individus.

Type de lien	$g(\mu_i)$
Identité	μ_i
Inverse	$\frac{1}{\mu_i}$
Log	$\log(\mu_i)$
Logit	$\log\left(\frac{\mu_i}{1-\mu_i}\right)$
Probit	$\text{probit}(\mu_i)$

TABLE 11 – Fonctions de liens $g(.)$

Chaque famille de lois possède une fonction de lien dite *canonique* permettant de relier leur(s) paramètre(s) à l'espérance μ. Par exemple, pour la loi Normale, la fonction canonique est le lien identité tandis que pour la loi de Bernoulli c'est le lien logit. L'estimation des paramètres des lois se fait par maximum de vraisemblance.

3.3.1 Exemple de la régression de Poisson

Nous considérons un GLM Poisson où les descripteurs sont des variables numériques. Ce modèle s'applique quand les Y_i sont des variables de comptage. Nous rappelons que la loi de Poisson $P(\lambda)$ est définie par la distribution suivante :

$$P(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!}$$

Ainsi, nous cherchons à estimer le paramètre λ grâce à la matrice X :

$$\begin{aligned}\lambda_i &= g^{-1}(\beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_n X_{i,n}) \\ &= g^{-1}(\eta_i)\end{aligned}$$

Finalement, η_i est le produit scalaire entre le vecteur des descripteurs et le vecteur des paramètres β :

$$\eta_i = (X_{i,0}\ ,\ldots,\ X_{i,n}) \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_n \end{pmatrix} =< X_{i,.}, \beta >$$

Par convention, $X_{i,0} = 1$. La fonction de lien doit être inversible et différentiable. De plus, pour le modèle Poisson, g est la fonction log. Ainsi,

$$\log \lambda_i = \eta_i =< X_{i,.}, \beta > \quad \Longrightarrow \quad Y_i \sim P(\exp < X_{i,.}, \beta >).$$

Nous avons donc la relation suivante :

$$\lambda_i = \mathbb{E}(Y_i \mid X_i) = \text{Var}(Y_i \mid X_i) = \exp(\beta_0 + \beta_1 X_{.,1} + \beta_2 X_{.,2} + \cdots + \beta_n X_{.,n}).$$

Écrivons la log vraisemblance associée à maximiser en fonction de β :

$$\begin{aligned}\log \mathcal{L}(\beta; y) &= \sum_{i=1}^{n} \left[y_i \log(\lambda_i) - \lambda_i - \log(y_i!)\right] \\ &= \sum_{i=1}^{n} y_i \cdot [X_i'\beta] - \exp[X_i'\beta] - \log(y_i!)\end{aligned}$$

Ainsi, en dérivant par rapport à β :

$$\frac{\partial \log \mathcal{L}(\beta; y)}{\partial \beta} = \sum_{i=1}^{n} X_i' \left(y_i - \exp[X_i'\beta]\right)$$

Il n'existe pas de solution explicite de $\hat{\beta}$, il est donc déterminé numériquement.

3.3.2 Choix du modèle

L'importance des variables X est prépondérante dans la qualité du modèle. Pour une même famille de lois (ie pour la même loi de Y), l'estimation peut être très différente en fonction des $X_{.,i}$ choisis. Plusieurs indicateurs permettent de juger la qualité d'un modèle et de le comparer aux autres candidats :

1. $\mathbf{R^2}$: le coefficient de détermination linéaire de Pearson mesure la qualité de la régression. Compris entre 0 (modèle faible) et 1 (modèle parfait), il mesure la part de variance expliquée par le modèle sur la variance totale. Il est égal à

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

   où y_i est la i$^{\text{ème}}$ valeur observée du label Y, $\bar{y}$ la moyenne des y_i et $\hat{y}_i$ l'estimation de y_i par le modèle.

2. **AIC :** le critère d'information d'Akaike permet de pénaliser les modèles en fonction du nombre de paramètres. Le modèle disposant de l'AIC le plus faible est choisi. Ce critère est égal à :

$$AIC = -2 \log \mathcal{L} + 2k$$

   avec k le nombre de variables.

3. **BIC :** le critère d'information bayésien est un dérivé de l'AIC et est égal à :

$$BIC = -2 \log \mathcal{L} + k \log(n)$$

   avec k le nombre de variables et n la taille de l'échantillon. À la différence de l'AIC, ce critère prend en compte la taille de l'échantillon sur lequel le modèle est calculé. Nous cherchons aussi à le minimiser.

Finalement, si la loi sous-jacente des Y est connue, le GLM est un bon candidat pour prédire notre label. Cependant, l'hypothèse forte réside dans le fait que ce modèle est paramétrique. Ainsi, une bonne estimation du paramètre naturel de la loi des Y est primordiale. Il existe une autre famille de modèles dits non paramétriques basés sur des calculs empiriques autour des données. Ils présentent l'avantage de proposer des hypothèses moins fortes mais sont sujets à l'effet *boîte noire*[20]. Introduisons ces approches.

3.4 Le Machine Learning

L'Apprentissage Statistique (Machine Learning) est un ensemble de méthodes d'Intelligence Artificielle axées sur des algorithmes qui apprennent en fonction des données fournies. Trouvant son origine dans la moitié du XX$^{\text{ème}}$ siècle, ces algorithmes sont capables d'accumuler un grand nombre de connaissances basées sur l'expérience des données sans qu'aucune indication humaine ne leur soit donnée, tout le processus est automatique. Cette discipline est assez récente et son développement ne cesse de croître. Elle est utilisée dans de nombreux domaines de la vie courante (médecine, robotique ou encore finance). L'arrivée du Big Data est un vrai plus car l'émergence de quantité massive de données a permis à ces méthodes de se perfectionner. De plus, nos ordinateurs étant de plus en plus performants, le temps de calcul de ces algorithmes s'en voit réduit. Le lien avec la théorie non paramétrique de la Statistique est fort. Le Machine

20. L'effet boîte noire se caractérise par le manque d'explicabilité des algorithmes.

Learning est ainsi une discipline à la croisée des chemins entre informatique et mathématique.

Le point de départ de l'Apprentissage Statistique est, comme pour le GLM, un échantillon de données. Notons $Z_i = (X_i, Y_i)$ les observations dont nous disposons. Nous regroupons ces algorithmes en deux types :

— **Apprentissage supervisé :** les observations $Z_i = (X_i, Y_i)$ sont composées de descripteurs $X_i \in \mathbb{R}^d, d \in \mathbb{N}$, et d'un label Y_i (parfois appelé étiquette) appartenant à $\mathbb{R}$ ou à un ensemble fini. L'objectif est de prédire, pour un nouveau $x \in \mathbb{R}^d$, la valeur de son label $\hat{y}$. Par exemple en médecine, nous pouvons donner en entrée une liste de radios présentant ou non des tumeurs. Le but sera alors de prédire, pour une nouvelle radio d'un patient, le risque de développer une tumeur. C'est par exemple le but d'une IA du MIT utilisée pour prédire le cancer du sein jusqu'à quatre ans avant que les tumeurs ne soient visibles à l'imagerie [21].

— **Apprentissage non-supervisé :** les observations $Z_i = (X_i)$ ne disposent pas de label. Dans ce cas, le but est de trouver des *clusters* (groupes) plus ou moins homogènes des Z_i. L'exemple classique est celui de la voiture et de la moto. Nous donnons en entrée plusieurs images (voitures ou motos) et l'algorithme élabore des clusters en fonction du type de véhicule. Il va apprendre à différencier une voiture d'une moto sans qu'aucune indication humaine ne lui soit donnée (quatre roues, deux roues, nombre de phares, ...), il reconnaît seul les différences.

Dans un premier temps, concentrons-nous sur l'apprentissage supervisé.

3.4.1 L'Apprentissage supervisé

Pour rappel, nous observons une base de données composée de n couples $Z_i = (X_i, Y_i)$ que nous supposons être des réalisations indépendantes d'une même loi P inconnue. L'espace des descripteurs noté $\mathcal{X}$ (ie $X \in \mathcal{X}$) est généralement égal à $\mathbb{R}^d$ où $d \in \mathbb{N}$. Les Y_i appartiennent à $\mathcal{Y}$ égal à $\mathbb{R}$ ou à un sous-ensemble de $\mathbb{R}$ ou encore à un ensemble fini. Le but est de prévoir le label $\hat{Y}$ associé à une nouvelle entrée X, en supposant la paire (X, Y) étant des réalisations de la loi P et indépendantes des observations précédentes.

Définissons alors :

1. L'ensemble des données $\mathcal{D}$:

$$\mathcal{D} = \mathcal{D}_n = \{(X_1, Y_1) \mid \cdots \mid (X_n, Y_n)\}$$

   où $(X_i, Y_i) \underset{iid}{\sim} P$.

2. Le **prédicteur :**

$$f : \mathbb{R}^d \longrightarrow \mathcal{Y}.$$

3. La **prédiction :**

$$\hat{Y} = f(X).$$

   Lorsque $\mathcal{Y} = \{0, 1, \ldots, k\}$ nous sommes en classification (classification binaire si $\mathcal{Y} = \{0, 1\}$) tandis que si $\mathcal{Y} = \mathbb{R}$ nous sommes en régression.

21. Lien vers l'article

La qualité de l'estimation est calculée par le **risque de prédiction**. Aussi nommé le coût de prédiction, celui-ci est composé de la fonction $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ qui s'applique à l'observation Y et à sa prédiction $\hat{Y} = f(X)$. Par exemple, en classification binaire nous parlons de perte *0-1* et l est définie par

$$l(Y, \hat{Y}) = \mathbb{1}\left\{Y \neq \hat{Y}\right\} = \begin{cases} 1 \text{ si } Y \neq \hat{Y} \\ 0 \text{ sinon} \end{cases}$$

En régression, nous pouvons utiliser la perte L^2 définie sur $\mathbb{R}^d$ par

$$l(Y, \hat{Y}) = \|Y - \hat{Y}\|_2^2.$$

Alors, le risque de prédiction R du prédicteur f est défini comme l'espérance mathématique de l :

$$\begin{aligned} R(f) &= \mathbb{E}_{(X,Y)\sim P}[l(Y, \hat{Y})] \\ &= \mathbb{E}_{(X,Y)\sim P}[l(Y, f(X))]. \end{aligned}$$

Ainsi le prédicteur f^* est optimal si $f^* = \underset{f \in \mathcal{F}}{\operatorname{argmin}}\, R(f)$. Le prédicteur optimal est appelé fonction **oracle**. Il est possible de trouver la forme explicite de f^*. Lorsque nous considérons la perte quadratique, c'est-à-dire quand $R(f) = \mathbb{E}[(Y - f(X))^2]$,

$$f^* = \mathbb{E}[Y \mid X].$$

Cet oracle est appelé le **régresseur de Bayes** avec $R^* = \mathbb{E}[(Y - \mathbb{E}[Y|X])^2]$. Généralement, le but de chaque algorithme est de minimiser le risque empirique $\hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n}(l(Y_i, f(X_i))$. Il est d'usage de poser $\eta(X) = \mathbb{E}[X \mid Y]$ puis de l'intégrer dans f^*. C'est la méthode du *plug-in*. Par la suite, l'objectif est d'estimer η par $\hat{\eta}$:

— En régression, avec perte quadratique :

$$f^*(X) = \eta(X).$$

— En classification binaire, avec perte 0-1 :

$$f^*(X) = \mathbb{1}\left\{\eta(X) \geq \frac{1}{2}\right\}.$$

Introduisons dès maintenant les algorithmes qui seront utilisés dans les prochaines parties : les KNN, les arbres de décision et les forêts aléatoires. Nous nous focaliserons uniquement sur ces algorithmes car ils font partie des méthodes les moins sujettes à l'effet boîte noire. Cependant, leur performance peut tout de même se révéler très convaincante (notamment pour les forêts aléatoires).

3.4.1.1 KNN (K-plus proches voisins)

L'algorithme des K-plus proches voisins ou K-nearest neighbors (KNN) appartient à la classe des méthodes dites à partitions. Cette approche est très populaire car sa conception est naturelle et simple à mettre en œuvre. De plus, son temps de calcul est très faible comparé à des méthodes plus complexes. Son principe est simple : un nouveau point X a un label estimé $\hat{Y}$ qui ressemble à ses voisins.

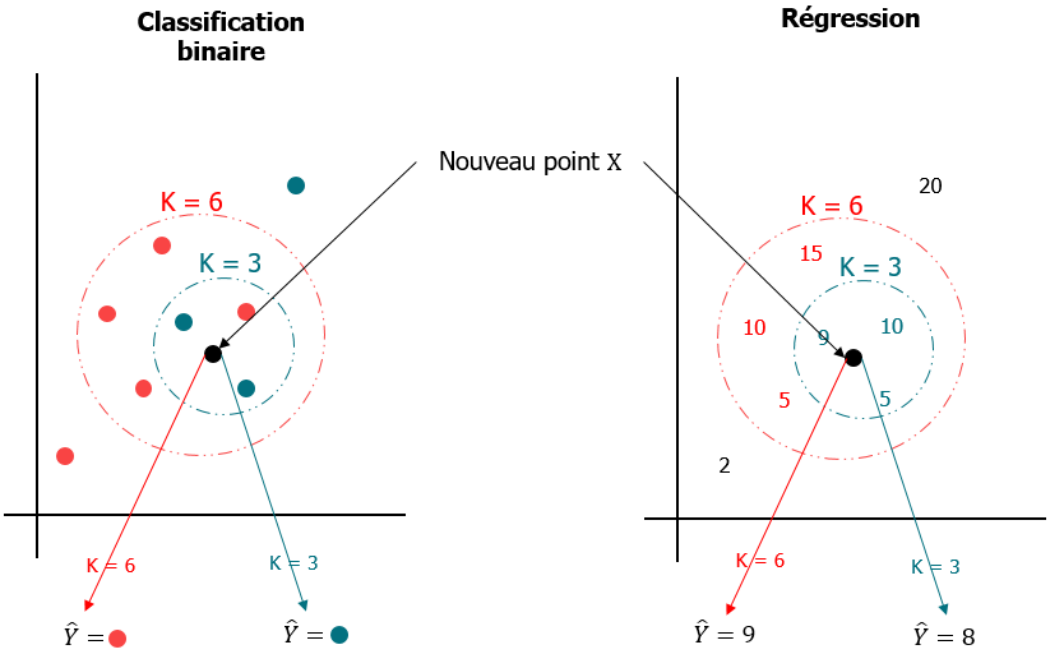


FIGURE 26 – Fonctionnement du KNN

En classification binaire nous posons $\mathcal{Y} = \{0, 1\}$ et $\mathcal{Y} \subset \mathbb{R}$ en régression. La notion de voisin se comprend au sens de la distance qui sépare les points entre eux. Pour simplifier, nous supposons que la distance d entre deux points X_i et X_j est la distance euclidienne :

$$\begin{aligned} d(X_i, X_j) &= \|X_i - X_j\|_2 \\ &= \sqrt{\sum_{k=1}^{n} (X_{i,k} - X_{j,k})^2} \end{aligned}$$

où $X_{i,k}$ est la valeur de la k$^{\text{ème}}$ colonne de l'individu i. Ainsi, pour un nouveau point X, son plus proche voisin est le X_i ayant la distance la plus faible avec lui. Posons V_k l'ensemble des k X_i les plus proches de X. Alors le prédicteur empirique optimal est égal à :

— **Régression :**

$$\hat{f}^*(X) = \hat{\eta}^{KNN}(X) = \frac{1}{k} \sum_{i:X_i \in V_k} Y_i$$

— **Classification binaire :**

$$\hat{f}^*(X) = \mathbb{1}\left\{\hat{\eta}^{KNN}(X) \geq \frac{1}{2}\right\}$$

Lorsque que nous sommes en classification générale ($\mathcal{Y} \subset \mathbb{N}$) le prédicteur choisit le label majoritaire dans V_k.

3.4.1.2 CART (arbre de décision)

L'algorithme CART pour Classification And Regression Trees est aussi un modèle à partitions basé sur un arbre de décision.

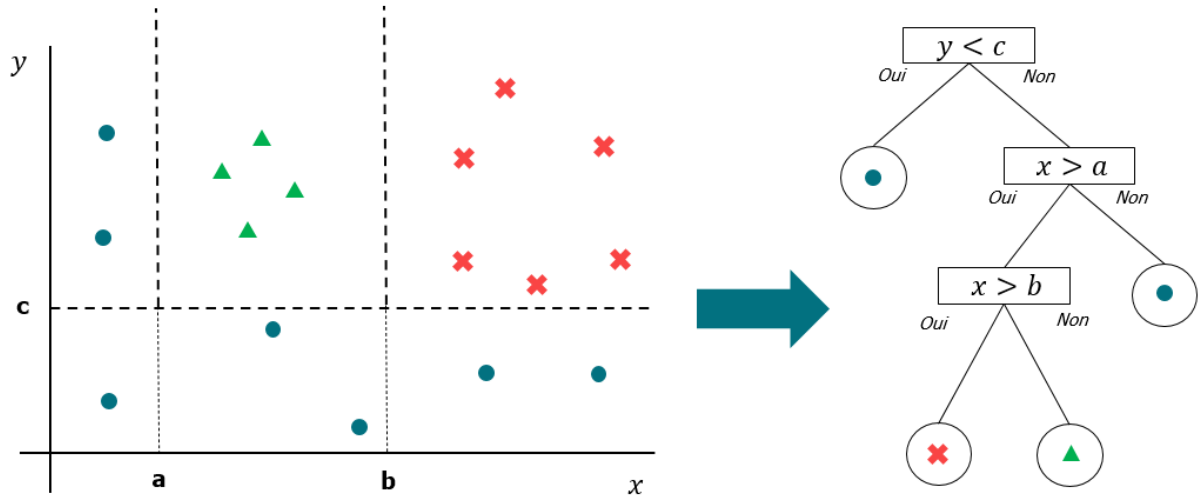


FIGURE 27 – Fonctionnement du CART

Ici, l'algorithme tente de créer des classes homogènes avec des seuils (lignes en pointillés sur le schéma). Un seuil se représente comme un nœud dans un arbre, c'est une paire d'un sous-ensemble A de $\mathcal{X}$ et d'un test T (appelé critère de segmentation) auquel on soumet les variables explicatives $X \in \mathcal{X}$. Si le test peut donner lieu à K résultats différents, c'est-à-dire $T : A \to \{1, \ldots, K\}$, alors le nœud correspondant à (A, T) donne naissance à k nœuds-fils, tel que l'ensemble $A^{(k)}$ associé au $k^{\text{ème}}$ fils est $A^{(k)} = \{X \in A : T(X) = k\}$. Ainsi, un arbre de décision peut être vu comme une fonction récursive initialisée à la racine où $A = \mathcal{X}$. Nous pouvons décrire ce processus par un algorithme contenant deux phases : une d'expansion et une d'élagage.

Algorithm 1 CART

Arbre ← *Racine*
for $N \in$ *Arbre* **do**
  if N vérifie la condition d'arrêt **then**
    $T \leftarrow$ critère de segmentation
    $(A, T) \leftarrow$ application de T à A
    Arbre ← *Arbre* + (A, T) *(Arbre = Arbre + noeuds fils)*
  end if
end for
for $N \in$ *Arbre* **do**
  if N vérifie la condition d'élagage **then**
    $M \leftarrow$ N + descendants
    Arbre ← *Arbre* − M
  end if
end for
return *Arbre*

Le critère d'arrêt consiste généralement à vérifier que l'une de ces trois conditions est vérifiée :

— la profondeur de l'arbre dépasse un certain seuil.
— le nombre de feuilles dans l'arbre dépasse un certain seuil.
— l'effectif du nœud est inférieur à un certain seuil.

Le critère d'élagage permet de supprimer une branche lorsque ce procédé ne nuit pas significativement à la qualité de l'arbre. Par qualité nous entendons par exemple le risque empirique.

Dans l'optique de trouver le label $\hat{Y}$ d'un nouvel X issu d'une nouvelle base de données, nous faisons parcourir l'arbre à X jusqu'à arriver à une feuille. En classification, $\hat{Y}$ sera alors égal au label majoritaire des Y de la feuille qui était présent lors de la phase de construction. En régression, $\hat{Y}$ sera la moyenne de ces Y. Présentons maintenant une autre méthode qui compile les CART : les forêts aléatoires (Random Forests).

3.4.1.3 Forêts aléatoires

Un arbre de décision seul est souvent biaisé et peu robuste s'il rencontre des valeurs spéciales telles que des outliers. Pour y remédier les forêts aléatoires ont été créées. Comme leur nom l'indique, elles sont composées de plusieurs arbres de décision construits aléatoirement par une méthode dite de **bagging**. Proposé en 2001 par Leo Breiman et Adèle Cutler, cet algorithme est entraîné sur des sous-ensembles aléatoires différents afin de créer de multiples arbres. Le nombre d'arbres de la forêt peut varier de plusieurs centaines à des milliers.

La méthode du bagging consiste à extraire K sous-ensembles $\mathcal{D}_n^1, \ldots, \mathcal{D}_n^K$ de $\mathcal{D}_n$. Chaque $\mathcal{D}_n^i$ est construit par un tirage aléatoire avec remise dans $\mathcal{D}_n$, cette approche se nomme le **bootstraping**. Nous construisons alors un arbre pour chaque sous-ensemble. Ainsi, nous disposons de K prédicteurs de la forme $\hat{f}^k(X, \mathcal{D}_n^k)$.

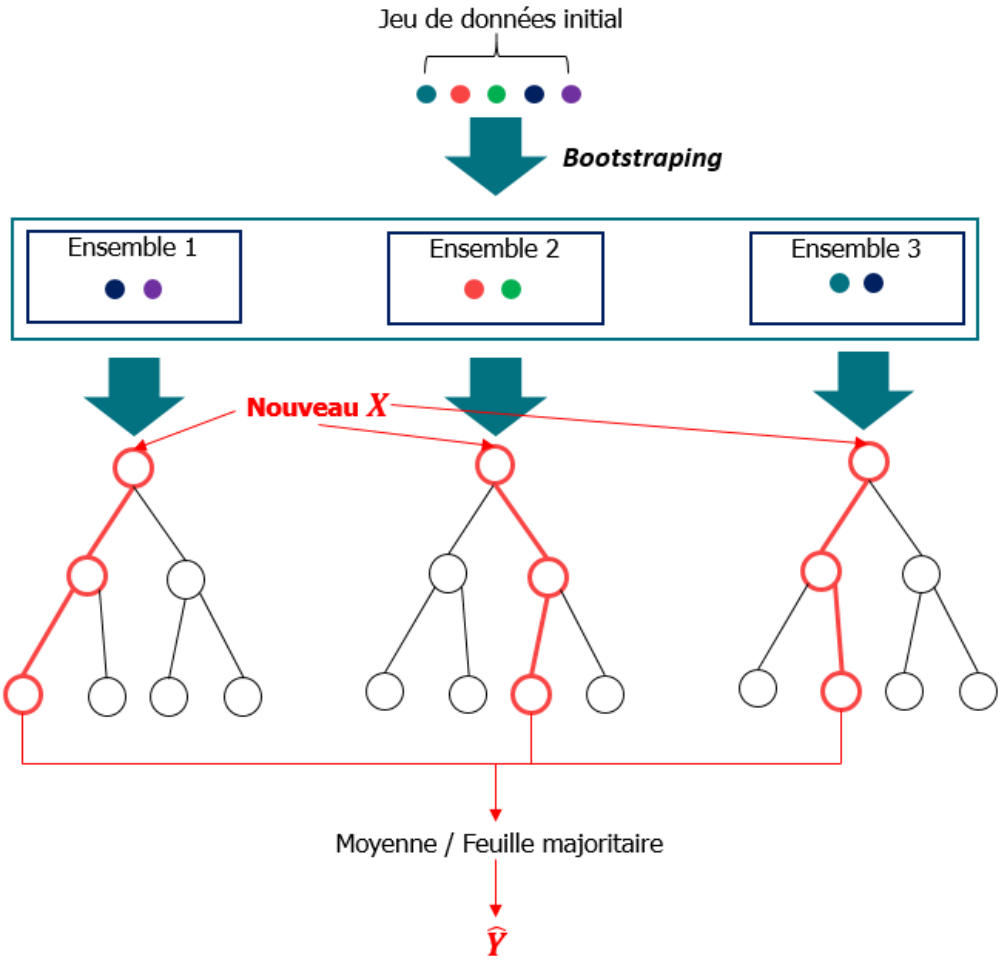


FIGURE 28 – Fonctionnement des forêts aléatoires

Si le nombre d'arbres est faible et n grand (nombre d'individus), il existe une probabilité forte qu'un individu ne soit pas contenu dans les $\mathcal{D}_n^i$, nous nommons ce phénomène l'**erreur o.o.b** pour out of bag. Le nombre de colonnes m de chaque $\mathcal{D}_n^i$, c'est-à-dire le nombre de variables à choisir lors du bootstraping, est souvent égal à $\sqrt{d}$ en classification et $\lfloor d/2 \rfloor$ en régression, où d est le nombre total de variables des descripteurs. Lorsque l'on souhaite prédire le label $\hat{Y}$ d'un nouveau X, le processus est identique aux CART. Nous faisons parcourir à X tous les arbres de la forêt. $\hat{Y}$ vaut alors la moyenne des feuilles atteintes par X de chaque arbre en régression. En classification, $\hat{Y}$ est égal à la feuille majoritaire.

3.4.1.4 Optimisation des algorithmes

L'étape cruciale dans le machine learning supervisé est le choix des valeurs des paramètres des algorithmes : le nombre de voisins, la profondeur d'un arbre, le nombre d'arbres d'une forêt, etc. Prenons l'exemple des KNN. Il est aisé de comprendre les impacts d'un choix de $k = 1$: le voisin le plus proche d'un nouveau point est choisi. Cependant, le résultat peut être très différent pour $k = 2$ si le deuxième voisin est éloigné du premier, la variance de la prédiction est très forte. Dans le cas où $k = n$ c'est-à-dire quand, pour tous les nouveaux points, le label prédit est égal à la moyenne (en régression) de tous les points du jeu de données initial. Pour ce k, le choix d'un algorithme de prédiction n'est même plus utile puisque que la prédiction est une fonction constante déjà connue. Nous comprenons donc bien que les choix extrêmes de k ne sont pas cohérents. Cependant, une fois ces valeurs écartées, quelle valeur optimale de k doit-on choisir ?

Une méthode assez naturelle est de tester un grand nombre de k et de choisir celui qui minimise le risque empirique $\hat{R}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n} l(Y_i, \hat{f}(X_i))$. Cette approche est appelée **ERM** pour Error Risk Minimization. Néanmoins, il existe un nouveau risque si nous procédons uniquement de cette façon : l'**overfitting**.

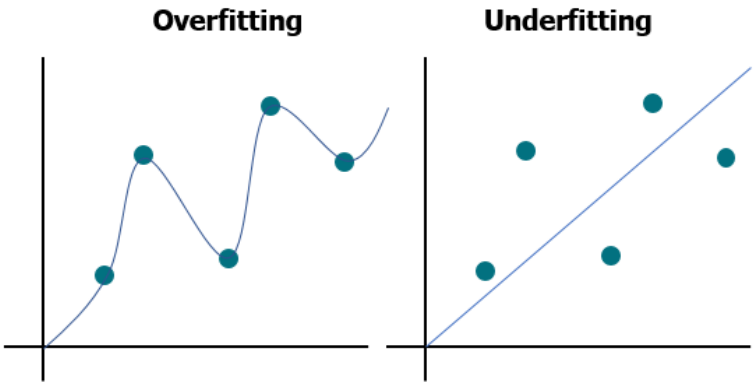


FIGURE 29 – Overfitting et underfitting

Ce phénomène se produit lorsque l'algorithme *sur-apprend* sur les données. Ainsi, la prédiction sera très efficace pour prédire exactement les mêmes labels que ceux du jeu de données d'entraînement. Cependant, toutes les nouvelles données seront mal prédites car le prédicteur ne sait pas s'adapter à un jeu de données différent. A contrario, l'underfitting est le cas inverse où l'algorithme ne prend pas assez en compte la base d'entraînement pour construire ses prédictions. Pour contrer ces effets, il est d'usage d'utiliser la **cross validation**.

La validation croisée (cross validation) est une méthode d'échantillonnage très populaire en machine learning. Son utilité réside dans l'indépendance des échantillons de validations et d'entraînement. Le but est d'effectuer l'apprentissage sur un sous-échantillon du jeu de données permettant ainsi d'estimer le prédicteur. Ensuite, le risque empirique est calculé sur un autre sous-échantillon (échantillon de validation). Nous ne détaillerons que la **K-fold cross validation**, qui peut se résumer au schéma suivant :

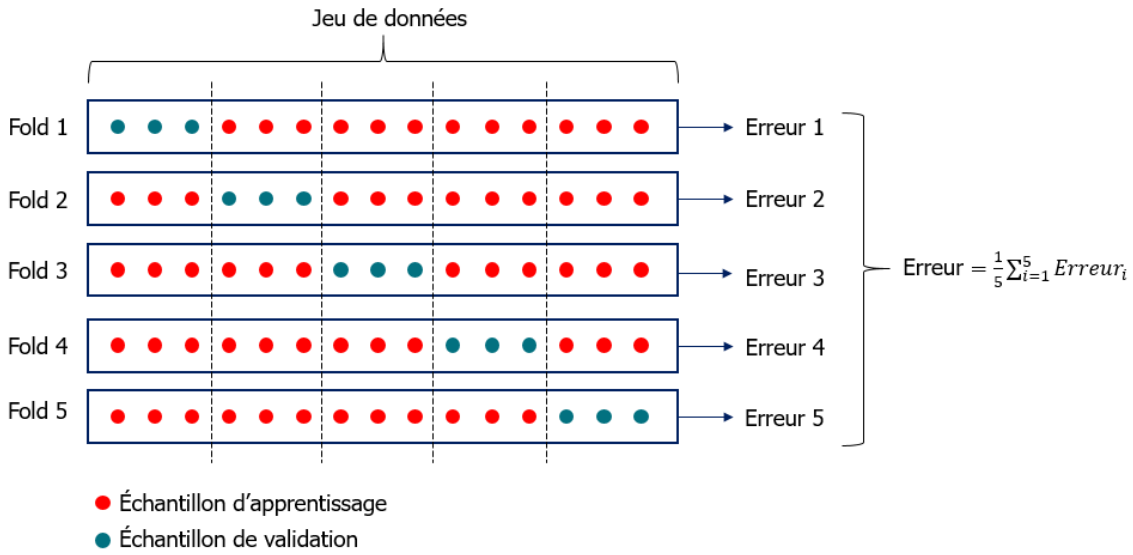


FIGURE 30 – Exemple de 5-fold cross validation

Dans cet exemple, notre jeu de données est divisé en 5 sous-groupes égaux (de taille 3). Le modèle apprend sur $4/5^{\text{ème}}$ des données et calcule ses prédictions sur $1/5^{\text{ème}}$ de l'échantillon. Finalement, l'erreur du modèle est la moyenne des erreurs calculées dans chaque *fold*. Ainsi, pour chaque valeur d'un paramètre, cette opération est répétée. Le paramètre optimal est celui qui minimise l'erreur empirique moyenne. Toutes ces étapes permettent de *valider* le modèle avec des paramètres optimaux. L'étape finale se trouve dans le calcul de la précision du modèle pour le valider définitivement.

Usuellement, nous coupons notre jeu de données en une partie réservée à l'estimation du modèle (généralement 75 % des données) et une autre pour calculer la précision (généralement 25 % des données). La partie estimation du modèle sert à estimer les paramètres avec K-fold cross validation tandis que l'échantillon de test est réservé à une estimation finale de la précision du prédicteur. Ceci marque la fin de la modélisation. Toutes ces étapes peuvent se résumer ainsi :

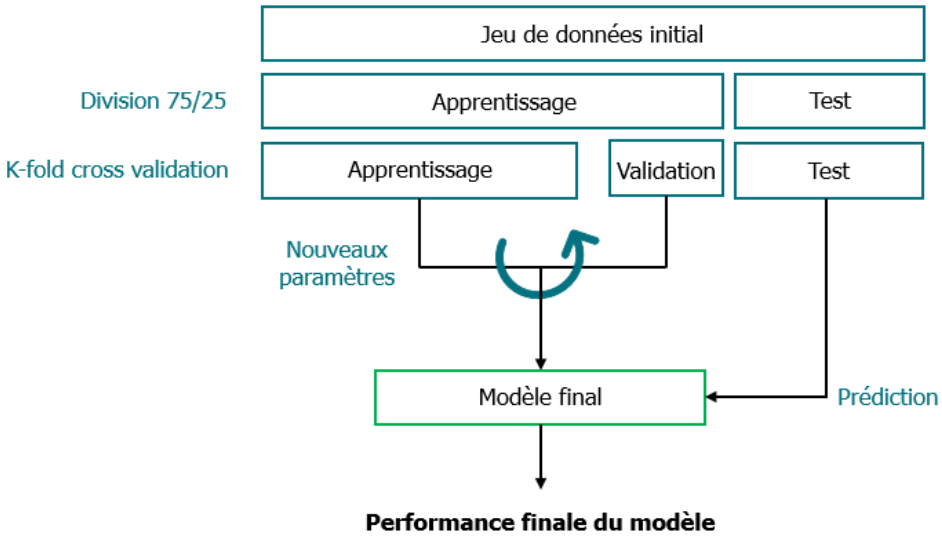


FIGURE 31 – Processus de construction d'un algorithme de machine learning supervisé

Une fois la validation croisée effectuée ainsi que les paramètres optimaux trouvés, nous utilisons le modèle final pour prédire les labels sur une base test issue du jeu de données initial. Le résultat de l'estimation sur cette base est considéré comme la performance finale du modèle. Le dernier choix important est la sélection des individus de la base d'apprentissage et de la base de test (la division 75/25). Sélectionner arbitrairement les 75 % premiers individus pour apprendre et les 25 % pour tester est une mauvaise approche. Prenons le cas où les données sont triées selon le label. Le modèle apprendra sur des observations très différentes de celles à tester. Ainsi, usuellement, les individus sont sélectionnés par tirage aléatoire afin de prévenir un effet de mauvaise représentativité de la base d'apprentissage. En classification, il est aussi important de s'assurer que le tirage aléatoire choisisse bien toutes les classes possibles du label.

Remarque sur l'optimisation du nombre d'arbres d'une forêt aléatoire

Par la suite, nous utiliserons l'algorithme des forêts aléatoires. Bien qu'il semble naturel d'optimiser le nombre d'arbres de la forêt, ce paramètre ne le sera pas car cela n'est pas nécessaire. En effet, une des explications est issue d'un ouvrage de Hastie et al. [22] :

Another claim is that random forests "cannot overfit" the data. It is certainly true that increasing B does not cause the random forest sequence to overfit; like bagging, the random forest estimate (15.2) approximates the expectation

$$\hat{f}_{\mathrm{rf}}(x) = \mathrm{E}_{\Theta} T(x;\Theta) = \lim_{B\to\infty} \hat{f}(x)_{\mathrm{rf}}^{B}$$

with an average over B realizations of Θ. The distribution of Θ here is conditional on the training data. However, this limit can overfit the data; the average of fully grown trees can result in too rich a model, and incur unnecessary variance. Segal (2004) demonstrates small gains in performance by controlling the depths of the individual trees grown in random forests. Our experience is that using full-grown trees seldom costs much, and results in one less tuning parameter.

En d'autres termes, augmenter le nombre d'arbres ne peut pas causer d'overfitting. Cependant, les autres paramètres eux peuvent être source d'overfitting. Une démonstration mathématique rigoureuse est apportée dans l'article de Philipp Probst et Anne-Laure Boulesteix : *To tune or not to tune the number of trees in random forest?* Nous pouvons tenter d'apporter une explication intuitive à ce phénomène. La performance d'une forêt aléatoire est la moyenne des prédictions des arbres. Par la Loi des Grands Nombres, cette performance se stabilisera à mesure que le nombre d'arbres tend vers l'infini. L'objectif est donc plutôt de trouver une valeur minimum d'arbres à partir de laquelle la variance de l'erreur de prédiction est faible.

En pratique, comment fixer le nombre d'arbres ? Il n'est pas nécessaire de le choisir par optimisation. Il faut le fixer comme une valeur raisonnable en termes de temps de calcul de l'algorithme, en général il est égal à 500. Leo Breiman propose lui une valeur de 1000. Idéalement, une fois que les autres paramètres sont fixés, nous pouvons essayer quelques valeurs du nombre d'arbres (100, 200, 500, 1000) et regarder à partir de quelle valeur la performance semble se stabiliser. Désormais, étudions une autre discipline du Machine Learning : l'apprentissage non supervisé.

3.4.2 L'Apprentissage non supervisé

Contrairement à l'apprentissage supervisé, l'apprentissage non supervisé ne cherche pas à prédire un label. La base de données utilisée pour ce type de machine learning ne contient

22. lien de l'ouvrage (citation page 596)

que des descripteurs ($Z_i = X_i$). Le but est de trouver des groupes homogènes dans les données. L'algorithme analyse alors automatiquement les valeurs des différentes variables afin de détecter certains schémas/structures entre elles. Dans le cadre de ce mémoire, nous utiliserons ce type d'approche uniquement en *clustering*[23]. La méthode la plus populaire pour l'effectuer est celle des **k-means** (k-moyens).

3.4.2.1 K-means

Créé en 1957 par Hugo Steinhaus et baptisé k-means par James MacQueen en 1967, cet algorithme a été développé dans un but de recherche sur les modulations d'impulsions codées au cœur des Laboratoires Bell. Cette approche sera révélée au grand public en 1982. Le principe des k-means est d'identifier des groupes distincts dans les données sans qu'aucune indication humaine ne soit donnée.

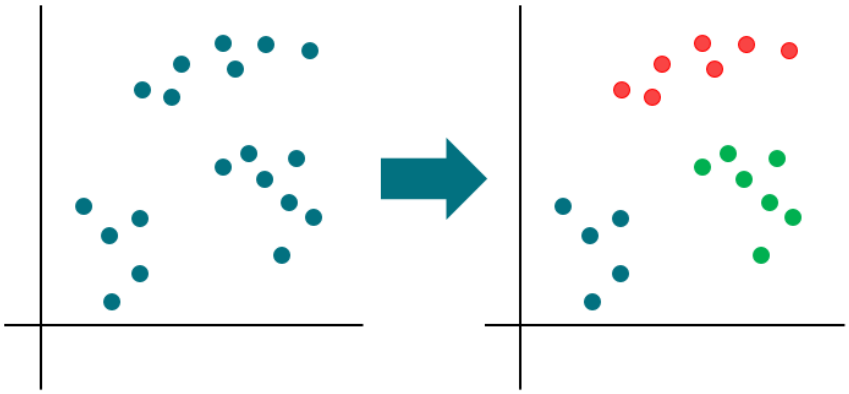


FIGURE 32 – Fonctionnement des k-means

Sur cet exemple en deux dimensions, trois groupes homogènes triviaux sont visibles. Ces clusters sont compacts et homogènes. Ceci est un exemple de 3-means, où $k = 3$. Comment l'algorithme procède-t-il pour trouver ces groupes ?

La première étape est de définir une distance entre les points. Nous utilisons généralement la distance euclidienne préalablement définie dans la partie KNN 3.4.1.1. Nous initialisons alors k points aléatoirement qui sont les centroïdes des clusters. Un centroïde est un individu fictif de référence d'un groupe, c'est le point milieu qui minimise la distance entre les autres individus du même groupe. Chaque donnée est alors associée à son centroïde dans le but de former des clusters. Deux étapes sont ensuite répétées :

1. Grouper chaque individu autour du centroïde le plus proche.
2. Redéfinir le centroïde de chaque cluster selon les individus du cluster.

Une fois les groupes stables, l'algorithme a convergé. Le paramètre à optimiser ici est le nombre de voisins k. Pour ce faire, il existe plusieurs méthodes.

23. Cluster = groupe

3.4.2.2 Optimisation des k-means

Méthode du coude

La méthode du coude consiste à analyser l'inertie des clusters. Le k optimal est celui qui forme un *coude* sur le graphique. L'inertie W d'un cluster est définie comme la somme des distances euclidiennes au carré entre chaque point et son centroïde associé, par exemple pour le cluster C de centroïde c :

$$\begin{aligned} W(C) &= \sum_{i:X_i \in C} \sum_{j=1}^{n} (X_{i,j} - c_j)^2 \\ &= \sum_{i:X_i \in C} \|X_i - c\|^2 \end{aligned}$$

L'inertie totale notée $Wtot_k$ est donc la somme des inerties des k clusters définie par :

$$Wtot_k = \sum_{m=1}^{k} W(C_m) = \sum_{m=1}^{k} \sum_{i:X_i \in C_m} \sum_{j=1}^{n} (X_{i,j} - c_j)^2$$

Le but est donc de trouver la valeur k minimisant $Wtot_k$ et tel que $Wtot_{k+1}$ est proche de $Wtot_k$. Logiquement, au plus k est grand au plus l'inertie est faible car nous augmentons le nombre de centroïdes. Les individus ont donc une plus grande probabilité d'être proche d'un centroïde.

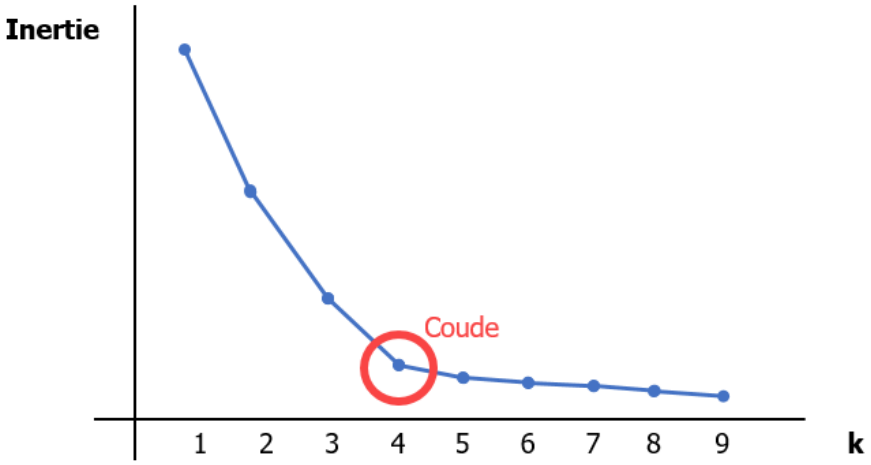


FIGURE 33 – Méthode du coude

Dans cet exemple, le k optimal vaut 4 car c'est celui qui forme le coude c'est-à-dire le point de la courbe où le gain en inertie est négligeable. Cette approche a l'avantage d'être simple et rapide en termes de temps de calcul. Il est aussi possible d'ajouter une seconde méthode afin de la valider : celle de la silhouette.

Méthode de la silhouette

Plus coûteuse en calcul que la méthode du coude, elle permet néanmoins d'ajouter plus de robustesse au choix optimal de k. Introduisons en premier lieu le coefficient de la silhouette égal à :

$$s = \frac{b - a}{\max(a, b)}$$

où a est la moyenne des distances entre les individus et b la distance moyenne au cluster le plus proche. Le coefficient s varie entre -1 et 1. Plus il est proche de 1 plus l'individu est bien situé dans son cluster tandis que s'il est proche de -1, l'individu n'est pas associé au bon cluster. Dans le cas où $s = 0$, l'individu est proche de la frontière de son groupe. Le but est donc de maximiser s en fonction de k.

Toutes les méthodes de prédiction (paramétrique et non paramétriques) étant présentées, nous pouvons dès à présent introduire la partie estimation de la prime commerciale. Pour rappel, nous nous plaçons dans le cadre où l'objectif est d'estimer la prime commerciale de réassurance à l'aide d'une formule de tarification basée sur l'écart type égale à :

$$PC = \frac{\bar{R} + \beta \times \sigma_R}{Rec_{factor} \times (1 - \alpha)}.$$

Le but est d'estimer β.

3.5 Méthode par chargement constant

Une des façons d'estimer β est de choisir celui qui minimise la différence entre la prime commerciale estimée en interne d'un traité et celle cotée par les réassureurs. Cette approche est nommée *méthode par chargement constant* car le coefficient de chargement estimé finalement est le même (donc constant) pour tous les traités concernés par cet algorithme. Notons $\hat{\beta}^*$ le coefficient de chargement estimé optimal. Soit a le plus petit coefficient à tester, b le plus grand, $step$ le pas désiré entre a et b et $f(a, b, step)$ la fonction calculant le vecteur des valeurs entre a et b avec un pas $step$. Posons T l'espace des traités de réassurance où $t \in T$ est un traité avec $Pcote_t$ la cotation moyenne du traité t. Ainsi, $Pcote$ représente le vecteur de l'ensemble des primes moyennes cotées pour tous les traités. L'algorithme de minimisation est le suivant :

Algorithm 2 Calcul de $\hat{\beta}^*$

Require: $a > b,\ step > 0$

$Pcote \leftarrow Pcote$
$\beta_{min} \leftarrow a$
$\beta_{max} \leftarrow b$
$step \leftarrow step$
$\beta \leftarrow [f(\beta_{min}, \beta_{max}, step)]$
$\hat{PC} \leftarrow [\text{length}(T)]$
$Error \leftarrow [\text{length}(\beta)]$
for $b \in \beta$ **do**
  $i \leftarrow 1$
  for $t \in T$ **do**
    $\hat{PC}[i] \leftarrow (\bar{R}_t + b \times \sigma_{R_t})/(Rec_{factor_t} \times (1 - \alpha))$
    $i \leftarrow i + 1$
  end for
  $Error[b] \leftarrow E(Pcote, \hat{PC})$
end for
$\hat{\beta}^* \leftarrow \underset{b}{\text{argmin}}(Error)$
return $\hat{\beta}^*$

Cet algorithme permet de calculer la différence entre les primes cotées et les primes estimées avec un coefficient β qui varie. Finalement, $\hat{\beta}^*$ minimise cet écart et est le coefficient de chargement optimal à considérer dans notre formule de tarification. La fonction $E(x, y)$ (ligne 14) est la mesure de l'erreur entre x et y. Nous choisissons quatre fonctions d'erreurs pour notre algorithme :

— **MAE** (Mean Absolute Error) : moyenne de la différence absolue moyenne entre l'échantillon estimé et celui d'origine.

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

— **MSE** (Mean Squared Error) : moyenne de la différence au carré entre l'échantillon estimé et celui d'origine.

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

— **MARE** (Mean Absolute Relative Error) : moyenne de la différence relative absolue entre l'échantillon estimé et celui d'origine.

$$MARE(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{\hat{y}_i - y_i}{y_i} \right|$$

— **MSRE** (Mean Squared Relative Error) : moyenne de la différence relative au carré entre l'échantillon estimé et celui d'origine.

$$MSRE(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\hat{y}_i - y_i}{y_i} \right)^2$$

En résumé, nous calculons une prime estimée pour chaque valeur choisie de β et sélectionnons celle qui minimise la fonction E, ici définie comme les quatre fonctions précédemment citées. Nous avons ainsi quatre valeurs possibles de $\hat{\beta}^*$. L'intérêt d'utiliser plusieurs mesures d'erreurs réside dans la façon même de définir un *écart*. Une différence au carré permet de fortement pénaliser les grandes erreurs comparée à une simple différence en valeur absolue. Cependant, dans certains cas, les ordres de grandeur peuvent grandement différer.

Nos primes cotées peuvent varier du simple au double (ou même triple) impliquant alors une forte disparité dans les valeurs de nos tarifs. Dans ce cas, le MAE ou le MSE ne sont pas nécessairement adaptés. En effet, une différence égale à 100 entre $y = 100$ et $\hat{y} = 200$ ne représente pas la même qualité d'estimation qu'entre un $y = 100\,000$ et $\hat{y} = 100\,100$. Dans le second cas $\hat{y}$ est bien plus proche de y. Or, le MAE et MSE ne feront aucune distinction entre ces deux écarts ce qui n'est pas forcément logique en termes de qualité d'estimation. Si nous ne pouvons pas assurer un même ordre de grandeur entre tous les y et $\hat{y}$, il est préférable de favoriser des écarts faibles en pourcentage en choisissant une mesure comme le $MSRE$ ou le $MARE$. Si nous reprenons notre exemple, $(\hat{y}-y)/y$ vaut 100% dans le premier cas et seulement 1% pour le second. Nous remarquons donc que la mesure de la différence est bien plus réaliste ici. Le choix entre ces quatre mesures dépend alors de l'importance que l'on souhaite apporter aux grands écarts :

	$\hat{y}$	y	$\hat{y}$	y
	350	200	15 000	12 000
MAE	150		3 000	
MSE	22 500		9 000 000	
MARE	75.00 %		25.00 %	
MSRE	56.25 %		6.25 %	

TABLE 12 – Différences entre les mesures d'erreurs

Cette approche a l'avantage d'être transparente et simple à utiliser en entreprise car elle permet d'obtenir une formule de tarification fixe [24] et *actuariellement* bien connue (principe de l'écart type). Nous appliquons cet algorithme à nos données.

3.5.1 Segmentation par risque

Dans un premier temps, nous appliquons cet algorithme en effectuant une segmentation par risque. L'ordre de grandeur des β étant très différent selon les risques, cette segmentation est nécessaire pour obtenir un coefficient de chargement cohérent. Pour mieux comprendre cette approche, illustrons-la avec le cas du GTPL. Pour chaque cotation moyenne et chaque type de fonction d'erreurs, nous estimons β.

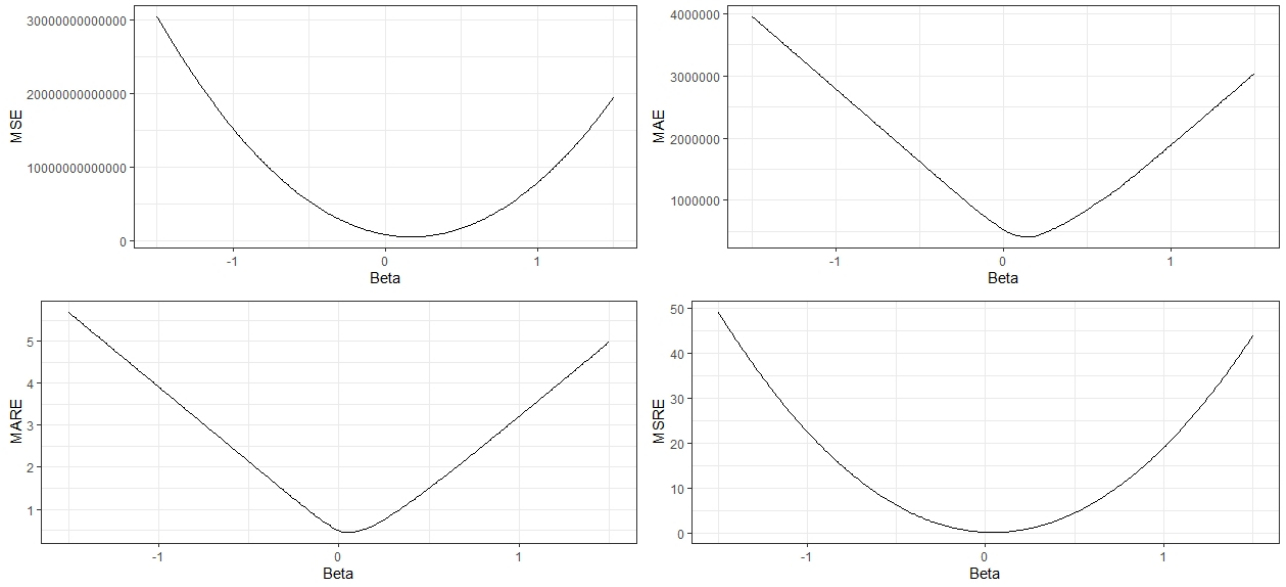


FIGURE 34 – Courbes d'erreurs de $\hat{\beta}$ en GTPL

Le résultat de ces calculs est une sortie de quatre fonctions convexes dont le minimum est atteint en $\hat{\beta}^*$ pour chaque mesure d'erreurs. Nous réalisons donc ces calculs pour chaque risque :

24. Où le coefficient de chargement est constant pour un risque (ou un ensemble de traités) donné.

Risque	Proportion	$\hat{\beta}^*$ (MSE)	$\hat{\beta}^*$ (MAE)	$\hat{\beta}^*$ (MARE)	$\hat{\beta}^*$ (MSRE)
CAT	0.31	-0.03	-0.02	-0.02	-0.05
GTPL	0.20	0.17	0.15	0.06	0.04
Marine	0.10	-0.32	-0.26	-0.05	-0.05
MTPL	0.21	0.08	0.05	0.00	-0.01
PTY	0.18	0.05	0.03	0.03	0.02

TABLE 13 – $\hat{\beta}^*$ après segmentation par risque

Nous remarquons une forte disparité de $\hat{\beta}^*$ entre risques suggérant que cette segmentation est judicieuse. De plus, en fonction du type d'erreur, le coefficient de chargement optimal change très nettement de valeur. En MTPL, celui-ci vaut 8 % lorsque la moyenne des écarts au carré est choisie comme fonction d'erreur (MSE). Cette valeur nous suggère qu'il existe une grande variance entre les ordres de grandeur des cotations. L'algorithme optimise alors la valeur de $\hat{\beta}$ sur les points donnant des écarts relatifs au carré forts. Ces cotations importantes nécessitant un grand coefficient de chargement, le $\hat{\beta}^*$ retenu avec la mesure du MSE est élevé.

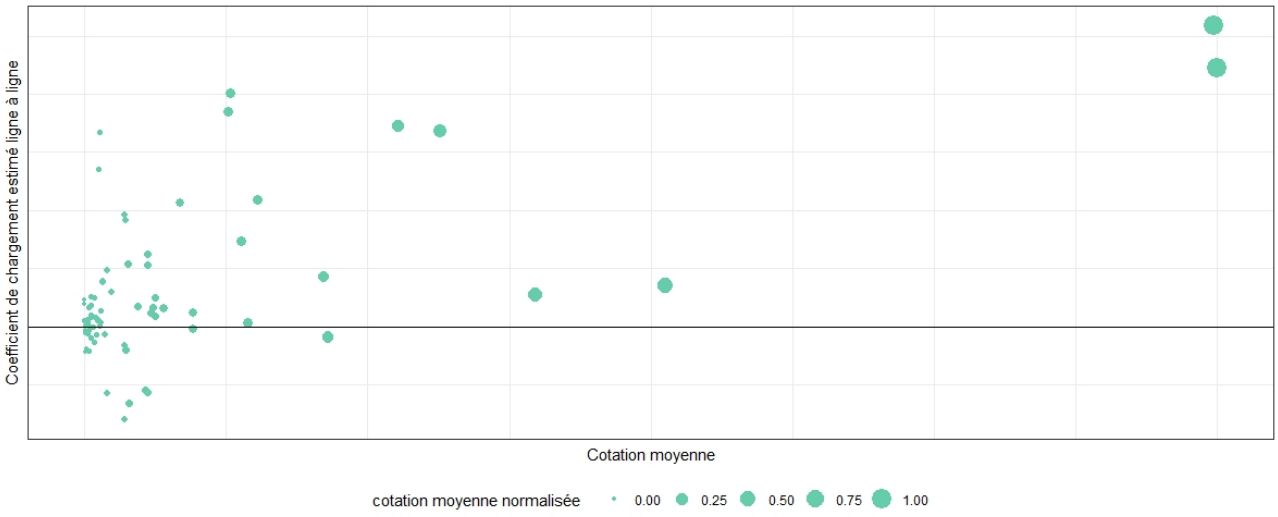


FIGURE 35 – Tendance de la cotation moyenne en fonction du chargement en MTPL

Interprétation de la figure : par souci de confidentialité, les cotations moyennes ont été normalisées. Le coefficient de chargement sur l'axe des ordonnées est celui calculé ligne à ligne comme dans la phase de nettoyage précédente où $\beta = (Cotation\ moyenne \times Rec_{factor} \times (1 - \alpha) - \bar{R})/\sigma_R$. *La taille du point est proportionnelle à la valeur de la cotation moyenne.*

En effet, comme expliqué ci-dessus, le risque MTPL a deux cotations moyennes très éloignées du reste avec en plus un coefficient de chargement [25] supérieur à tous les autres. Par conséquent, le MSE donne bien plus d'importance à l'estimation de ces points. Néanmoins, deux cotations moyennes fortes sont assez proches de la ligne médiane où $\beta = 0$, ce qui force une valeur de $\hat{\beta}^*$ moins importante. Sans ces deux points, l'ordre de grandeur de $\hat{\beta}^*$ est aux alentours de 30 %. Inversement, le risque Marine suit les mêmes tendances mais négativement. Cependant, le $\hat{\beta}^*$ obtenu par MSE est assez proche de celui calculé par MAE. Ceci indique une présence forte de cotations moyennes élevées ayant un coefficient de chargement négatif élevé.

25. Le coefficient de chargement considéré est celui calculé ligne à ligne.

Le risque CAT quant à lui est assez constant dans la valeur des $\hat{\beta}^*$, peu importe la mesure d'erreurs choisie (à 3 % près). Cette stabilité nous indique que, tout d'abord, l'ordre de grandeur des cotations moyennes est assez similaire pour tous les points et que leur coefficient de chargement est centré autour de $0-$. Le $\hat{\beta}^*$ par MSRE étant le plus élevé, cela indique que les cotations moyennes assez faibles ont des chargements bas. L'erreur relative au carré les prend alors davantage en considération que les autres. De plus, ce risque représente 31 % des données, nous pouvons donc le considérer comme le plus fiable en termes de qualité d'estimation.

La tendance en PTY semble être similaire à celle du CAT mais positivement. Enfin, le risque GTPL possède un $\hat{\beta}^*$ très différent entre les mesures absolues et relatives. Cela indique une forte présence de cotations moyennes élevées ayant un chargement important. Or, les erreurs relatives indiquant un $\hat{\beta}^*$ autour de 5 %, soit environ 10 % de moins que pour les erreurs absolues, cette valeur nous signale qu'il existe une forte part de cotations faibles ayant des coefficients de chargement bas.

Finalement, les différences entre les signes de $\hat{\beta}^*$ et son ordre de grandeur selon la mesure d'erreurs indiquent qu'une segmentation par risque ne suffit pas.

3.5.2 Segmentation par risque et par signe du chargement

Nous réappliquons le même algorithme mais cette fois en créant deux sous-groupes par risque : un où le coefficient de chargement estimé ligne à ligne est positif (groupe 2) et un autre où il est négatif (groupe 1). Ceci permet d'obtenir une plus grande cohérence dans la valeur finale de $\hat{\beta}^*$. Cependant, une segmentation par signe diminuant le nombre de points pour chaque sous-groupe, la qualité d'estimation peut être impactée négativement. Le gain de cohérence n'est quant à lui pas négligeable. En général, il existe toujours un compromis à faire entre transparence et performance de l'approche. Cette balance se résume bien par la fameuse phrase de Jim Jarmush : *fast, cheap and good... pick two.* Lors de la tarification d'un nouveau traité, il est nécessaire de lui attribuer le bon groupe. Or, celui-ci n'étant pas connu à l'avance, il est nécessaire de le prédire. Pour ce faire, nous pouvons par exemple utiliser un algorithme de prédiction en classification. L'intérêt de cette partie résidant dans la transparence de la méthode, l'usage de Machine Learning n'est pas nécessairement adapté en fonction des algorithmes choisis. Dans l'optique de limiter un maximum l'effet boîte noire, nous nous concentrons uniquement sur des méthodes simples (KNN) ou bien visuelles (arbre de décision) afin que la simplicité de la méthode ne soit pas trop impactée. Ces méthodes sont généralement nommées *méthodes boîtes blanches*.

3.5.2.1 Estimations de $\hat{\beta}^*$ après la nouvelle segmentation

Dans un premier temps, analysons les résultats de cette nouvelle segmentation :

Risque	Groupe	Proportion	$\hat{\beta}^*$ (MSE)	$\hat{\beta}^*$ (MAE)	$\hat{\beta}^*$ (MARE)	$\hat{\beta}^*$ (MSRE)
CAT	1	0.19	-0.09	-0.07	-0.08	-0.09
	2	0.12	0.03	0.03	0.03	0.04
GTPL	1	0.03	-0.15	-0.14	-0.06	-0.06
	2	0.17	0.21	0.16	0.09	0.07
Marine	1	0.05	-0.32	-0.26	-0.08	-0.08
	2	0.05	0.12	0.06	0.06	0.09
MTPL	1	0.06	-0.5	-0.05	-0.03	-0.03
	2	0.15	0.28	0.14	0.03	0.02
PTY	1	0.05	-0.13	-0.10	-0.05	-0.06
	2	0.13	0.20	0.15	0.06	0.06

TABLE 14 – $\hat{\beta}^*$ après segmentation par risque et par signe

Pour rappel, le groupe 1 (respectivement 2) correspond aux cas où le β estimé ligne à ligne est négatif (respectivement positif). Comme attendu, la proportion de données pour estimer β de certains groupes a fortement diminué notamment pour le groupe 1.

Le risque Marine est divisé en deux, avec autant de points dans le groupe négatif que dans le positif. Nous remarquons alors l'intérêt de cette nouvelle segmentation car pour une même mesure d'erreur, en l'occurrence le MSE, $\hat{\beta}^*$ vaut -32 % pour le groupe 1 et 12 % pour le 2. Lors de notre précédente étude, sans segmentation par signe, $\hat{\beta}^*$ valait -32 %. En effet, tous les points nécessitant un chargement positif n'étaient pas visibles.

La stabilité entre les mesures d'erreurs pour les deux groupes est observable seulement pour le risque CAT. Les $\hat{\beta}^*$ calculés par erreurs relatives ou absolues sont presque égaux. Ceci est le signe que l'ordre de grandeur des primes commerciales estimées est similaire, indiquant une segmentation sensée. Ainsi, pour ce risque en particulier et avec cette méthode, nous pouvons par exemple retenir un $\hat{\beta}^*$ de -8 % pour le groupe 1 et 3 % pour le groupe 2.

Les autres risques ne suivent pas la même tendance. Les différences entre mesures relatives et absolues sont toujours élevées. Par exemple, pour le groupe positif en MTPL, $\hat{\beta}^*$ varie de 28 % (MSE) à 2 %(MSRE). Ceci indique qu'il existe encore des points atypiques ne permettant pas de fixer un β constant par cette méthode pour ce groupe. Les risques GTPL, Marine et MTPL présentent exactement le même problème de volatilité du $\hat{\beta}^*$.

Supposons néanmoins que nous validons cette segmentation et que nous choisissons de fixer $\hat{\beta}^*$ pour chaque risque et groupe. Comment connaître le groupe d'un nouveau traité afin de lui attribuer le bon coefficient de chargement ? Une manière naturelle de faire est d'utiliser un algorithme de prédiction. Comme explicité par avant, n'oublions pas que l'intérêt de cette section réside dans la simplicité de son utilisation. Pour respecter ce cadre, les algorithmes utilisés doivent donc être simples à comprendre et éviter de faire tourner des programmes informatiques complexes. Deux algorithmes peuvent répondre en partie à ces exigences opérationnelles : les KNN et les CART.

3.5.2.2 Prédiction du groupe par KNN

Très peu coûteux en termes de temps de calculs, simple à optimiser et à comprendre, l'algorithme des KNN est un bon candidat. Malheureusement, un inconvénient majeur de cet

algorithme, qui peut fortement nuire à la pertinence de ses prédictions, est le calcul des distances qui ne peut se faire qu'entre des variables quantitatives. En effet, calculer une distance entre deux risques n'a pas de sens. Il convient donc de faire un modèle par risque. Or, le nombre de données n'étant pas forcément suffisant sur certains risques, la qualité de prédiction peut être insatisfaisante et instable. Entraînons tout de même cet algorithme sur nos données. Les variables présélectionnées sont les suivantes :

Type de Variable	Variable	Référence
Descripteurs	Limite	Limite/portée du traité
	Priorité	Priorité/rétention du traité
	Récupération moyenne	$\bar{R} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{M_i} X_j^T$
	Écart type des récupérations	$\sigma_R = \sqrt{\frac{\sum_{i=1}^{n}(\sum_{j=1}^{M_i} X_j^T - \frac{1}{M_i}\sum_{j=1}^{M_i} X_j^T)^2}{n}}$
	Facteur de reconstitution	$Rec_{factor} = 1 + \alpha + \beta + ... + \delta$
	Probabilité d'attachement	$Attachement_{prob} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{\sum_{j=1}^{M_i} X_j^T > 0\right\}$
	Probabilité d'épuisement	$Epuisement_{prob} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{\sum_{j=1}^{M_i} X_j^T > Perte_{max}^T\right\}$
Label	Groupe	$\{1, 2\}$

TABLE 15 – Variables retenues pour la prédiction du groupe par KNN

L'étude préalable de ces variables est primordiale afin d'analyser si certaines d'entre-elles sont corrélées. Pour calculer les corrélations entre variables numériques, nous utilisons le test de Pearson. Concernant la variable cible *Groupe*, celle-ci étant catégorielle, nous effectuons une analyse graphique par boxplot [26].

Rappel des corrélations de Pearson

Le coefficient de Pearson est une valeur qui analyse une potentielle relation linéaire entre deux variables continues. Celui-ci varie entre -1 et 1. Dans le cas où il vaut -1 (respectivement 1) les deux variables sont négativement (respectivement positivement) corrélées c'est-à-dire que lorsque l'une des deux croît, la seconde décroît (respectivement croît). Lorsque ce coefficient est égal à 0, il n'y a pas de corrélation. Ce coefficient se calcule de la façon suivante, en posant X et Y nos deux variables d'intérêt :

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}$$

Le test de Pearson est régit par deux hypothèses : $[H_0 : r = 0]$ contre $[H_1 : r \neq 0]$.

Procédons à l'analyse des relations entre nos descripteurs :

26. Boîte à moustache

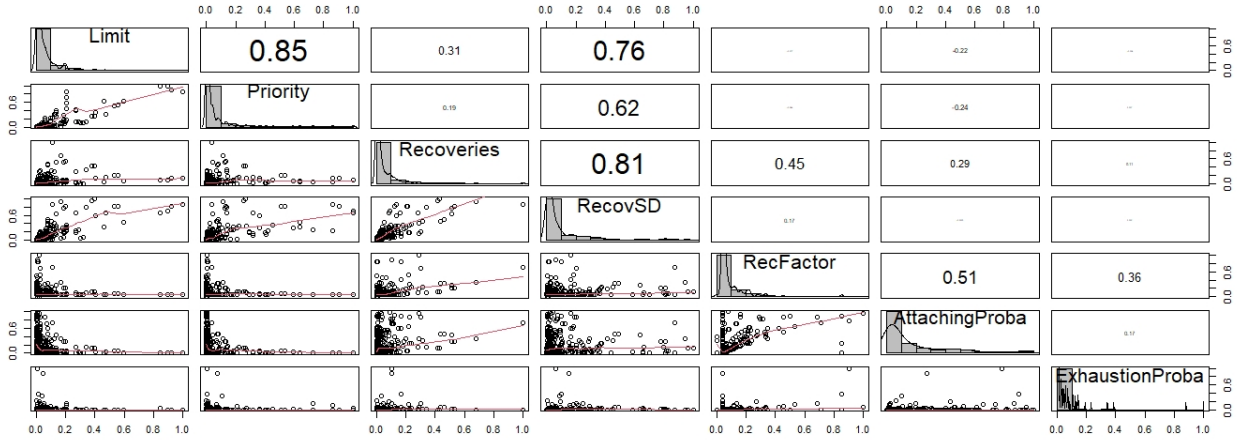


FIGURE 36 – Analyse des descripteurs

Interprétation de la figure : Par souci de confidentialité, les axes snt normalisés. La partie basse du graphique (triangle bas) réalise des nuages de points entre les variables de la diagonale. Par exemple, le graphique tout en bas à gauche est la limite (abscisse) en fonction de la probabilité d'épuisement (ordonnée). La ligne rouge est une droite de régression. La diagonale représente les histogrammes de chaque descripteur. De plus, une densité non paramétrique est ajoutée au-dessus des histogrammes. Enfin, la partie en haut à droite (triangle haut) représente les corrélations de Pearson. Par exemple, la corrélation entre l'écart type des récupérations et les récupérations moyennes est de 0.81. À noter que la taille d'affichage de ces coefficients est proportionnelle à leur valeur. À noter également que les cas où la limite est infinie ne sont pas affichés. Ces observations sont très peu nombreuses et limitées à deux risques. De plus, étant donné leurs grandes valeurs, ils rendent complexe la lecture des graphiques.

Nous observons une forte disparité dans les ordres de grandeur des corrélations. La limite et la priorité sont fortement corrélées (85 %). Ceci traduit le fait que la priorité et la limite semblent augmenter simultanément. La récupération moyenne et l'écart type des récupérations suivent aussi la même tendance avec 81 % de corrélation. Ainsi, un traité dont les récupérations sont volatiles, semble aussi avoir une récupération moyenne élevée. Sans surprise, la probabilité d'attachement est corrélée positivement à 51 % au facteur de reconstitution. Un traité dont la probabilité d'être touché par un sinistre est élevée a plus de chances de consommer des reconstitutions.

Dans nos modèles, il est important de choisir des variables indépendantes. Si deux descripteurs sont trop corrélés nous pouvons les considérer comme des doublons et les algorithmes peuvent sur-considérer l'importance sous-jacente de ceux-ci. Ainsi, si nous observons des variables dans ce cas, nous devons être sûrs que la corrélation entre-elles n'implique pas qu'elles soient *identiques*, que leur information soit bien différente. Dans notre cas, la récupération moyenne et l'écart type apportent des informations différentes. L'une en termes de comportement moyen et la seconde sur la volatilité du traité. Celles-ci étant complémentaires et non pas identiques, leur remplacement doit se faire par un indicateur n'entraînant pas la perte d'information. Une mesure triviale combinant ces deux variables est la récupération moyenne standardisée égale à $\bar{R}/\sigma_R$.

Concernant la limite et la priorité, le sujet de la différence d'information est plus flou. En effet, ces deux variables indiquent une information de seuillage (plafond dans un cas, plancher

dans l'autre), elles sont donc du même *type*. Ainsi, nous nous proposons de remplacer la limite et la priorité par le RMP (Relative Median Point), qui pour rappel, vaut :

$$\text{RMP} = \frac{1}{2}\text{Limite} + \text{Priorité}$$

Cet indicateur équivaut au point milieu de la tranche du traité XS. Analysons désormais les boxplots de nos variables en fonction des groupes. Pour rappel, un boxplot s'analyse de cette façon :

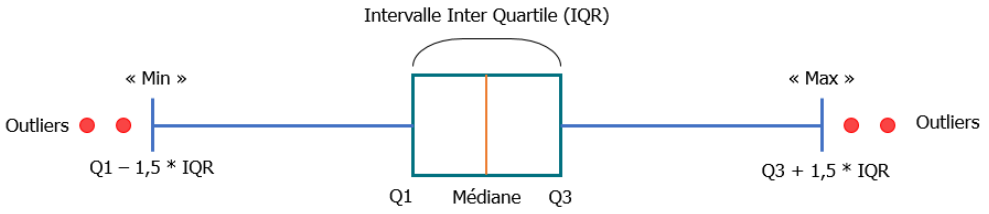


FIGURE 37 – Fonctionnement du boxplot

Dans un boxplot (boîte à moustache en français), $Q1$ (respectivement $Q3$) est le quantile à 25 % (respectivement 75 %). La définition du maximum et du minimum est assez spécifique à ce type de graphique puisqu'elle est égale à $Q3 + 1.5 \times IQR$ pour le maximum et $Q1 - 1.5 \times IQR$ pour le minimum. Le boxplot considère les points trop éloignés de l'intervalle interquartile comme des outliers (points rouges). Finalement, ce type de graphique permet de représenter la distribution des données assez simplement. Visualisons les boxplots de nos données, en fonction du groupe 1 et 2 :

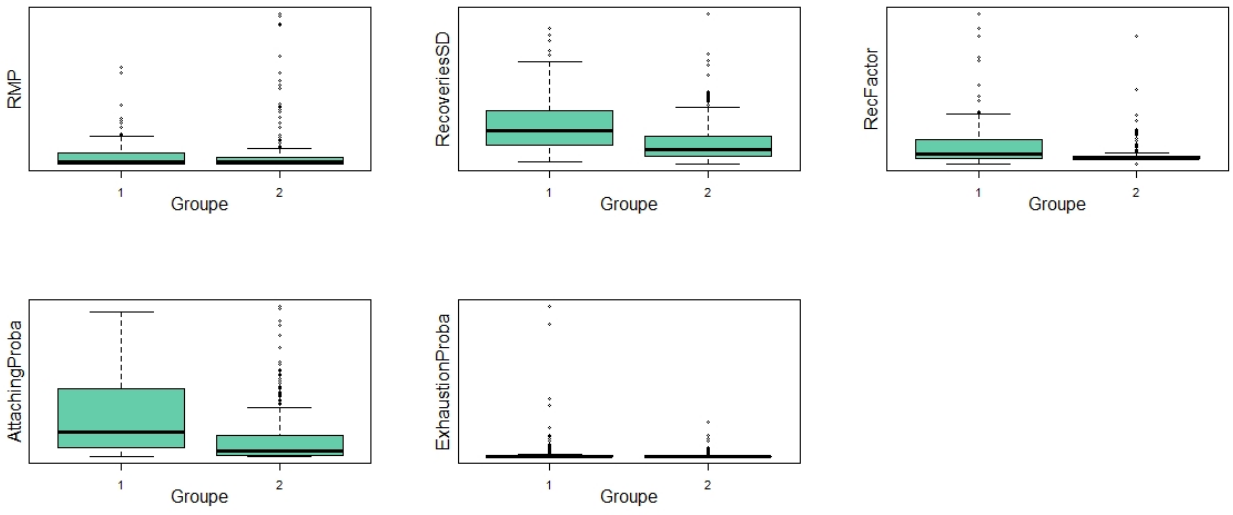


FIGURE 38 – Groupes 1 et 2 en fonction des descripteurs

Interprétation de la figure : Par mesure de confidentialité, les valeurs des axes ne sont pas affichées. La variable RecoveriesSD représente la récupération moyenne standardisée égale à $\bar{R}/\sigma_R$.

La probabilité d'attachement semble être la variable disposant de la plus grande différence de distribution entre les deux groupes. Pour le premier, le 3$^{\text{ème}}$ quartile (quantile à 75 %) est bien au-dessus de celui du boxplot du deuxième groupe. La médiane suit aussi la même tendance. Les maximums par contre sont eux équivalents. Concernant le RMP, il n'y a pas de différence nette visible, mis à part pour les outliers du boxplot. La récupération standardisée et le facteur de reconstitution semblent être plus dispersés pour le groupe 1 que pour le groupe 2. Cependant, les médianes sont assez proches pour les deux groupes. Enfin, les outliers de la probabilité d'épuisement compliquent la lecture du graphique, ils *écrasent* le boxplot.

L'approche de modélisation est la suivante : pour chaque risque, nous entraînons un algorithme KNN par 10-fold cross validation. Chaque modèle est découpé en 75 % pour la base d'apprentissage et 25 % pour la base de test. Les variables sont toutes centrées et réduites [27] afin de réduire l'importance des différences d'échelles entre les différents descripteurs. Par exemple, les limites et les probabilités d'attachements ont des ordres de grandeur très différents. Ainsi, un calcul de distance sans prise en compte de cette différence peut engendrer une plus grande prise en compte de la limite et une sous considération de la probabilité d'attachement. Les résultats sont les suivants :

Risque	k optimal	Précision sur la base d'apprentissage	Précision sur la base de test
CAT	1	0.86	0.81
GTPL	6	0.92	0.81
Marine	1	0.57	0.75
MTPL	9	0.67	0.65
PTY	8	0.71	0.60

TABLE 16 – Performance du KNN en prédiction des groupes 1 et 2

Pour rappel, la précision est calculée de la sorte :

$$\text{Précision} = \frac{\#\text{Prédictions correctes}}{\#\text{Prédictions}}$$

Les scores sont très différents en fonction du risque. En CAT et GTPL, l'algorithme fonctionne de façon très correcte. Les autres risques sont plus décevants, aucun d'entre eux ne dispose d'une performance satisfaisante. Une façon d'expliquer ce phénomène est le manque de données de chaque algorithme. Diviser la prédiction en cinq sous-modèles ne favorise pas les risques où il y a moins de données. En effet en Marine, l'algorithme performe très peu avec 57 % sur la base d'apprentissage et 75 % sur celle de test. Cette différence flagrante de précision entre échantillon d'apprentissage et de test peut indiquer que les données de test ont été bien estimées par *chance* ; c'est-à-dire que le faible nombre de points proposés en test sont ceux correctement modélisés par la base d'apprentissage. Tout ceci nous pousse à changer de méthode afin d'améliorer nos prédictions.

La façon la plus naturelle de procéder est d'établir un modèle général pour tous les risques. Ceux-ci étant des variables catégorielles, il est nécessaire d'opter pour un algorithme capable d'apprendre sur ce type de variable. De plus, afin de respecter notre besoin de transparence et de simplicité de prédiction, notre méthode ne doit pas comporter d'effet boîte noire trop important. L'approche répondant à tous ces critères est le CART.

27. Soustraction de la moyenne de la colonne et division par l'écart type de la colonne ($\frac{X-\bar{X}}{\sigma_X}$).

3.5.2.3 Prédiction du groupe par CART

La modélisation de cette partie est similaire à celle des KNN. La division entre apprentissage et test est de 75 % - 25 %, avec 15-fold cross validation. Dans un CART, le paramètre à optimiser est la profondeur de l'arbre. Nous fixons cinq valeurs possibles : 1, 3, 4, 6 et 10. Celles-ci sont faibles pour préserver une certaine simplicité dans l'arbre et limiter le risque d'overfitting. Les résultats sont les suivants :

Profondeur maximale	Précision	Kappa	Précision SD	Kappa SD
1	0.7094	0.3957	0.1083	0.2247
3	0.7131	0.3357	0.0795	0.2024
4	0.6946	0.3043	0.0844	0.2146
6	0.6946	0.3149	0.0818	0.2034
10	0.6946	0.3159	0.0762	0.1971

TABLE 17 – Performances du CART en prédiction des groupes 1 et 2

Comme nous pouvons l'observer, la profondeur optimale est 3 puisque c'est elle qui maximise la précision. À noter que la précision est celle calculée par cross validation, sur l'échantillon de validation.

Remarque sur Kappa

Le Kappa de Cohen est calculé selon la formule suivante :

$$\kappa = \frac{p_0 - p_e}{1 - p_e}$$

où p_0 est la performance du modèle et p_e la performance espérée. Attardons-nous quelques instants sur p_e. Plus clairement, la performance espérée est celle obtenue par un classifieur aléatoire calculée à partir de la matrice de confusion. Prenons par exemple le cas usuel suivant :

Prédit \ Réel	Malade	Sain
Malade	10	7
Sain	5	8

Lecture de la matrice : 10 personnes malades ont été prédites comme malades. 7 personnes saines ont été prédites comme malades. 5 personnes malades ont été prédites comme saines. 8 personnes saines ont été prédites comme saines.

Pour calculer la performance espérée de notre matrice de confusion, nous devons multiplier la fréquence de prédiction de malades (10+7 = 17) par la fréquence réelle de malades (10+5 = 15) puis diviser ce produit par la fréquence totale de la matrice (réel et prédiction, soit 10+7+5+8 = 30) soit un résultat de 8.5 ($\frac{15\times17}{30}$). Nous effectuons les mêmes calculs avec le cas sain et nous obtenons 6.5 ($\frac{(7+8)\times(5+8)}{30}$). Finalement, nous sommons ces deux valeurs et les rediviser par 30 soit $(8.5 + 6,5)/30 = 0.5$. Ainsi, $p_e = 0.5$. Ceci est toujours le cas lorsque les deux classes contiennent autant de données (15 chacune ici). La performance réelle p_0 est la précision égale à $(10 + 8)/30 = 0.6$. Ainsi,

$$\kappa = \frac{0.6 - 0.5}{1 - 0.5} = 0.2$$

Si nous prenons un exemple légèrement différent :

Prédit \ Réel	Malade	Sain
Malade	22	9
Sain	7	13

Les résultats sont les suivants :
— **Fréquence réelle :** 29 malades, 22 sains
— **Fréquence de prédiction :** 31 malades, 20 sains
— **Total :** 51
— $p_0 = ((22 + 13)/51) = 0.69$
— $p_e = \frac{29\times31+22\times20}{51^2} = 0.51$
— $\kappa = 0.37$

Cet indice évalue en somme la qualité de prédiction de notre classifieur par rapport au hasard. Il n'existe pas d'interprétation universelle de la valeur de κ. Cependant, Landis et Koch [28] ont tenté de proposer des ordres de grandeur mais qui ne font pas consensus dans la communauté scientifique, notamment à cause du fait que le nombre de classes influe sur la valeur de κ (plus leur nombre est faible, plus κ est élevé et inversement). Leur conclusion est de considérer des tranches de κ variant par pas de 0.2, de 0 à 1, 0 étant très faible et 1 parfait. Fleiss [29] considère lui une valeur entre 0.4 et 0.7 comme bonne et au-dessus comme excellente. Revenons désormais à nos données.

Le tableau 17 indique un kappa proche de 0.33, ce qui est, selon l'échelle arbitraire évoquée juste avant, plutôt une mauvaise valeur. De plus, son écart type est assez élevé (0.2024). Cette valeur est issue des 15-fold cross validation effectuées. Cependant, la volatilité de la précision semble elle plus acceptable (autour de 8 %). Cette valeur de κ paraît nous indiquer que notre classifieur a mieux performé que sa réelle capacité à prédire. Afin de prendre connaissance des classes les mieux prédites, affichons la matrice de confusion de notre modèle :

Prédit \ Réel	1	2
1	16.0 %	5.9 %
2	22.7 %	55.4 %

TABLE 18 – Matrice de confusion du CART final (classification des groupes 1 et 2)

La fréquence est en pourcentage car nous observons ici les prédictions moyennes après les 15-fold validations croisées. De plus, toutes les valeurs sont affichées en proportion de la fréquence totale (la somme des cellules vaut 100 %). Le groupe des β positifs (groupe 2) est bien mieux prédit que celui des β négatifs avec uniquement 16 % de prédictions moyennes correctes contre 22.7 % fausses. Ceci signifie que notre classifieur prédit correctement le groupe 1 en moyenne seulement 41 % du temps $(16/(16 + 22.7))$. Notre arbre n'est donc pas précis pour anticiper les traités disposant d'une vision interne plus pessimiste que celle des réassureurs ($\beta < 0$). Inversement, la prédiction du groupe 2 est elle très bonne avec une précision de 90 % (55.4/(55.4+5.9)).

Afin d'évaluer la pertinence de nos variables, nous pouvons entraîner un arbre disposant de toutes les variables citées auparavant (15), dans les mêmes conditions que celui-ci. Les résultats sont sensiblement équivalents avec une précision optimale de 73 % et un kappa de 0.41. Le gain

28. Landis, J. R. and Koch, G. G. (1977) pp. 159–174
29. Tableau comparatif des valeurs de κ

de précision est négligeable. Notre choix de réduction de dimension en intégrant le RMP et les récupérations moyennes standardisées se révèle donc pertinent car nous gagnons en transparence (moins de variables) sans perdre en précision.

Il convient désormais d'afficher notre arbre afin de prendre connaissance de sa forme.

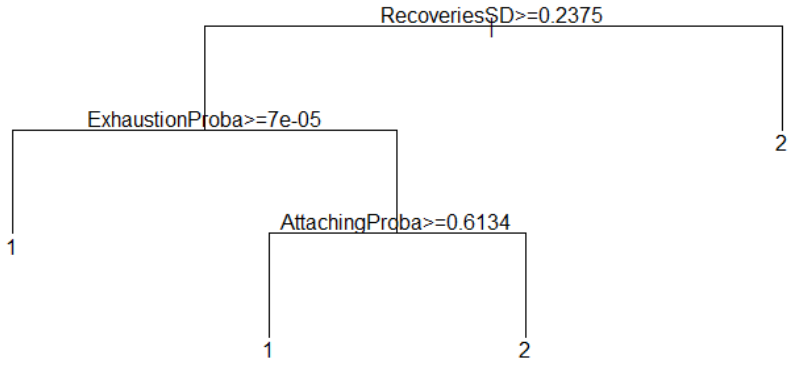


FIGURE 39 – CART final en segmentation binaire

Comme nous le constatons, cet arbre est très simple et dispose d'une profondeur de 3. Les variables les plus pertinentes dans le seuillage sont l'écart type des récupérations, la probabilité d'épuisement et d'attachement. Ces colonnes semblent expliquer au mieux le signe du coefficient de chargement. De plus, grâce à sa petite taille, un tel arbre est facilement utilisable en entreprise de par sa simplicité dans sa compréhension et sa prédiction. Enfin, dans l'optique de valider la précision attendue en pratique, il convient de prédire notre base test (25 % des données) et d'observer la matrice de confusion ainsi que les indicateurs associés.

Prédit \ Réel	1	2
1	12.5 %	5.7 %
2	26.1 %	55.7 %

Précision	68.18 %
Sensitivité	32.35 %
Spécificité	90.74 %

TABLE 19 – Prédiction sur l'échantillon de test du CART en segmentation binaire

Par souci de confidentialité, nous affichons ici les fréquences en pourcentage du nombre total d'individus de la base de test. La prédiction semble suivre la tendance de la base d'apprentissage avec un score de 68 % contre 71 % en phase d'entraînement. Cette différence de 3 points indique une bonne adaptation du modèle avec une quasi-inexistence d'overfitting. De plus, la prédiction du groupe 1 semble une nouvelle fois poser problème avec 32.35 % (sensitivité) de taux de succès contre 90.74 % (spécificité) pour le groupe 2. Comme attendu, le modèle ne sait pas bien prédire le groupe des $\beta < 0$. La liste de tous les indicateurs fournis en classification binaire par la matrice de confusion (par la librairie **caret** sur R) est disponible en annexe ici 77.

En conclusion, cette méthode nous permet d'estimer la prime commerciale de réassurance par risque selon le signe du coefficient de chargement β. Pour chaque nouveau traité à tarifer,

nous estimons le signe du coefficient puis nous attribuons le coefficient de chargement adéquat en fonction du risque. Les résultats 14 indiquent cependant une différence forte dans les valeurs de $\hat{\beta}^*$ en fonction de la mesure d'erreur, suggérant que des valeurs atypiques sont présentes. Un choix naturel est donc de créer un groupe atypique.

3.5.3 Segmentation par risque et par type de chargement

La forte volatilité des estimations de $\hat{\beta}^*$ en fonction de la mesure d'erreur suggère qu'il est nécessaire d'effectuer deux groupes : atypique et attritionnel. Attritionnel pour les β standards et atypique pour ceux extrêmes. Alors, comment identifier le seuil permettant de diviser un groupe en ces deux types de chargement ? Pour parvenir à estimer ce seuil extrême nous utilisons une partie de la théorie des valeurs extrêmes (TVE), plus particulièrement les méthodes d'estimation de queue de distribution.

3.5.3.1 Notions autour de la théorie des valeurs extrêmes

En TVE, il existe de fameuses méthodes graphiques permettant de fixer un seuil extrême d'une distribution. Généralement, cette théorie est utilisée en assurance pour modéliser les sinistres au comportement atypique. Un indice de queue d'une loi extrême (GEV ou GPD) est estimé afin de simuler ces sinistres. Dans notre cas, l'approche est assez similaire puisque nous souhaitons identifier par risque, à partir de quelle valeur de β [30] celui-ci devient bien plus grand. Pour réaliser cette étude, nous utilisons trois indicateurs : le **mean excess plot**, le **hill plot** et le **gertensgarbe plot**. Ces trois méthodes graphiques ont l'avantage d'être *facilement* réalisables et assez aisément interprétables (selon les cas).

3.5.3.1.1 Mean Excess Plot (MEP)

Le MEP est un graphique mesurant l'espérance de X au-delà d'un seuil u sachant u. Posons $M(u)$ la valeur de cette espérance au-delà de u, alors $M(u)$ est égal à :

$$M(u) = E[X - u \mid X > u]$$

Empiriquement, cette valeur vaut donc :

$$\widehat{M}(u) = \frac{\sum_{X_i > u} (X_i - u)}{\sum_{i=1}^{n} \mathbb{1}\{X_i > u\}}, u \geq 0$$

En fonction de la distribution de X, il existe certaines formes du MEP connues, nous permettant par la même occasion de modéliser la loi de X. Dans notre cas, nous nous limitons à l'estimation du seuil u afin de fixer un groupe attritionnel et atypique.

30. Celui estimée ligne à ligne

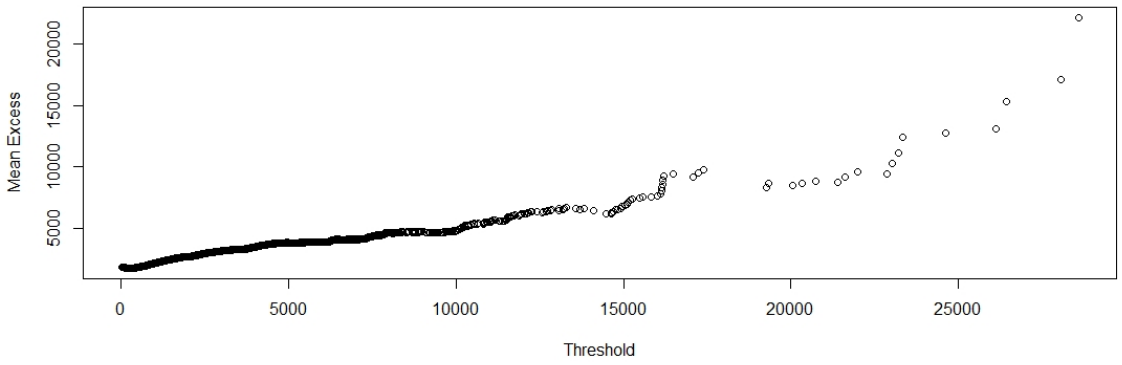


FIGURE 40 – Exemple de MEP en assurance auto

Ci-dessus un exemple de mean excess plot effectué sur la variable *PAID* des données *AutoClaims* du package *insuranceData* disponible sous R. Cette forme montre une volatilité croissante de $M(u)$ au fur et à mesure que u augmente puisque le nombre de points diminue. Nous constatons ici que lorsque u atteint environ 15 000, la forme linéaire du MEP change, indiquant que l'indice de la queue de distribution des sinistres se trouve probablement vers cette valeur. Définissons désormais le hill plot.

3.5.3.1.2 Hill Plot

Introduit en 1975, l'estimateur de Hill est défini par :

$$\widehat{\gamma}_{k,n}^{(H)} = \frac{1}{k}\sum_{j=1}^{k} \log X_{n-j+1,n} - \log X_{n-k,n}$$

L'idée est ici de choisir une bonne valeur de k (équivalent au u du MEP). La notion de *bonne* valeur repose une fois de plus sur une interprétation graphique du hill plot. Ce graphique représente $\widehat{\gamma}_{k,n}^{(H)}$ en fonction de k (où n est le nombre total d'observations). Une valeur optimale de k est atteinte quand $\widehat{\gamma}_{k,n}^{(H)}$ semble se stabiliser. Affichons ce graphique sur les données utilisées pour le MEP :

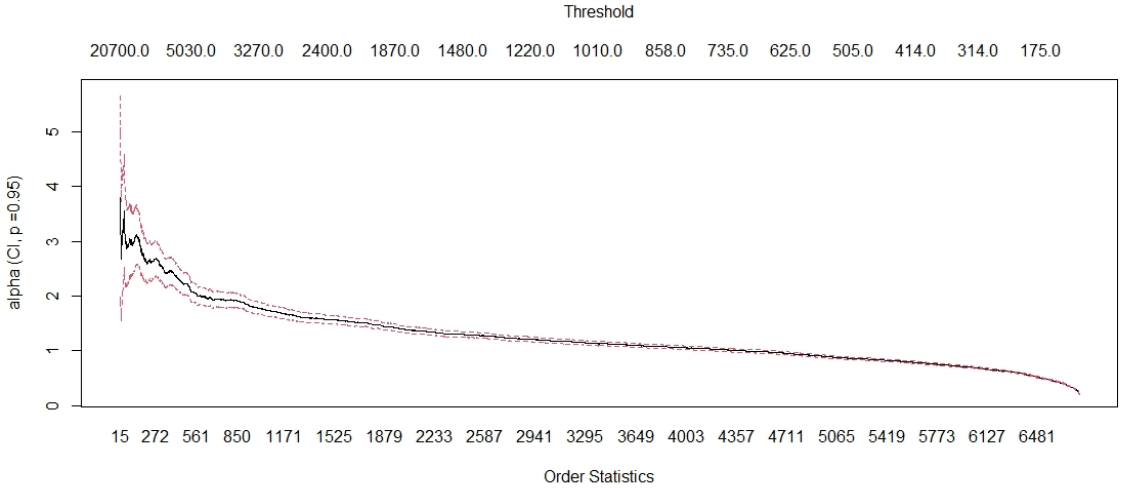


FIGURE 41 – Exemple de Hill plot en assurance auto

En abscisse nous retrouvons la valeur de k et celle de $\widehat{\gamma}_{k,n}^{(H)}$ en ordonnée. En haut du graphique est affiché le montant de sinistre associé à la $k^{\text{ème}}$ valeur du jeu de données (trié par ordre décroissant). Sur ce graphique, l'estimateur est très volatil lorsque le nombre de données k est faible. Celui-ci semble tout de même être stable lorsque k est autour 561 (soit 5 030 en montant de sinistre).

3.5.3.1.3 Gertensgarbe Plot

La méthode de Gertensgarbe permet de fournir un outil graphique de deux courbes où l'intersection de celles-ci se trouve au seuil optimal. Posons $\Delta_i = \beta_{(i)} - \beta_{(i-1)}$ pour $i = 2, \ldots, n$ le vecteur des différences des β ordonnés par ordre croissant [31]. Cette procédure repose sur l'idée qu'il paraît raisonnable de penser que le comportement de Δ change lorsqu'il est calculé sur des valeurs extrêmes. Il doit exister un seuil $\hat{k}$ marquant l'entrée de Δ (et donc de β) dans la région extrême. Pour identifier ce seuil, nous utilisons deux séries inspirées du test de Mann-Kendall calculées sur les Δ pour l'une et sur les Δ rangés dans l'ordre inverse (décroissant) pour l'autre. Les séries sont les suivantes :

$$U_i = \frac{\sum_{k=1}^{i} n_k - \frac{i(i-1)}{4}}{\sqrt{\frac{i(i+1)(2*i+5)}{72}}} \text{ et } \quad \tilde{U}_i = \frac{\sum_{k=1}^{i} \tilde{n}_k - \frac{i(i-1)}{4}}{\sqrt{\frac{i(i+1)(2*i+5)}{72}}},$$

avec $n_k = \sum_{j=1}^{k} I\left(\Delta_j < \Delta_k\right)$ et $\tilde{n}_k = \sum_{j=1}^{k} I\left(\Delta_{n-j} < \Delta_{n-k}\right)$. L'intersection entre U et $\tilde{U}$ est $\hat{k}$.

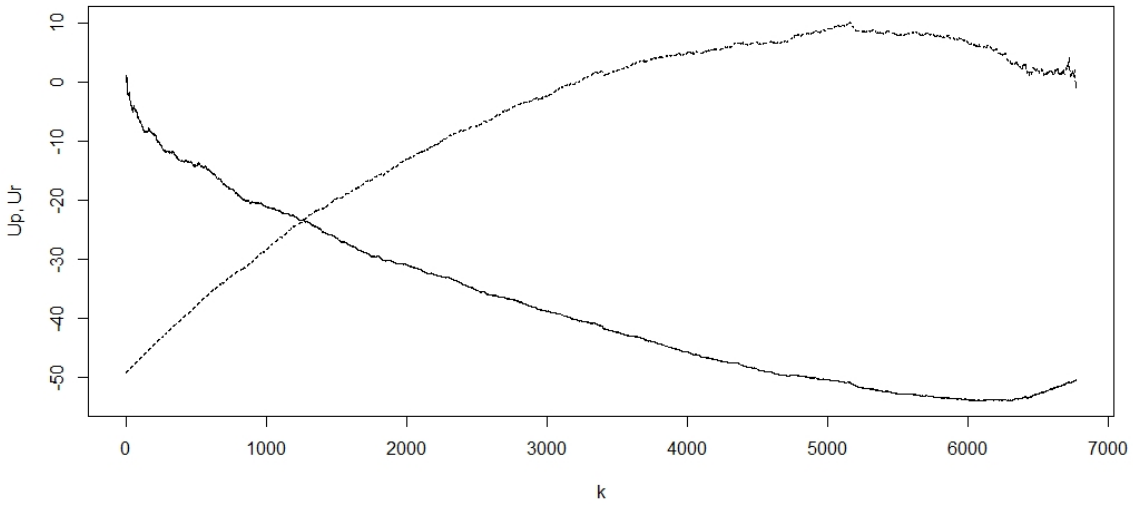


FIGURE 42 – Exemple de Gertensgarbe plot en assurance auto

Toujours sur les mêmes données *AutoClaims*, est tracé le gertensgarbe plot [32]. Le point d'intersection des deux droites est le seuil extrême.

Les résultats entre MEP, hill plot et gertensgarbe plot peuvent différer en fonction des données utilisées. Ainsi, un avis d'expert pour trancher sur la bonne valeur de k est toujours nécessaire d'autant plus qu'une méthode universelle pour estimer une valeur optimale de k est

31. $\beta_{(1)} \leq \beta_{(2)} \leq \cdots \leq \beta_{(n)}$
32. Fonction *ggplot* du package *tea* sous R.

inexistante. L'inconvénient majeur de ces méthodes réside donc dans l'interprétation graphique qui est le seul moyen proposé pour choisir un bon seuil.

3.5.3.2 Choix d'un β atypique

Pour rappel, nous avions trouvé les valeurs de $\hat{\beta}^*$ suivantes lors de notre segmentation par signe :

Risque	Groupe	Proportion	$\hat{\beta}^*$ (MSE)	$\hat{\beta}^*$ (MAE)	$\hat{\beta}^*$ (MARE)	$\hat{\beta}^*$ (MSRE)
CAT	1	0.19	-0.09	-0.07	-0.08	-0.09
	2	0.12	0.03	0.03	0.03	0.04
GTPL	1	0.03	-0.15	-0.14	-0.06	-0.06
	2	0.17	0.21	0.16	0.09	0.07
Marine	1	0.05	-0.32	-0.26	-0.08	-0.08
	2	0.05	0.12	0.06	0.06	0.09
MTPL	1	0.06	-0.5	-0.05	-0.03	-0.03
	2	0.15	0.28	0.14	0.03	0.02
PTY	1	0.05	-0.13	-0.10	-0.05	-0.06
	2	0.13	0.20	0.15	0.06	0.06

La présence de grandes différences dans la valeur de $\hat{\beta}^*$ selon les mesures d'erreurs est le signe de β atypiques. En effet, lorsque des erreurs relatives et absolues donnent les mêmes ordres de grandeur cela signifie que les points sur lesquels elles sont calculées sont équivalents en terme de montants. Le risque CAT est le seul disposant d'une stabilité dans la valeur de $\hat{\beta}^*$ peu importe la mesure choisie. Pour le groupe 1, $\hat{\beta}^*$ vaut environ -8 % et 3 % pour le groupe 2. Le groupe 1 en GTPL semble lui aussi stable avec un $\hat{\beta}^*$ aux alentours de -4 %. Le groupe 2 en Marine semble lui plus sujet à débat car la différence entre MSE et MSRE n'est que de 3 %. De plus le nombre de points est bas pour ce risque.

Les autres risques et groupes présentent eux des valeurs assez différentes entre erreurs relatives et absolues, indiquant un besoin d'une nouvelle segmentation. Cependant, le nombre de points est trop faible pour le groupe 1 des risques présentant un $\hat{\beta}^*$ volatile (GTPL, Marine et PTY). Une modélisation atypique des $\beta < 0$ n'est donc pas possible. Concentrons-nous alors dans la modélisation des $\beta \geq 0$ atypiques. Notons le groupe 1 celui des $\beta < 0$, le groupe 2 celui des $\beta \geq 0$ attritionnels et le groupe 3 celui des $\beta \geq 0$ atypiques. Pour plus de clarté, ci-dessous les caractéristiques de chaque nouveau groupe :

Groupe	Signe de β	Caractère de β
1	< 0	Divers
2	≥ 0	Attritionnel
3	≥ 0	Atypique

TABLE 20 – Caractéristiques de chaque nouveau groupe

Le besoin d'une nouvelle segmentation par groupe et par risque est donc le suivant :

Risque	Groupes après la nouvelle segmentation
CAT	1
	2
GTPL	1
	2
	3
Marine	1
	2
MTPL	1
	2
	3
PTY	1
	2
	3

TABLE 21 – Nouvelle segmentation atypique par risque

Nous passons donc à l'étude de chaque risque et groupe nécessitant une modélisation atypique.

3.5.3.2.1 Modélisation atypique de β

Une façon naïve de réaliser cette modélisation atypique est de construire individuellement un seuil atypique pour chaque risque. Les risques concernés sont le GTPL, MTPL et PTY. Il peut être plus judicieux de construire un seuil atypique pour l'ensemble de ces risques, s'ils possèdent une distribution de β proche. Pour vérifier la similitude des lois sous-jacentes, plusieurs approches sont possibles. La première méthode naturelle est de construire le boxplot des $\beta \geq 0$ de chaque risque.

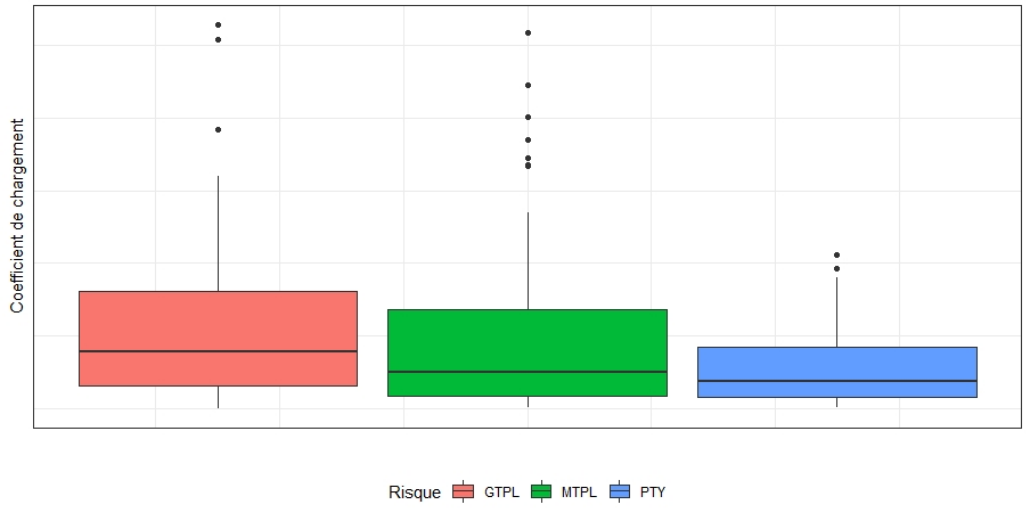


FIGURE 43 – Boxplots des $\beta \geq 0$ en GTPL, MTPL et PTY

D'emblée nous remarquons une forte similitude de distribution en GTPL et MTPL. Les quantiles et médianes sont assez proches. Le risque PTY semble avoir une médiane équivalente aux autres risques mais son maximum et son $3^{\text{ème}}$ quartile sont plus faibles. Une modélisation commune entre GTPL et MTPL semble être judicieuse tandis que le risque PTY nécessite

d'être étudié davantage pour décider s'il faut l'inclure avec les deux autres risques. Pour trancher, nous utilisons le test de Mann-Whitney-Wilcoxon.

Test de Mann-Whitney-Wilcoxon

Le test de Mann-Whitney-Wilcoxon repose sur deux hypothèses : $[H_0 : X$ et Y ont la même distribution] contre $[H_1 : X$ et Y n'ont pas la même distribution]. Il permet ainsi de tester si deux échantillons X et Y proviennent de la même distribution. Similaire au test de Student, celui-ci présente l'avantage d'être non paramétrique et ainsi de ne pas faire l'hypothèse de normalité des vecteurs. De plus, il est adapté aux échantillons contenant peu de données. Plus exactement, le test de MWW se base sur les rangs des valeurs de X et Y c'est-à-dire l'emplacement des valeurs lorsque celles-ci sont triées par ordre croissant. Pour illustrer ce test, nous utilisons les risques GTPL et MTPL. Posons $U = \{\beta \mid \beta \geq 0 \cap (\beta \in GTPL \cup \beta \in MTPL)\}$ l'ensemble des $\beta \geq 0$ en GTPL et MTPL. Les rangs permettent d'obtenir une distribution moins affectée par les valeurs extrêmes. Pour rappel, pour un vecteur A disposant de n valeurs, le rang de A_i est égal à :

$$\text{rang}(A_i) = \sum_{k=1}^{n} \mathbb{1}\{A_i \geq A_k\}.$$

Plus généralement, les rangs de nos deux risques se présentent sous cette forme :

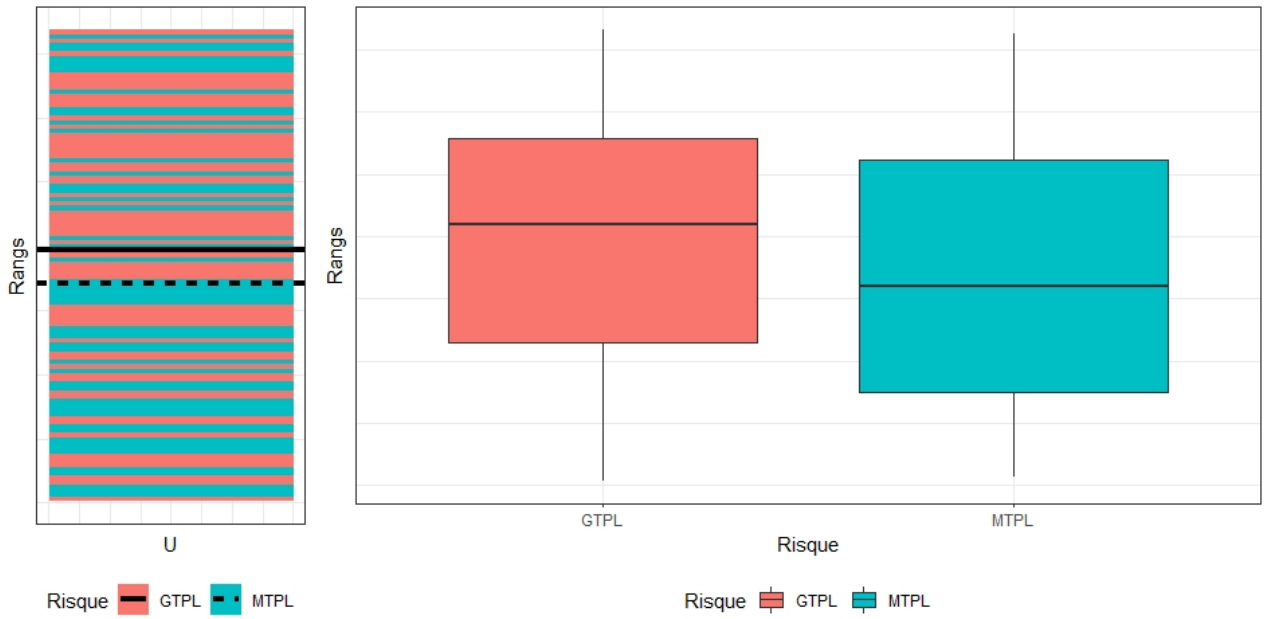


FIGURE 44 – Boxplots des rangs de $\beta \geq 0$ en GTPL et MTPL

Interprétation de la figure : Le graphique de gauche est la liste des rangs de U, la ligne noire pleine (respectivement pointillée) est le rang moyen du β en GTPL (respectivement MTPL). Le graphique de droite est le boxplot des rangs de β par risque.

Nous remarquons une certaine similarité entre les rangs de nos risques. Les rangs de U semblent uniformément répartis pour chaque risque. Les β en MTPL dominent légèrement la partie basse de U (plus de bleu que de rouge) et inversement en GTPL. Le rang moyen du risque GTPL est plus élevé que celui du risque MTPL. Notre test se base sur les deux métriques suivantes :

$$W_{GTPL} = \left[\sum_{i=1}^{n} \mathbb{1}\{U_i \in GTPL\} R_{U_i}\right] - n_{GTPL}\left(n_{GTPL}+1\right)/2$$

$$W_{MTPL} = \left[\sum_{i=1}^{n} \mathbb{1}\{U_i \in MTPL\} R_{U_i}\right] - n_{MTPL}\left(n_{MTPL}+1\right)/2$$

où R_{U_i} est le rang de U_i et n_{GTPL} le nombre de données en GTPL. La valeur W utilisée pour le test dépend de l'échantillon de référence choisi au départ. Dans notre cas, nous posons arbitrairement le risque GTPL comme référence. Calculons tout de même ces deux métriques.

Risque	n (en % du total)	$\sum R_{U_i}$	W
GTPL	54 %	3481	1711
MTPL	46 %	2624	1298

TABLE 22 – Métriques du test MWW en MTPL et GTPL

Finalement, notre statistique de test W est de 1711. Pour plus de clarté, $W = W_{GTPL}$ n'est rien d'autre que la somme du nombre de fois où le β en GTPL est supérieur à celui en MTPL. La p-value représente la probabilité d'observer une valeur de W au moins aussi élevée que celle calculée (1711) sachant H_0 (les deux échantillons suivent la même loi). Une valeur faible de p indique une improbabilité que l'expérience ne suive pas l'hypothèse nulle. Le seuil de significativité α est fixé à 5 %. Ainsi si $p > \alpha$ nous acceptons H_0 et nous supposons donc que les deux échantillons ont la même distribution. Pour calculer p il est nécessaire de simuler un grand nombre de U aléatoires en effectuant une redistribution au hasard des rangs de β entre les deux risques. La densité empirique obtenue (l'histogramme des simulations) après 10 000 simulations est la suivante :

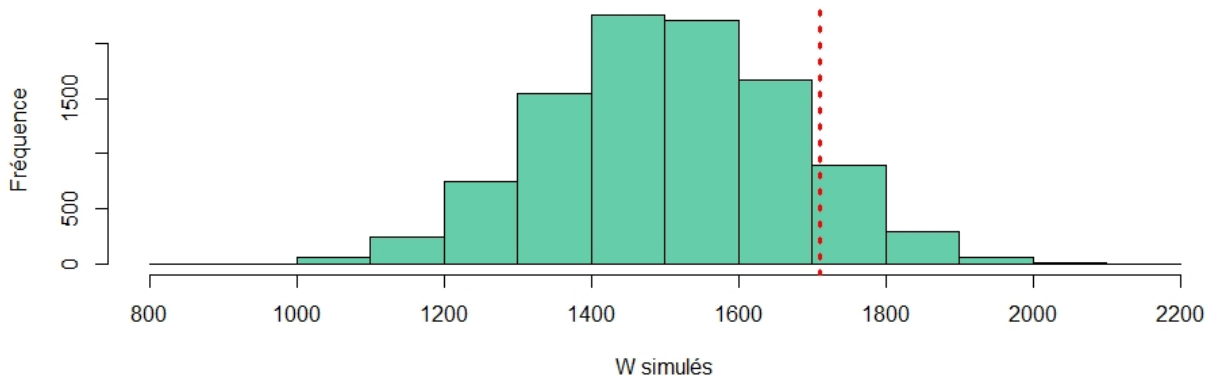


FIGURE 45 – Histogramme des simulations de W

La ligne pointillée rouge représente la *vraie* valeur de W (1711). Le test à effectuer pour calculer p est bilatéral. L'hypothèse nulle est en fait un ensemble des deux sous-hypothèses suivantes :

1. Le β en GTPL n'a pas une localisation particulièrement élevée ET
2. Le β en GTPL n'a pas une localisation particulièrement faible.

En posant $\tilde{W}$ le vecteur des W simulés, la valeur approximative de p est donc

$$p = 2 \times \frac{\sum_{i=1}^{10000} \mathbb{1}\{\tilde{W}_i \geq W\}}{10000} \simeq 0.229.$$

En fait, la distribution de $\tilde{W}$ est approximativement normale lorsque le nombre de données dans chaque échantillon est grand. Il n'est donc pas forcément nécessaire d'effectuer beaucoup de calculs pour approximer p[33] [34]. Dans notre cas, nous utilisons la densité de la statistique de Wilcox (*pwilcox* sur R) pour calculer notre p-value. La valeur de p trouvée en utilisant cette fonction est $p = 0.2178618$. Afin de s'assurer de la vraisemblance de nos résultats, nous les comparons à la fonction *wilcox.test* disponible sur R :

```
> wilcox.test(rangs~risque,data=U,exact = T)

        Wilcoxon rank sum exact test

data:  rangs by risque
W = 1711, p-value = 0.2179
```

La statistique de test W est bien égale à 1711 et notre p-value est équivalente à celle calculée avec la densité de Wilcox et est très proche de celle trouvée *à la main*. Finalement, nous pouvons accepter H_0 et conclure que la distribution des β de ces deux risques est équivalente. Les résultats de ce test pour nos trois risques sont les suivants :

Risques testés	W	p-value
GTPL et MTPL	1711	0.2179
GTPL et PTY	1783	0.01179
MTPL et PTY	1326	0.3665

TABLE 23 – Test de Wilcoxon des risques GTPL, MTPL et PTY.

Le test de même distribution est non significatif entre le GTPL et le PTY. De plus, étant donné leur grande différence de boxplot, nous supposons que la loi de β de ces deux risques est différente. Notons également que le test de MWW semble très significatif pour les risques MTPL et PTY. Nous choisissons tout de même de modéliser les β atypiques en isolant le PTY des deux autres risques. Estimons par exemple le seuil atypique des β en MTPL et GTPL. Ci-dessous les sorties graphiques du MEP, gertensgarbe plot et hill plot :

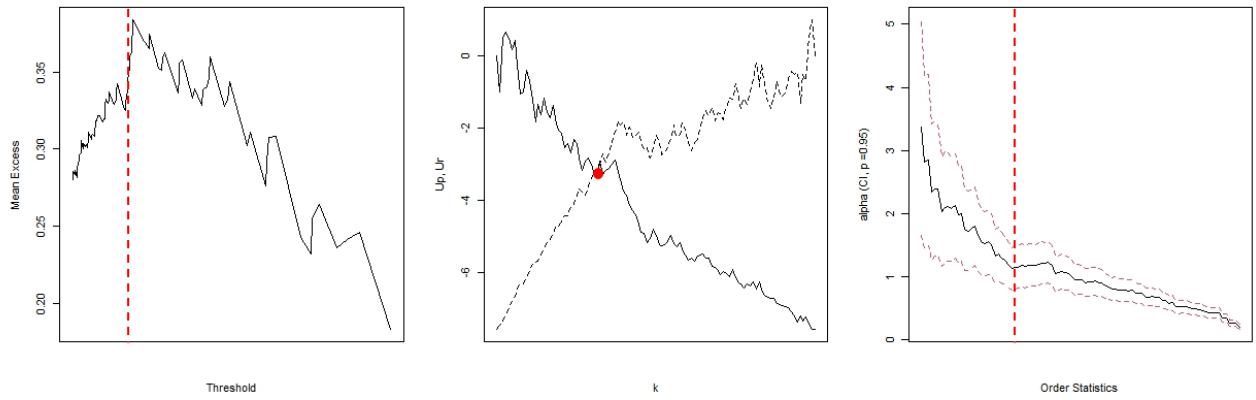


FIGURE 46 – MEP, gertensgarbe plot et hill plot des $\beta \geq 0$ en MTPL et GTPL

33. Normal Approximations to the Distributions of the Wilcoxon Statistics
34. Densité *pwilcox* de W sur R.

Interprétation du graphique : Les axes des abscisses ont été supprimés par mesure de confidentialité. Ils indiquent tous soit un rang soit une valeur de β. De gauche à droite nous retrouvons le MEP, le gertensgarbe plot et le hill plot. Les lignes en pointillés rouges représentent le seuil choisi pour le MEP et le hill plot. Le point rouge représente l'intersection des deux courbes sur le gertensgarbe plot.

Finalement, l'étude de ces trois graphiques nous donne un seuil optimal atypique différent. Nous retenons le seul atypique moyen de chaque méthode par risque. Notre nouvelle distribution de β est la suivante :

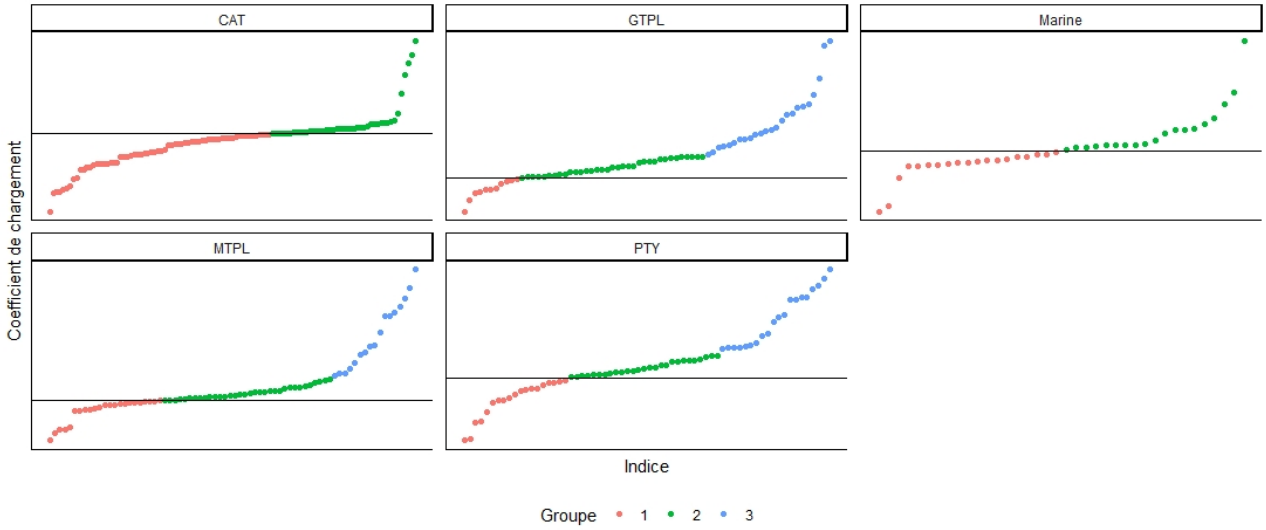


FIGURE 47 – Nuage de points des β après la nouvelle segmentation atypique

3.5.3.2.2 Estimations de $\hat{\beta}^*$ suite à la nouvelle segmentation atypique

L'approximation de $\hat{\beta}^*$ ce fait donc une nouvelle fois pour chaque risque et groupe à l'aide de notre algorithme de minimisation. Les résultats sont les suivants :

Risque	Groupe	Proportion	$\hat{\beta}^*$ (MSE)	$\hat{\beta}^*$ (MAE)	$\hat{\beta}^*$ (MARE)	$\hat{\beta}^*$ (MSRE)
CAT	1	0.19	-0.09	-0.07	-0.08	-0.09
	2	0.12	0.03	0.03	0.03	0.04
GTPL	1	0.03	-0.15	-0.14	-0.06	-0.06
	2	0.10	0.11	0.11	0.07	0.06
	3	0.07	0.51	0.42	0.38	0.40
Marine	1	0.05	-0.32	-0.26	-0.08	-0.08
	2	0.05	0.12	0.06	0.06	0.09
MTPL	1	0.06	-0.05	-0.05	-0.03	-0.03
	2	0.10	0.14	0.12	0.02	0.02
	3	0.05	1.02	0.87	0.46	0.41
PTY	1	0.05	-0.13	-0.10	-0.05	-0.06
	2	0.08	0.04	0.04	0.04	0.04
	3	0.05	0.30	0.30	0.17	0.20

TABLE 24 – $\hat{\beta}^*$ après la nouvelle segmentation atypique

Finalement, les groupes 2 et 3 en GTPL et PTY sont plutôt stables peu importe la mesure

d'erreurs. Nous remarquons que la segmentation en MTPL n'est pas assez précise. En effet, la valeur de $\hat{\beta}^*$ varie toujours selon les mesures relatives et absolues. Les MSE et MAE suggèrent une valeur de 13 % environ pour le groupe 2 tandis que le MSRE et le MARE trouvent 2 % pour $\hat{\beta}^*$. Ceci indique que la volatilité des β reste tout de même assez élevée. Néanmoins, une segmentation plus fine nuirait fortement à la qualité de prédiction des groupes. En effet, plus nous augmentons le nombre de classes plus nous diminuons le nombre de points et donc d'exemples par classe sur lesquels notre arbre de décision apprend. La segmentation atypique semble tout de même être un bon choix. Nous observons, en GTPL, des valeurs entre 38 % et 51 % pour le groupe atypique et des valeurs entre 17 % et 20 % en PTY. Le risque MTPL est celui disposant des valeurs les plus grandes avec un $\hat{\beta}^*$ de 102 % pour le MSE et de 87 % pour le MAE. Les mesures d'erreurs relatives estiment elles un $\hat{\beta}^*$ autour de 43 %.

Le choix du $\hat{\beta}^*$ final par groupe et par risque peut désormais se faire. Pour rappel, le but est de trouver un coefficient de chargement constant afin d'obtenir une formule de tarification fermée et simple pour un ensemble de traités communs. Dans notre cas, il existe quatre valeurs possibles de $\hat{\beta}^*$. Il faut donc choisir parmi ces résultats.

Une question se pose alors : quel ordre de grandeur de β souhaitons-nous privilégier ? Lorsque $\hat{\beta}^*$ est stable pour nos quatre mesures d'erreurs cette question ne se pose pas puisque les ordres de grandeurs sont identiques. Cependant, lorsque cette stabilité n'est pas acquise malgré une segmentation fine (comme pour le groupe 3 en MTPL) le choix final de $\hat{\beta}^*$ porte sur un choix plus subjectif. Opter pour le MSE ou le MAE indique un souhait de privilégier les primes commerciales les plus élevées en chargement et donc d'accepter une moins bonne estimation des traités bénéficiant de tarifs faiblement chargés. Le choix des mesures relatives indique que nous souhaitons une approximation plutôt satisfaisante peu importe le montant de la cotation moyenne du traité. Dans notre cas, nous préférons opter pour le $\hat{\beta}^*$ calculé par MARE car cette mesure d'erreur est celle la plus précise en moyenne, peu importe l'ordre de grandeur de la cotation. Notre $\hat{\beta}^*$ final par risque et par groupe est donc :

Risque	Groupe	$\hat{\beta}^*$
CAT	1	-0.08
	2	0.03
GTPL	1	-0.06
	2	0.07
	3	0.38
Marine	1	-0.08
	2	0.06
MTPL	1	-0.03
	2	0.02
	3	0.46
PTY	1	-0.05
	2	0.04
	3	0.17

TABLE 25 – $\hat{\beta}^*$ final par risque et par groupe

Ainsi, nous estimons notre prime commerciale avec les coefficients de chargement estimés ci-dessus. Le rendu est le suivant :

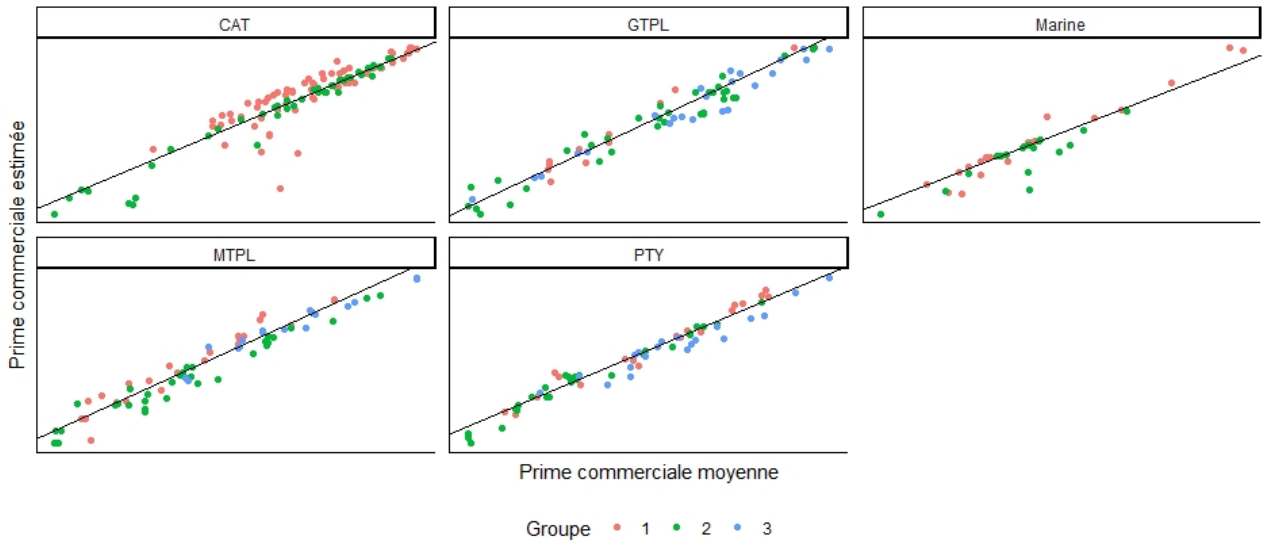


FIGURE 48 – Estimation de la cotation moyenne après segmentation par groupe

À noter que pour faciliter la lecture des primes faibles, l'échelle en abscisse et en ordonnée est en log10. Nous observons une bonne approximation de la prime cotée moyenne pour tous les traités. Il existe certains points où nous sous-estimons la prime commerciale, notamment en CAT et en Marine. Ceci est le résultat espéré puisque nous fixons un β constant par groupe. Il existe donc toujours des points mal estimés bien que nous remarquons un bon alignement autour de la droite médiane, et ce, pour tous les ordres de grandeur de primes. Ceci montre une certaine robustesse et flexibilité de l'algorithme autour de la grande variété de primes et de coefficients. La dernière étape de cette procédure est d'entraîner un nouvel arbre de décision pour prédire notre nouveau groupe atypique.

3.5.3.3 Prédiction du groupe par CART

Nous conservons les mêmes données que celles utilisées dans la partie 3.5.2.3. La division entre apprentissage et test est de 75 % - 25 %, avec 15-fold cross validation. La précision obtenue par profondeur est la suivante :

Profondeur maximale	Précision	Kappa	Précision SD	Kappa SD
1	0.5950	0.3125	0.0667	0.1070
5	0.6285	0.3970	0.0967	0.1482
7	0.6100	0.3698	0.1017	0.1584
12	0.6100	0.3698	0.1017	0.1584
13	0.6100	0.3698	0.1017	0.1584

TABLE 26 – Performances du CART en prédiction des groupes 1, 2 et 3

La profondeur optimale est atteinte lorsque celle-ci vaut 5 avec une précision de 62.85 %. Ce score obtenu est assez décevant, nous obtenons une baisse de 10 points environ par rapport au CART prédisant les groupes 1 et 2 avec une profondeur de 5 au lieu de 3, rendant notre CART plus complexe que l'ancien. Aussi, l'ajout d'un groupe, qui de plus est atypique, nuit nécessairement à la prédiction car nous diminuons le nombre d'exemples en augmentant le nombre de classes. Analysons tout de même la matrice de confusion du modèle :

Prédit \ Réel	1	2	3
1	23.8	7.8	4.5
2	11.2	32.7	6.3
3	3.7	3.7	6.3

TABLE 27 – Matrice de confusion du CART pour les groupes 1, 2 et 3

En sommant la diagonale de la matrice nous retrouvons bien notre précision moyenne de 62.8 % sur les 15 validations croisées. Le groupe 2 est encore celui qui bénéficie de la meilleure estimation avec 32.7 prédictions correctes contre $7.8 + 3.7 = 11.5$ fausses soit une précision d'environ 74 % (contre 90 % dans le CART précédent). Le groupe 1 obtient lui une précision de 61.5 % tandis que le 3 n'est prédit correctement que 36.9 % du temps. Ce résultat s'explique de plusieurs façons.

Tout d'abord le groupe 3 est atypique par essence. Ainsi le nombre de données pour cette classe est bien plus faible que les autres. De plus, celui-ci ne semble pas disposer de variables ayant des tendances suffisantes pour expliquer le caractère du coefficient de chargement associé à ce groupe. L'arbre peut donc avoir plus de difficultés à élaborer des seuils pertinents. De plus, lorsque le groupe réel est le 3, l'arbre prédit la valeur 1 4.5 fois soit environ 35 % $(4.5/(6.3 + 6.3 + 4.5))$ des prédictions.

L'arbre associe donc un β négatif au lieu d'un coefficient positif extrême ce qui est une valeur complètement opposée à celle réelle. En général, une prédiction d'un β négatif sachant qu'il est positif (ie groupes réels 2 et 3 et groupe prédit égal à 1) se présente dans 20 % des cas. Pour obtenir cette proportion, il faut sommer la part prédite de 1 dans les cas où le groupe est 2 ou 3 soit $7.8 + 4.5 = 12.3$ puis diviser par la somme des colonnes réelles du groupe 2 et 3 soit 61.3 ce qui donne $12.3/61.3 \simeq 20$ %. Ainsi, l'arbre prédit 2 fois sur 10 un coefficient négatif alors qu'il devrait être positif. Cette erreur de prédiction peut fortement nuire à la pertinence de la prime commerciale estimée avec ce coefficient. L'arbre obtenu est le suivant :

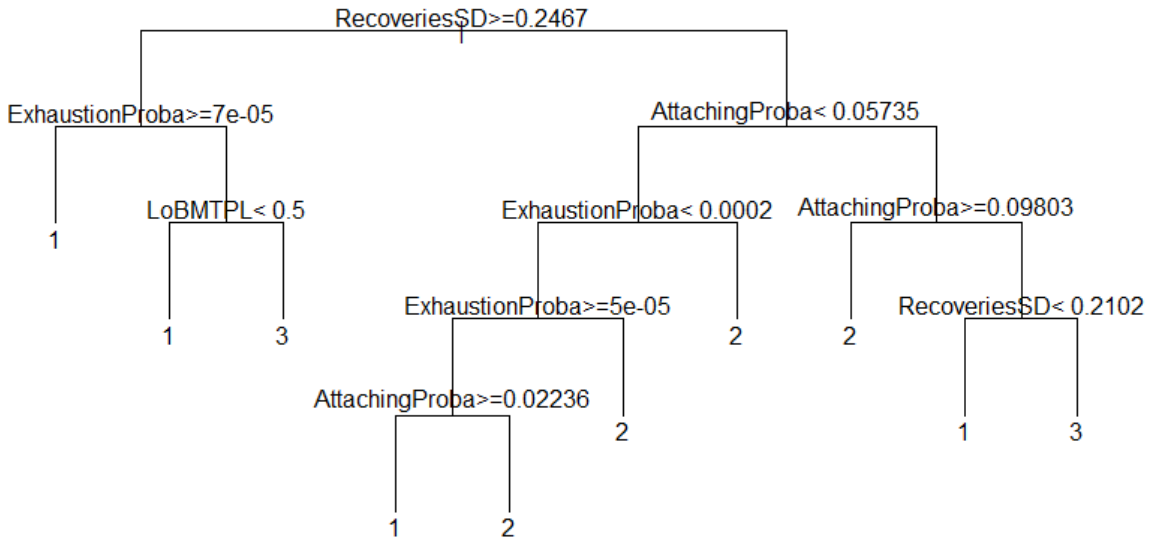


FIGURE 49 – CART en classification du groupe 1, 2 et 3

Lecture du CART : Pour chaque nœud, si le test est positif nous parcourons le fils gauche.

Par exemple, si $RecoveriesSD \geq 0.2467$*, nous testons le fils gauche égal à* $ExhaustionProba \geq 7e-05$*. De plus, le risque MTPL fait partie d'un nœud. Le test effectué sur celui-ci est* $LoBMTPL < 0.5$ [35] *résultant de la transformation par One Hot Encoding* [36] *c'est-à-dire en la construction d'une colonne binaire pour chaque risque. Ainsi,* $LoBMTPL < 0.5$ *signifie que le risque MTPL est égal à 0 donc que le risque n'est pas MTPL.*

Cet arbre d'une profondeur maximale de 5 indique plusieurs tendances. Tout d'abord, un β atypique peut être prédit si par exemple la récupération standardisée est supérieure à 0.2467 et que la probabilité d'épuisement est inférieure à $7e-05$. Si le traité respecte ces deux conditions et qu'il est en MTPL, le groupe 3 est prédit. Ceci peut expliquer les mauvaises prédictions de ce groupe car il n'y a que 3 risques sur 5 qui disposent de groupes atypiques. Ainsi, si le traité testé est en CAT mais qu'il respecte les deux conditions évoquées précédemment, il sera tout de même classé en groupe 3. Enfin, nous remarquons que le groupe 2, qui dispose du plus grande nombre de prédictions, possède autant de feuilles que le groupe 1 et moitié moins que le groupe 3. Analysons plus en détails la performance du modèle sur la base de validation.

Prédit \ Réel	1	2	3
1	17.1 %	4.5 %	1.1 %
2	15.9 %	34.1 %	9.1 %
3	5.7 %	5.7 %	6.8 %

	Groupe 1	Groupe 2	Groupe 3
Sensitivité	44.1 %	76.9 %	40.0 %
Spécificité	90.7 %	55.1 %	86.3 %

TABLE 28 – Performance finale du CART en classification des groupes 1, 2 et 3

La précision finale obtenue par cette matrice est la somme de la diagonale soit 58 %. La sensitivité par classe représente la probabilité de bien prédire cette classe sachant que l'individu d'origine était bien de cette classe. Par exemple, pour le groupe 2, ce calcul est le rapport des bonnes prédictions (34.1 %) divisé par le nombre d'individus de ce groupe (4.5 % + 34.1 % + 5.7 % = 44.3 %) soit une sensitivité égale à 76.9 % (34.1/44.3). La spécificité est équivalente à une erreur de type 2 en théorie des tests.

Cette erreur représente la probabilité de ne pas prédire cette classe sachant que l'individu n'y appartient pas. Pour le groupe 1, la spécificité est la probabilité de prédire 2 ou 3 sachant que l'observation appartient bien au groupe 2 ou 3. Plus clairement, c'est la probabilité de ne pas prédire 1 sachant qu'il ne fallait pas prédire 1. Dans notre cas, elle se calcule en divisant la part des mauvaises prédictions de 1 (4.5 % + 1.1 % = 5.6 %) par la part totale d'individus des groupes 2 et 3 (somme des deux colonnes *réel* 2 et 3, soit 55.7 %) soit une valeur d'environ 90.9 %. Celle-ci est légèrement différente de la valeur de référence du tableau 28 à cause de l'arrondi dans les pourcentages de données de la matrice de confusion.

Finalement, le groupe 2 a une bonne précision de prédiction avec 76.9 % de classifications correctes. Cependant, dans quasiment un cas sur deux ce groupe est prédit alors qu'il fallait prédire 1 ou 3. Ainsi, 55.6 % des prédictions du groupe 2 sont en réalité 1 ou 3. Notre arbre est donc performant pour la prédiction de cette classe mais il la prédit bien trop souvent au

35. LoB signifie *Line of Business* (le risque)
36. https ://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/

total. Le groupe 3 ne dispose pas d'une bonne précision (40 %) mais sa spécificité est bonne. Ceci signifie que 86.4 % des observations différentes de 3 ne sont pas prédites dans le groupe 3. Ainsi, lorsque 3 n'est pas la classe prédite par l'arbre, nous sommes plutôt confiants sur le fait que l'individu n'appartient pas au groupe 3.

En conclusion, la méthode par chargement constant avec prédiction du groupe par CART s'avère transparente et performante sur l'estimation du coefficient de chargement. Cependant, la prédiction du groupe pour un nouveau traité n'est pas nécessairement précise. Afin de palier à ce problème, une amélioration possible est de sélectionner un algorithme de prédiction du groupe plus performant mais bien plus boîte noire (comme les forêts aléatoires). Aussi, il peut être possible de trouver une méthode permettant d'identifier des groupes communs et de calculer directement le $\hat{\beta}^*$ sur ceux-ci, sans prédiction. Une façon de réaliser cette segmentation est d'utiliser des algorithmes de machine learning non supervisé car ils permettent de trouver les points milieux de clusters. Ainsi, la méthode de détermination d'un groupe pour un nouveau traité n'est plus une estimation mais complètement déterministe. Cependant, l'estimation de $\hat{\beta}*$ par groupe n'est plus aussi précise.

3.5.4 Segmentation par groupes homogènes (k-means)

Comme observé auparavant, la détermination de clusters en fonction du type de chargement nécessite une prédiction pour tous les nouveaux traités à tarifer. Cette méthode de tarification devant rester simple et assez transparente, l'algorithme idéal de prédiction est le CART. Cette approche montre de bons résultats dans la tarification mais peine à prédire avec précision le type de chargement à utiliser pour un nouveau traité. Ainsi, nous obtenons une formule de tarification par écart type optimale pour plusieurs sous-groupes mais il est probable que nos nouveaux traités ne soient pas correctement associés à leur formule.

Pour répondre à ce problème de prédiction, il peut être intéressant d'étudier un algorithme de clustering déterministe nous permettant à l'avenir de supprimer l'aléa sur le bon groupe d'un traité. La méthode potentiellement capable de répondre à nos besoins est celle des k-means introduite auparavant. Comme pour les KNN, cet algorithme se base sur une distance euclidienne. Ainsi, il est uniquement possible de donner en entrée des descripteurs numériques. Le nombre de données par risque n'étant pas suffisant, il est nécessaire de réaliser un modèle général pour tous les risques. Nous nous basons donc uniquement sur les descripteurs qui ont été introduits dans ce tableau 15. Avant de modéliser, il est essentiel de mettre nos données à la même échelle en **normalisant par minimum et maximum** [37]. Cette procédure permet de ramener un ensemble de valeurs entre 0 et 1. Posons $X = X_1, \ldots, X_n$ un vecteur de n valeurs numériques. Alors, la valeur normalisée $\bar{X}_i$ de X_i est égal à

$$\bar{X}_i = \frac{X_i - \min(X)}{\max(X) - \min(X)}.$$

Cette normalisation permet d'*effacer* les différences d'échelles entres variables et ainsi donner le même poids à chacune d'entre elles dans le calcul des distances. Elle est indispensable pour utiliser l'algorithme des k-means. Afin de trouver un nombre optimal de clusters, commençons par utiliser la méthode du coude. Pour rappel, celle-ci consiste à minimiser l'inertie (ou la somme de la variance intra-clusters) définie par :

37. Concernant la limite du traité, nous n'utilisons pas les traités ayant une limite infinie car la normalisation ne serait pas adaptée. Leur cluster sera attribué ultérieurement.

$$Wtot_k = \sum_{m=1}^{k} W(C_m) = \sum_{m=1}^{k} \sum_{i:X_i \in C_m} \sum_{j=1}^{n} (X_{i,j} - c_j)^2$$

La valeur optimale est atteinte lorsque $Wtot_k$ est faible et que $Wtot_{k+1}$ est proche de $Wtot_k$ ($Wtot_{k+1} - Wtot_k < \epsilon$ avec $\epsilon > 0$ un seuil à fixer). Généralement, la valeur de ϵ n'est pas fixée, nous effectuons simplement une analyse graphique du coude c'est-à-dire lorsque ϵ semble être petit. Effectuons cette méthode sur nos données.

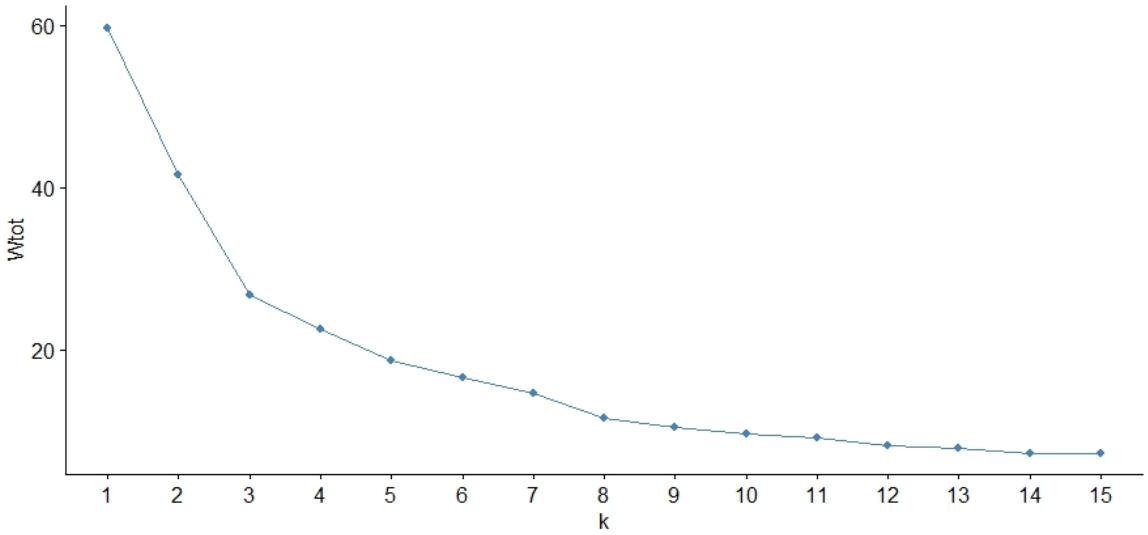


FIGURE 50 – Nombre optimal de clusters par méthode du coude

La localisation du coude ne semble pas évidente car elle n'est pas très nette. La valeur de 8 pour k semble tout de même être un seuil à partir duquel $Wtot$ diminue moins vite. Ainsi, par cette méthode, 8 clusters sont optimaux. À cela, nous couplons la méthode de la silhouette introduite auparavant. Pour rappel, le but est de maximiser le coefficient de la silhouette défini comme la somme des silhouettes de chaque cluster. Posons C_k le $k^{\text{ème}}$ cluster avec :

— la distance moyenne a du point X_i aux autres individus de son cluster C_k :

$$a(C_k, X_i) = \frac{1}{\#C_k - 1} \sum_{j \in C_k, j \neq i} \|X_i, X_j\|_2$$

avec $\#C_k$ l'effectif de C_k.

— la distance moyenne b du point X_i au cluster le plus proche :

$$b(C_k, X_i) = \min_{k' \neq k} \frac{1}{\#C_{k'}} \sum_{i' \in C_{k'}} \|X_i, X_{i'}\|_2$$

Le but est donc de minimiser b et de maximiser a pour un point X_i en particulier. Sa silhouette $s(C_k, X_i)$ est finalement égale à :

$$s(C_k, X_i) = \frac{b(C_k, X_i) - a(C_k, X_i)}{\max(a(C_k, X_i), b(C_k, X_i))}$$

Ainsi, la silhouette totale $stot_k$ est la somme des silhouettes moyennes individuelles de chaque cluster :

$$stot_k = \frac{1}{k} \sum_{m=1}^{k} \frac{1}{\#C_m} \sum_{i \in C_m} s(C_m, X_i).$$

Nous calculons $stot_k$ pour k variant de 1 à 15.

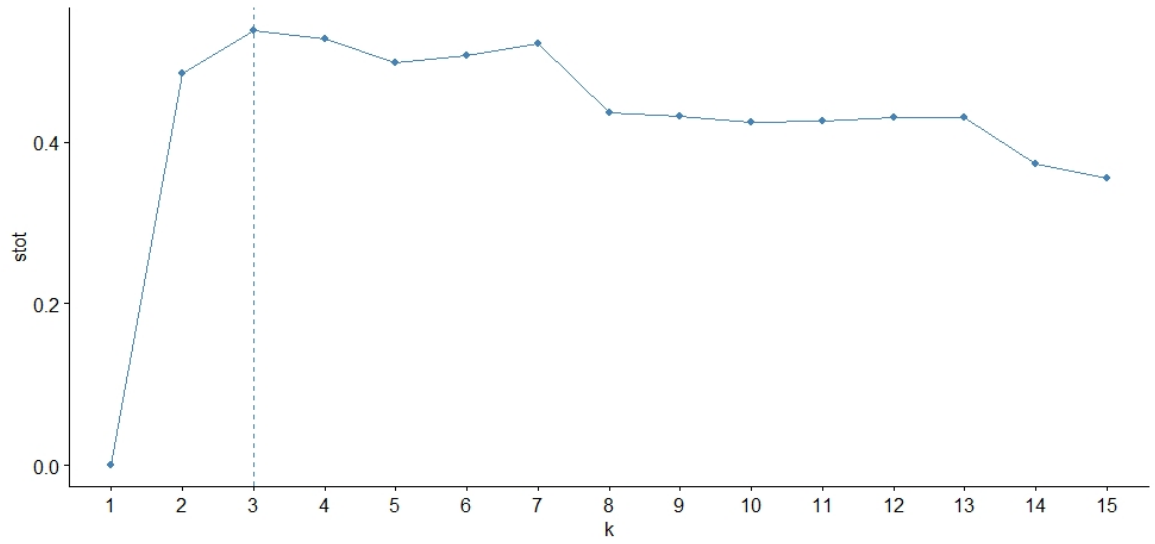


FIGURE 51 – Nombre optimal de clusters par méthode de la silhouette

Le nombre optimal de clusters est de 3. Ce résultat est très différent de celui obtenu par la méthode du coude où $k = 8$. Cependant, nous remarquons que la valeur de *stot* pour $k = 3$ est très proche de celle où $k = 7$. De plus, $Wtot$ est assez proche entre $k = 7$ et $k = 8$. Nous choisissons donc de faire un compromis entre ces deux méthodes afin d'obtenir un nombre optimal de clusters plus robuste en posant $k = 7$. Entraînons alors notre algorithme des k-means sur nos données. Afin de visualiser les caractéristiques de chaque cluster, nous pouvons afficher la **heat map** (littéralement la carte des chaleurs) de nos centroïdes.

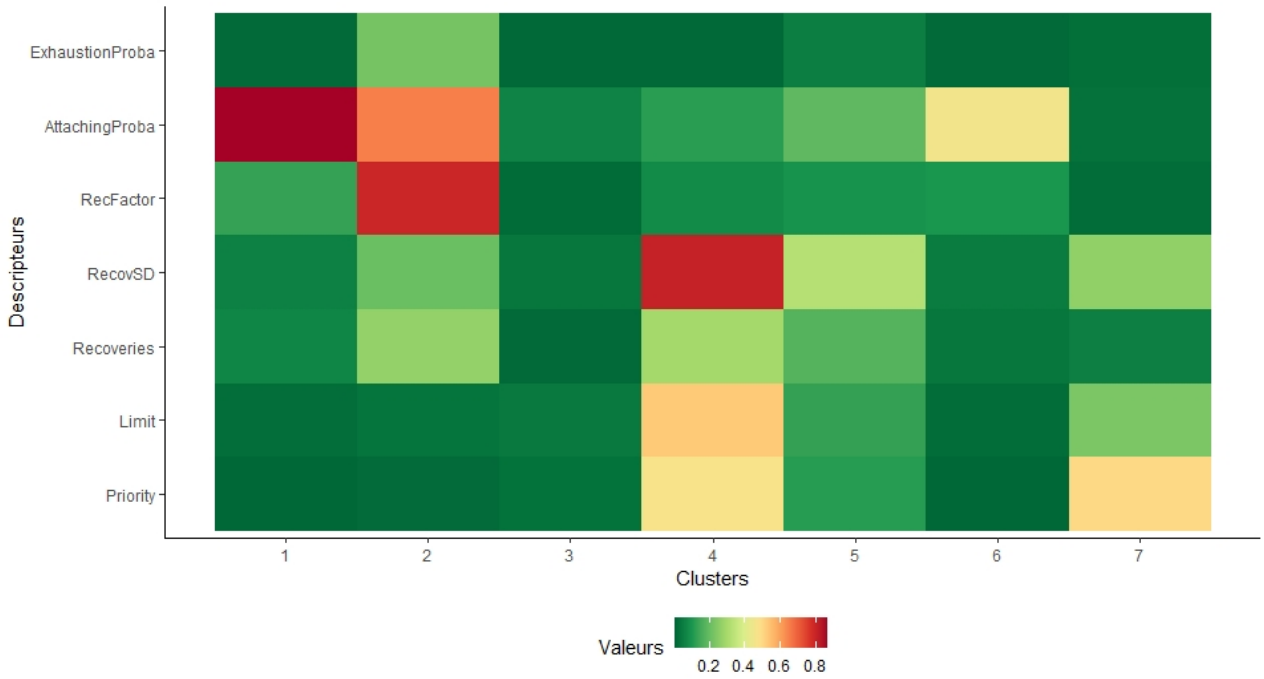


FIGURE 52 – Heat map par cluster après application des k-means

Interprétation du graphique : En abscisse de la heat map se trouvent les clusters. En ordonnée nous retrouvons les variables données en input des k-means. Une couleur rouge indique que la valeur est forte par rapport aux autres valeurs de la même variable. Les variables étant normalisées par colonne, celles-ci vont de 0 à 1. Par exemple, le centroïde du cluster 2 est celui possédant une des probabilités d'attachement les plus fortes (orange foncé) avec le cluster 1 (rouge). Le $2^{ème}$ cluster dispose aussi d'un facteur de reconstitution plutôt élevé.

La heat map représente l'*intensité* du centroïde de chaque cluster en fonction des différents descripteurs fournis aux k-means. Tout d'abord, nous observons que le cluster 1 est celui possédant la probabilité d'attachement la plus élevée et est proche du maximum observé dans les données (car sa couleur est rouge foncé). Ses autres variables étant plus en vert clair, ceci indique que le centroïde a une limite et une priorité assez faibles. Les récupérations moyennes (*Recoveries*) et l'écart type des récupérations (*RecovSD*) sont eux aussi proches des faibles valeurs observées dans les données. Le facteur de reconstitution quant à lui est plutôt proche de 0.3, il est donc d'ordre moyen-faible.

Finalement, le cluster 1 réunit des traités ayant pour caractéristiques communes d'avoir un RMP, des récupérations moyennes, un écart type, un facteur de reconstitution et une probabilité d'épuisement plutôt faibles avec une probabilité d'attachement très élevée. Ceci semble assez naturel puisqu'une couverture de réassurance avec une limite et une priorité faibles a de fortes chances d'être touchée par des sinistres. De plus, la couverture étant faible, les récupérations sont elles aussi assez basses. Ceci dépend tout de même du facteur de reconstitution qui est quant à lui moyennement faible.

Le cluster 4 est assez opposé au cluster 1. Il regroupe des traités très volatiles ayant un RMP plutôt important (jaune-beige) avec des récupérations moyennes autour de 0.5 soit une valeur médiane. Les autres variables sont faibles. Le cluster 2 se distingue lui aussi par son caractère élevé sur la probabilité d'attachement et sur le facteur de reconstitution. Ses récupérations moyennes et l'écart type sont aussi assez moyens tandis que le RMP est très faible. Les

traités considérés ici sont donc ceux avec un RMP bas mais des récupérations et un écart type proches de 0.3 soit une valeur moyenne-faible. Ainsi, ce type de traité dispose nécessairement d'une probabilité d'attachement forte et d'un facteur de reconstitution important car les récupérations sont élevées comparées au RMP. Les autres clusters sont d'autres combinaisons des variables mais dans des ordres de grandeurs moyens et faibles.

La fréquence par cluster est la suivante :

Cluster	Fréquence
1	5 %
2	3 %
3	56 %
4	4 %
5	10 %
6	17 %
7	5 %

TABLE 29 – Fréquence par cluster après k-means

Le cluster majoritaire est le 3 avec 56 % des données. Sur la heat map, celui-ci présente des descripteurs faibles. Ainsi, nous pouvons conclure qu'une partie importante de nos traités disposent de valeurs assez faibles comparées aux autres clusters. Les clusters 5 et 6 représentent quant à eux des traités avec des descripteurs moyens et bas mais restant tout de même très peu extrêmes. À eux trois (clusters 3, 5 et 6), ils composent nos observations standards avec 83 % des données. Les autres clusters sont assez atypiques par leur comportement extrême. La part des données comprise dans ces clusters est donc faible. Le but étant d'estimer β par cluster, nous appliquons une nouvelle fois notre algorithme de minimisation sur chaque cluster. Pour rappel, le coefficient final choisi est une nouvelle fois celui estimé par MARE.

Cluster	Proportion	$\bar{\beta}$	β médian	σ_β	$\hat{\beta}^*$ (MSE)	$\hat{\beta}^*$ (MAE)	$\hat{\beta}^*$ (MARE)	$\hat{\beta}^*$ (MSRE)	$\hat{\beta}^*$
1	0.05	-0.07	-0.08	0.28	-0.01	-0.05	-0.06	-0.11	-0.06
2	0.03	-0.11	-0.41	0.58	-0.52	-0.53	-0.41	-0.13	-0.41
3	0.56	0.13	0.06	0.23	0.02	0.02	0.01	0.00	0.01
4	0.04	-0.03	0.01	0.09	-0.03	0.01	0.02	0.00	0.02
5	0.10	-0.08	-0.06	0.12	-0.09	-0.06	-0.08	-0.10	-0.08
6	0.17	0.10	0.04	0.37	0.08	0.08	-0.10	-0.10	-0.10
7	0.05	0.03	0.03	0.04	0.02	0.03	-0.01	0.00	-0.01

TABLE 30 – $\hat{\beta}^*$ après segmentation par k-means

Les résultats sont très différents selon les clusters. Pour rappel, une segmentation par k-means dans l'optique d'estimer un coefficient de chargement par cluster nécessite une tendance homogène et stable de β pour chaque cluster créé. Cela ne semble pas être toujours le cas.

Le cluster 1 dispose d'un coefficient de chargement estimé ligne à ligne moyen ($\bar{\beta}$) égal à -7 % et un β median de -8 %. Le $\hat{\beta}^*$ final estimé par MARE est proche et vaut -6 %, indiquant une distribution de β potentiellement homogène pour ce cluster. Par ailleurs, un $\hat{\beta}^*$ de -0.1 % par MSE et une valeur de 0.28 pour σ_β indiquent qu'il existe des $\beta \geq 0$ dans ce cluster. Le cluster 2 est celui possédant la plus grande volatilité du coefficient de chargement avec $\sigma_\beta = 0.58$. Le coefficient final estimé $\hat{\beta}^*$ est quant à lui égal à la médiane.

Le 3ème cluster est nécessairement un de ceux qui souffre le plus de la qualité d'estimation. Le nombre de points étant important, il existe une grande variété dans les valeurs de β. Ainsi, l'algorithme considère que ne pas charger est la meilleure solution ($\hat{\beta}^* = 0.01 \simeq 0$). Ce résultat est induit par le mélange de différents signes de β au sein du même cluster, avec une volatilité non négligeable de 0.23.

Le cluster 6 se démarque par sa forte différence dans la valeur et le signe de $\hat{\beta}^*$ entre erreurs relatives et absolues. Cette différence vient de la grande variété de risques (quatre en tout) et des ordres de grandeurs des $\hat{\beta}^*$ en fonction de la mesure d'erreur. Ce cluster contient des primes élevées nécessitant un coefficient de chargement grand. Ainsi, une erreur relative a tendance à porter beaucoup plus de poids sur ces valeurs du fait des ordres de grandeur forts. D'un autre côté, les erreurs relatives portent plus d'importance sur les erreurs faibles en pourcentage donc aussi sur les petites primes. Or, celles-ci nécessitent un chargement négatif soit une valeur négative pour le MSRE et MARE. Cette différence de signe indique simplement que ce cluster est vraisemblablement homogène au sens des k-means mais que cela ne suffit pas à créer une tendance stable de β. Enfin, les clusters 5 et 7 sont plutôt bien estimés avec un $\hat{\beta}^*$ proche de celui médian et moyen.

Pour palier à ce manque de robustesse des k-means, nous pouvons envisager plusieurs solutions :

— **Augmenter le nombre de clusters :** notre nombre actuel de clusters permet de gagner en transparence et en simplicité. Cependant, ce faible nombre a un coût sur la qualité d'estimation du chargement de certains clusters induit par le manque d'homogénéité de β.
— **Ajouter des variables :** les k-means imposent une distance euclidienne. Cet inconvénient majeur nous empêche d'utiliser d'autres informations comme la risque ou la région du traité. L'idée est donc de changer la façon de calculer la distance.

Un moyen d'inclure ces deux solutions est d'utiliser le **PAM**.

3.5.5 Segmentation par groupes homogènes (PAM)

Utiliser des variables numériques et catégorielles dans l'optique de créer des clusters nous apporte un gain d'information important. Une méthode permettant d'utiliser différents types de données est le **Partitioning Around Medoids** (PAM) basé sur la distance de **Gower**.

3.5.5.1 Distance de Gower

L'idée de la distance de Gower est assez naturelle. Pour chaque variable, une distance particulière conforme à son type est utilisée, prenant des valeurs entre 0 et 1. Une combinaison linéaire (souvent une moyenne) est utilisée pour calculer la matrice finale des distances. Les métriques sont les suivantes :

— **Variable continue numérique :** la distance de Manhattan normalisée est utilisée.
— **Variable catégorielle non ordonnée :** ce type de variable est par exemple un pays, un risque ou une région soit un type de variable qualitatif mais qui ne contient pas d'ordre (à la différence d'une note sur 20 par exemple). La distance est une fonction indicatrice.

Finalement, la distance de Gower est généralement définie comme la moyenne des différences partielles entre les individus :

$$D_{Gower}\left(X_{1,.}, X_{2,.}\right) = \frac{1}{p}\sum_{j=1}^{p} s_j\left(X_{1,j}, X_{2,j}\right)$$

où s_j est la fonction des différences partielles qui dépend du type de la variable j et p le nombre total de variables. Ainsi, pour les variables continues :

$$s_j\left(X_{1,j}, X_{2,j}\right) = \frac{|X_{1,j} - X_{2,j}|}{R_j}$$

où R_j est le coefficient de normalisation égal à $\max_i X_{i,j} - \min_i X_{i,j}$ (*longueur* de la $j^{\text{ème}}$ variable). Pour les variables qualitatives, cette distance est :

$$s_j\left(X_{1,j}, X_{2,j}\right) = \mathbb{1}\{X_{1,j} \neq X_{2,j}\}.$$

Exemple de calcul de la distance de Gower

Prenons un jeu de données fictif donnant des informations sur trois personnes.

	Age	Yeux	Taille	Sexe	Salaire
	30	bleu	160	1	3000
	27	marron	180	1	1600
	24	vert	170	2	1400
R	*6*		*20*		*1600*

TABLE 31 – Données fictives de trois personnes

R est le coefficient de normalisation introduit précédemment.
La distance entre l'individu 1 et 2 est :

— **Age :** $|30 - 27|/6 = 0.5$
— **Yeux :** $\mathbb{1}\{\text{bleu} \neq \text{marron}\} = 1$
— **Taille :** $|160 - 180|/20 = 1$
— **Sexe :** $\mathbb{1}\{1 \neq 1\} = 0$
— **Salaire :** $|3000 - 1600|/1600 = 0.875$
— **Distance de Gower :** $(0.5 + 1 + 1 + 0 + 0.875)/5 = 0.675$

Nous calculons donc cette distance entre tous nos individus, la matrice des distances est de la forme :

$$M_{(Gower,X)} = \begin{pmatrix} D_{Gower}(X_{1,.}, X_{1,.}) & D_{Gower}(X_{1,.}, X_{2,.}) & \cdots & D_{Gower}(X_{1,.}, X_{n,.}) \\ D_{Gower}(X_{2,.}, X_{1,.}) & \ddots & \cdots & D_{Gower}(X_{2,.}, X_{n,.}) \\ \vdots & \vdots & \ddots & \vdots \\ D_{Gower}(X_{n,.}, X_{1,.}) & \cdots & \cdots & D_{Gower}(X_{n,.}, X_{n,.}) \end{pmatrix}$$

$M_{(Gower,X)}$ est donc une matrice symétrique et sa diagonale est égale à 0. Nous calculons cette matrice sur nos données qui nous permettent d'obtenir la distance entre chaque traité. L'algorithme de clustering que nous allons appliquer à cette matrice est donc le PAM.

3.5.5.2 Partitioning Around Medoids (PAM)

L'algorithme PAM est similaire à celui des k-means à l'exception près que PAM utilise des médoïdes. Un médoïde est une observation de référence. Elle est différente du centroïde des k-means qui est le barycentre des clusters mais qui n'existe pas en tant qu'observation. Le médoïde est lui une vraie observation qui joue le rôle de l'individu de référence du cluster dans le sens où c'est ce point qui minimise les distances entre lui et les autres individus du cluster. L'algorithme fonctionne en quatre étapes :

1. Choisir k médoïdes aléatoirement dans les observations.
2. Assigner chaque individu au médoïde le plus proche afin de former k clusters en utilisant la matrice des distances $M_{(Gower,X)}$.
3. Pour chaque cluster, choisir l'observation qui minimise la distance entre elle et les autres points. Celle-ci est le nouveau médoïde.
4. Si au moins un médoïde est réassigné, retourner à l'étape 2. Sinon, l'algorithme est terminé.

Ces étapes sont très similaires à celles des k-means. Le PAM possède l'avantage d'être plus polyvalent et il permet de choisir une donnée réelle comme centre du cluster. Néanmoins, le temps de calcul est quadratique (complexité en $O(n^2)$). Pour sélectionner le nombre de clusters, nous utilisons une nouvelle fois les méthodes de la silhouette et du coude :

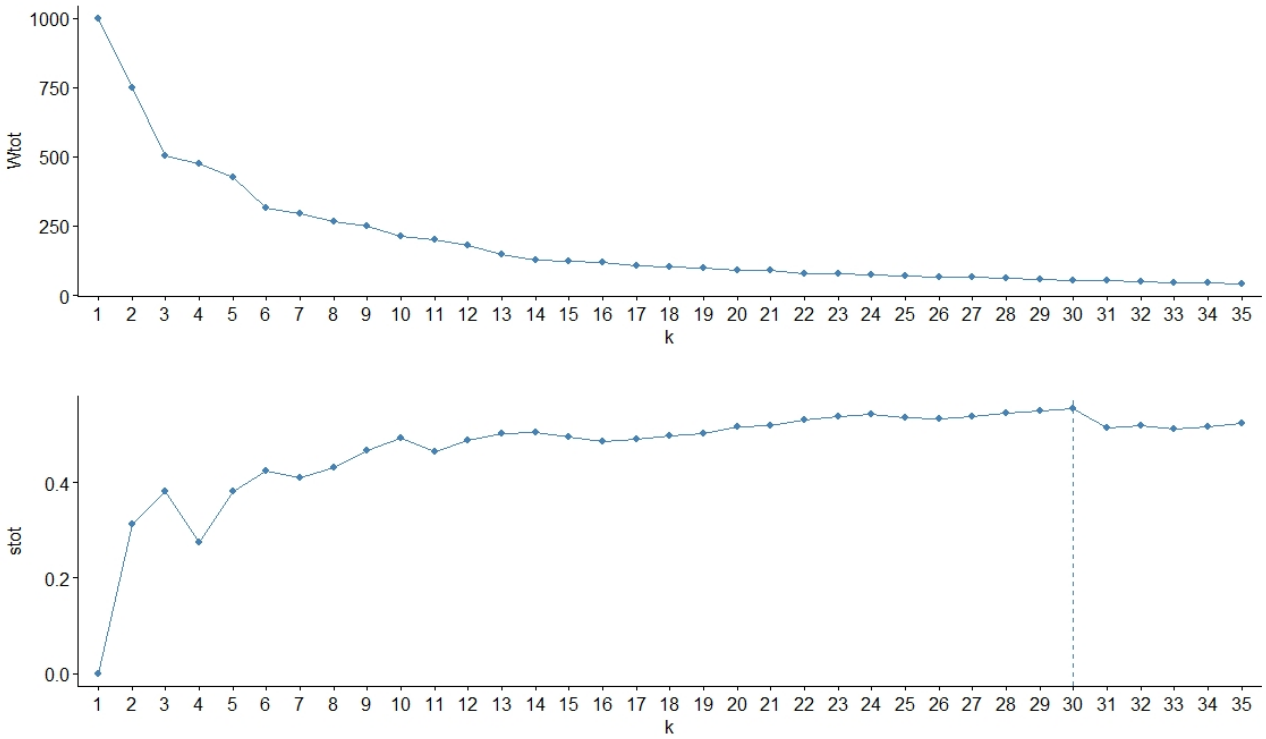


FIGURE 53 – Méthode du coude et de la silhouette en PAM

Le nombre optimal de clusters est bien plus important qu'en k-means. La méthode du coude (graphique du haut) semble afficher un k optimal de 14 tandis que celle de la silhouette (graphique du bas) donne $k = 30$. Cependant, le gain de *stot* entre 14 et 30 est assez faible. Ainsi nous choisissons $k = 14$, soit deux fois plus de clusters qu'auparavant. Afin d'observer le

comportement interne de chaque cluster, affichons une nouvelle fois la heat map. Par ailleurs, cette fois nous n'allons plus afficher la heat map du centroïde mais plutôt la heat map du comportement moyen de chaque cluster.

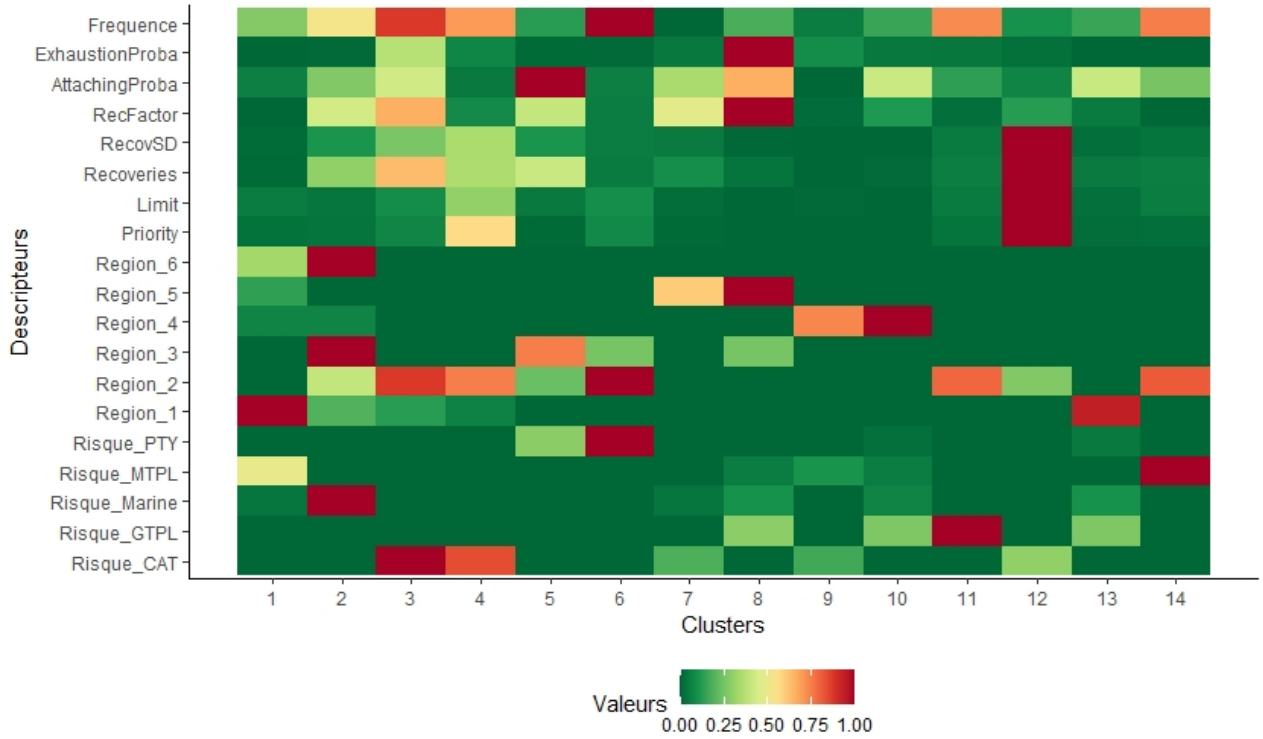


FIGURE 54 – Heat map par cluster après application du PAM

Pour des raisons de confidentialité, la variable région est numérotée. Chaque région correspond à une partie du monde, par exemple à l'Europe ou l'Amérique du Sud. La variable fréquence est le nombre d'observations par cluster.

Nous pouvons observer une fréquence bien mieux répartie entre chaque cluster. Le cluster 1 contient en majorité des traités MTPL et une forte proportion de la région 1 (la plus grande de tous les clusters). Ses autres variables sont plus faibles et comprises entre 0 et 0.25. Le cluster 3 quant à lui contient une bonne proportion de données car la fréquence est de couleur orange foncé. Le risque majoritaire est le CAT avec principalement la région 2. Les traités de ce cluster sont plutôt assez sinistrés avec une récupération moyenne et un facteur de reconstitution autour de 0.6 soit légèrement au-dessus de la valeur médiane.

Le cluster 12 est le plus extrême avec le RMP, les récupérations moyennes et l'écart type des récupérations les plus forts. Ainsi, il regroupe des traités avec les tranches les plus hautes mais aussi ceux qui sont les plus sinistrés et volatiles. La proportion de données est quant à elle assez faible (vert clair soit 0.10 environ). Le cluster contenant le plus de données est le 6. Les régions majoritaires sont la 2 et la 5. Les autres variables de ce cluster sont assez proches du minimum, elles sont entre 0 et 0.25.

Finalement, cet algorithme nous permet de déceler bien plus d'informations entre clusters en augmentant leur nombre et les descripteurs choisis. À présent, nous appliquons notre méthode de minimisation sur chaque cluster. Les résultats sont les suivants :

Cluster	Proportion	$\bar{\beta}$	β median	σ_β	$\hat{\beta}^*$ (MSE)	$\hat{\beta}^*$ (MAE)	$\hat{\beta}^*$ (MARE)	$\hat{\beta}^*$ (MSRE)	$\hat{\beta}^*$
1	0.06	0.14	0.03	0.21	0.11	0.06	0.02	0.00	0.02
2	0.08	0.04	0.01	0.31	-0.32	-0.26	-0.05	-0.05	-0.05
3	0.12	-0.13	-0.13	0.20	-0.13	-0.08	-0.15	-0.14	-0.15
4	0.10	-0.01	-0.00	0.06	-0.03	-0.01	-0.02	-0.04	-0.02
5	0.04	-0.02	-0.04	0.20	-0.02	-0.10	-0.10	-0.09	-0.10
6	0.14	0.10	0.05	0.15	0.10	0.04	0.03	0.02	0.03
7	0.02	-0.12	-0.17	0.15	-0.16	-0.16	-0.17	-0.16	-0.17
8	0.04	-0.04	0.07	0.57	0.31	0.06	0.05	-0.02	0.05
9	0.03	0.22	0.07	0.28	0.04	0.05	0.00	-0.01	0.00
10	0.04	0.03	-0.05	0.19	0.01	-0.06	-0.03	0.00	-0.03
11	0.11	0.31	0.20	0.33	0.16	0.15	0.11	0.09	0.11
12	0.04	-0.01	0.01	0.06	-0.01	0.01	0.01	0.00	0.01
13	0.04	0.14	0.11	0.19	0.21	0.21	0.08	0.05	0.08
14	0.13	0.15	0.06	0.36	0.08	0.05	-0.01	-0.02	-0.01

TABLE 32 – $\hat{\beta}^*$ après segmentation par PAM

Une nouvelle fois, tous les clusters ne sont pas aussi stables dans l'estimation de β. Ils le sont néanmoins majoritairement. Le $\hat{\beta}^*$ final (obtenu par MARE) semble plutôt bien suivre la valeur réelle médiane du β de certains clusters. De plus, les erreurs sont elles aussi assez stables entre elles. Naturellement, les clusters volatiles, avec un σ_β important, sont ceux bénéficiant des plus grands écarts entre erreurs relatives et absolues.

Par exemple le cluster 8 avec un σ_β de 0.57 présente de grandes différences entre MSE (0.31) et MSRE (-0.02). Ce résultat s'explique par une grande volatilité, un $\bar{\beta}$ négatif et un β médian égal à 7 %. Globalement, l'ajout de clusters et de variables permet une meilleure estimation de $\hat{\beta}^*$. Néanmoins, notre clustering par PAM ne semble pas être suffisant pour expliquer intégralement les valeurs sous-jacentes de β par cluster. Tous ne sont pas bien estimés mais nous avons l'avantage de d'ores et déjà connaître les clusters non adaptés à cette méthode de tarification. En effet, le cluster d'un nouveau traité n'a pas besoin de prédiction mais simplement d'un calcul déterministe de distance (celle de Gower) déjà connue et transparente dans sa méthode. Cette méthode est bien plus sûre que celle où chaque groupe était prédit par CART. En effet, il existait, selon les groupes, un fort aléa sur la qualité de prédiction. Ici, ce n'est plus le cas.

Nous pouvons donc présélectionner les clusters *valides* en termes de qualité d'estimation. Une façon de faire est d'estimer la cotation moyenne (à partir de $\hat{\beta}^*$) et de calculer le MARE entre la cotation estimée et la cotation moyenne réelle par cluster. Ainsi, pour chaque cluster c, nous calculons :

$$MARE_c = \frac{1}{\#c} \sum_{i=1}^{n} \frac{|\text{cotation estimée}_i - \text{cotation moyenne}_i|}{\text{cotation moyenne}_i} \times \mathbb{1}\{\text{traité}_i \in c\}$$

Par des critères subjectifs, nous pouvons créer des tranches de qualité d'estimation basées sur cette erreur d'estimation. Nous considérons plusieurs niveaux de fiabilité par valeur de MARE :

— $0 \leq MARE \leq 0.25$ indique une bonne fiabilité.
— $0.25 < MARE \leq 0.4$ indique une fiabilité correcte.
— $0.4 < MARE \leq 0.6$ indique une fiabilité faible.
— $0.6 < MARE < \infty$ indique une mauvaise fiabilité.

Ces niveaux sont déterminés graphiquement après comparaison des valeurs de MARE et des estimations par cluster. Le résultat final par cluster est le suivant :

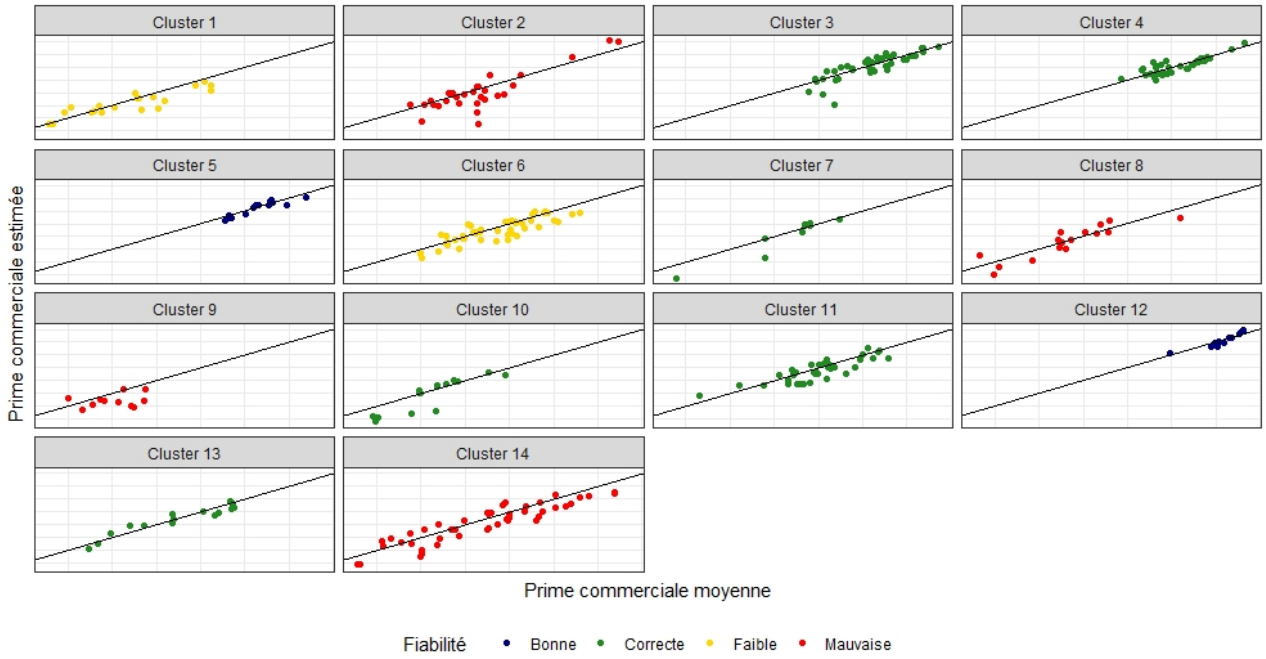


FIGURE 55 – Estimations des cotations par cluster après application du PAM

Les abscisses et ordonnées sont en échelle log10. Nous observons que les clusters 12 et 5 sont ceux disposant de la meilleure estimation mais leur nombre de traités est assez faible. Inversement, les clusters 2 et 14 ne sont pas fiables mais rassemblent bien plus de traités. Les clusters 11 et 3 quant à eux ont une fiabilité moyenne avec un nombre important de traités. Finalement, si nous n'exploitons que les clusters ayant un bon et moyen niveau de fiabilité, nous ne retenons pas les clusters 1, 2, 6, 8, 9 et 14 soit 6 clusters sur 14 (environ 42 %).

Une façon d'améliorer cette méthode réside donc une nouvelle fois dans un choix judicieux du cluster. Comme nous l'avons observé, un groupe basé sur le type du chargement β permet une bonne estimation de celui-ci en minimisant le nombre de groupes. Néanmoins, l'inconvénient majeur réside dans la faible précision de prédiction du groupe. Une amélioration possible est d'utiliser une méthode plus complexe mais potentiellement plus sujette à l'effet boîte noire. Or, ici nous cherchons à privilégier la transparence. Une partition basée sur les descripteurs sans prise en compte de β (PAM) permet de palier à ce souci de prédiction mais suppose que chaque cluster dispose d'une homogénéité de β mais comme nous l'observons, ce n'est pas toujours le cas. Malgré cela, nous pouvons déjà avoir connaissance des clusters qui ne sont pas bien estimés et ainsi assumer de ne pas tarifer les traités de ces groupes. Afin de maximiser notre qualité d'estimation de β nous pouvons essayer une autre approche : prédire directement β à l'aide d'un algorithme de prédiction.

3.6 Méthode par prédiction directe de β

Comme évoqué précédemment, la transparence et la simplicité du clustering utilisé précédemment a un impact négatif direct sur la qualité d'estimation de β. L'approche par chargement constant présente l'avantage d'être simple et naturelle. Cependant, elle n'a de sens que si les ordres de grandeur des chargements ainsi que le signe de β sont semblables. De plus, imposer un β constant permet une formule de tarification générale pour un grand nombre de traités. Malheureusement ce choix affecte notre qualité d'estimation de la cotation moyenne.

Idéalement, chaque nouveau traité à tarifer devrait bénéficier d'un coefficient de chargement unique et personnalisé. En effet, un traité est en réalité basé sur des conditions de marché assez différentes comme le nombre de réassureurs le cotant ou encore les *types* de réassureur en eux-mêmes. Les *petits réassureurs*[38] n'interviennent pas nécessairement sur les mêmes traités que les *réassureurs importants*. Ainsi, certains disposent d'une meilleure diversification/mutualisation avec un potentiel de tarif attractif plus ou moins fort. Tout ceci nous laisse à penser que l'imposition d'un chargement constant peut nuire à la qualité des estimations. Une idée naturelle est alors d'entraîner un algorithme à la prédiction du coefficient de chargement. Deux approches sont possibles : paramétrique et non paramétrique. La partie paramétrique repose sur un GLM tandis que celle non paramétrique utilise une forêt aléatoire.

3.6.1 Prédiction de β par GLM

L'utilisation du modèle linéaire généralisé repose sur une connaissance de la loi sous-jacente de nos données. Nous travaillons une nouvelle fois sur le coefficient de chargement estimé sur la cotation moyenne. En effet, il peut exister plusieurs cotations pour un même traité. Si nous gardons toutes les cotations telles quelles nous obtenons une base de la forme :

$$\begin{array}{cccc} X_{1,1} & \dots & X_{1,p} & Y_1 \\ X_{1,1} & \dots & X_{1,p} & Y_2 \\ X_{1,1} & \dots & X_{1,p} & Y_3 \\ \vdots & \vdots & \vdots & \vdots \\ X_{4,1} & \dots & X_{4,p} & Y_{15} \\ \vdots & \vdots & \vdots & \vdots \\ X_{m,1} & \dots & X_{m,p} & Y_n \end{array}$$

avec X les descripteurs et Y la variable cible (ici la cotation). Par exemple, nous observons dans ce cas trois Y_i différents pour l'individu $X_{1,\cdot}$ soit un cas de doublons. L'algorithme met alors bien plus de poids sur les individus disposant d'un nombre de cotations important pour un même $X_{i,\cdot}$. Poussée à l'extrême, notre approche estimera très bien uniquement les traités avec plusieurs cotations. Ceci explique finalement pourquoi il est nécessaire de moyenner la cotation Y par $X_{i,\cdot}$ bien cela nous fasse perdre l'information de la volatilité intra-cotation. Nous considérons cependant que prédire une prime commerciale proche de la moyenne des futures cotations est un objectif suffisant.

La première étape avant la modélisation est de choisir la famille de lois appropriée à notre coefficient de chargement. Tout d'abord le support est nécessairement compris dans $\mathbb{R}$ puisque β prend des valeurs négatives et positives. Cela restreint d'emblée notre choix de famille de lois. Pour rappel, le GLM modélise une relation entre le label Y et un ensemble de descripteurs $\boldsymbol{X} = X_1, \dots, X_n$. Plus précisément, le modèle estime une espérance conditionnelle de la forme

38. En termes de chiffre d'affaires

$E(Y \mid \boldsymbol{X}) = \mu = g^{-1}(\eta)$. En terme probabiliste, nous pouvons définir le GLM comme une estimation de

$$Y \mid \boldsymbol{X} \sim f\left(\mu, \sigma^2\right)$$

Il est important de noter que la densité ne dépend pas seulement de Y mais de Y sachant X. Ainsi le choix de la famille de lois dépend de notre connaissance du lien entre Y et $\boldsymbol{X}$. Si notre label est continu dans $\mathbb{R}$ le choix par défaut qui s'impose à nous est la distribution Normale. Le second choix est celui de la fonction de lien, si celui-ci est la fonction *identité* alors nous sommes en régression linéaire. Explorons alors le lien entre Y et $\boldsymbol{X}$ à travers un triangle de corrélation entre notre label et nos descripteurs numériques. A noter qu'il est possible de calculer une corrélation entre une variable catégorielle et une variable numérique si la catégorie est binaire (aussi appelée le *point-biserial correlation coefficient*). Nous appliquons donc un procédé de one hot encoding sur nos variables qualitatives afin d'observer les corrélations avec notre taux de chargement. La distribution de chaque variable n'étant pas nécessairement normalement distribué, nous utilisons les corrélations de Spearman [39] qui donnent des résultats sensiblement équivalents aux corrélations usuelles de Pearson :

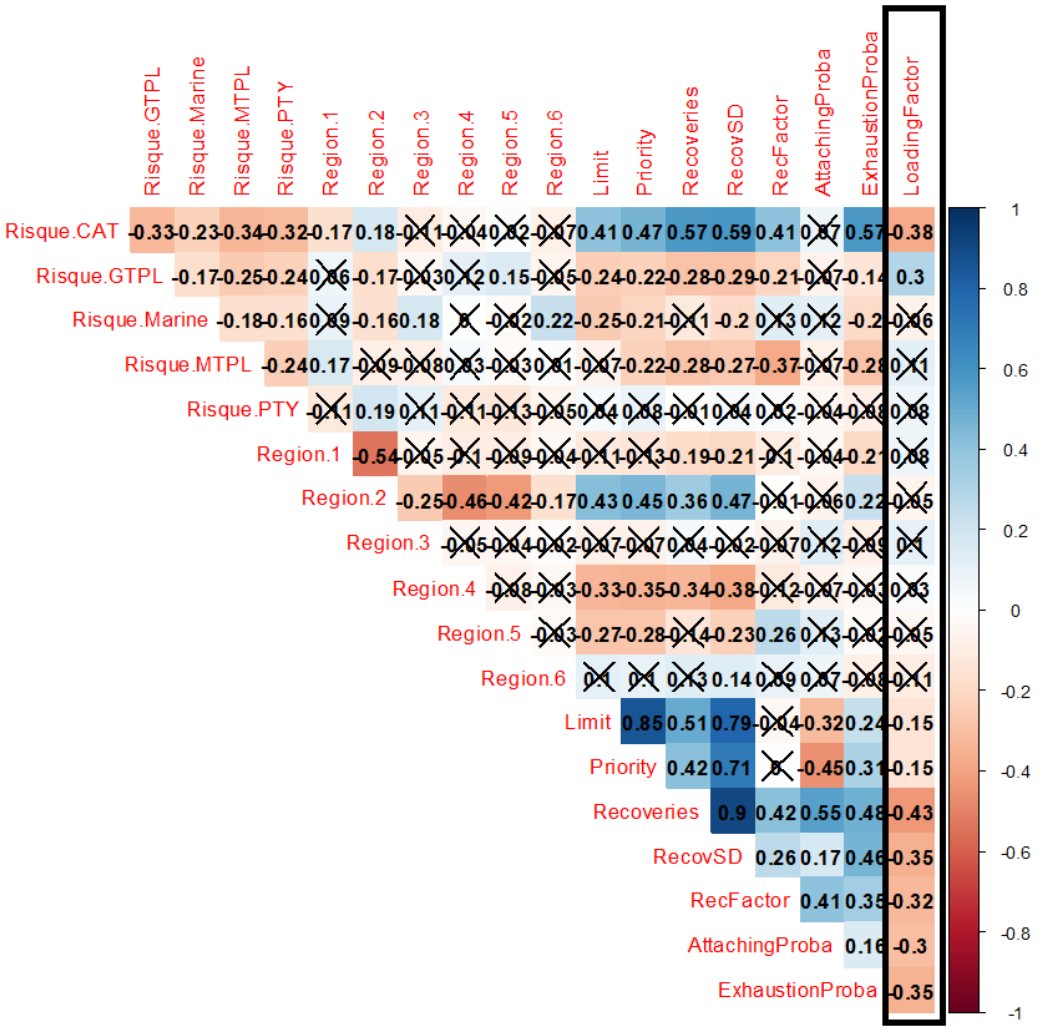


FIGURE 56 – Triangle de corrélation pour l'étude du lien label-descripteurs

39. $r = \frac{\text{cov}(\text{rang}_X, \text{rang}_Y)}{\sigma_{\text{rang}_X} \sigma_{rang_Y}}$

La colonne qui nous intéresse ici est la dernière (celle encadrée en noir). Elle montre les corrélations entre notre label et chaque descripteur. Nous observons une majorité de corrélations négatives. Tous nos descripteurs quantitatifs sont négativement corrélés avec notre label. Il n'y a que le risque GTPL qui est corrélé positivement avec le coefficient de chargement. Nous remarquons aussi que la quasi-totalité des risques et régions semble indépendante du loading factor. Le risque CAT est le seul à avoir une corrélation négative forte avec le loading factor. Ce constat était attendu puisque nous avions déjà remarqué que le β de ce risque avait tendance à être plus faible que pour les autres risques.

Ainsi, en général, β est plus faible lorsque $Risque.CAT = 1$. Cela s'explique aussi par le fait que nous n'avions pas détecté d'outliers pour ce risque et qu'il existait une forte proportion de $\beta < 0$ comparée aux quatre autres risques. Les cases du triangle marquées par une croix sont celles où la p-value du coefficient de corrélation est < 5 %. Dans ce cas, nous supposons $r = 0$ (pas de corrélation). Il semble exister un lien général de corrélations négatives entre Y et $\boldsymbol{X}$. Cela indique que lorsqu'un X_i augmente, Y diminue. De plus, les variables numériques sont aussi corrélées positivement entre elles. Cela indique qu'il existe une tendance générale entre nos descripteurs.

Ainsi, il semble possible de supposer un lien **inverse** entre Y et $\boldsymbol{X}$ ce qui signifie qu'une augmentation de $\boldsymbol{X}$ entraîne une diminution de Y. Cependant, le lien inverse étant défini par $\mu = g^{-1}(\eta) = 1/\eta$, il risque de ne pas fonctionner, et ce, pour plusieurs raisons. Il peut exister des cas où η est faible, notamment pour les traités bas et peu sinistrés. Ainsi, $1/\eta$ risque d'être assez grand et donc loin d'un coefficient de chargement qui est un pourcentage. En effet, les valeurs de β sont toutes comprises dans un intervalle de valeurs très faibles comparé à $\boldsymbol{X}$. Ainsi, un lien plus adapté est donc le lien canonique qui peut permettre d'estimer des valeurs correctes de β. Nous sommes donc en régression linéaire avec intercept. Le GLM choisi est donc un GLM gaussien avec lien identité.

Nous l'entraînons donc sur nos données, en remplaçant une nouvelle fois la limite et la priorité par le RMP ainsi que la récupération moyenne et l'écart type par la récupération moyenne standardisée (nommée ici *RecupStandardise* au lieu de *RecoveriesSD*). De plus, nous remplaçons la variable région par le pays. En effet, cette information étant plus précise il est intéressant de l'utiliser dans ce contexte. Nous ne l'affichons cependant pas sur le triangle de corrélation car elle contient trop de modalités ce qui compliquerait la lecture du graphique. Nous nommons ce descripteur *ExpoGeo* pour exposition géographique car parfois le pays est plutôt un ensemble de pays (région). Par confidentialité, celui-ci prend des valeurs numériques (1 pour France par exemple).

Enfin, nous choisissons notre modèle par minimisation de l'AIC grâce à la méthode *glmStepAIC* qui effectue toutes les combinaisons possibles entre les différentes variables. Le modèle final est disponible en annexe ici [78]. Il est entraîné sur une base d'apprentissage correspondant à 75 % des données comme dans les sections précédentes.

L'AIC obtenu vaut -16.62 avec un $R^2 = 0.07865117$, un RMSE de 0.3057044 et un MAE de 0.2002567 soit une erreur moyenne absolue d'environ 20 %. Pour rappel, le RMSE est la racine du MSE soit

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

Les variables les plus significatives sont le risque, l'exposition géographique, la probabilité d'épuisement et les récupérations standardisées. Les résultats obtenus sont peu convaincants. Le R^2 est faible et le MAE assez élevé bien que l'ordre de grandeur entre les β soit assez stable. Les fitted values (labels estimés par le modèle) sont donc mal modélisées. De plus, les résultats sur l'échantillon de test suivent la même tendance avec un MAE d'environ 18 % et un RMSE de 22 %.

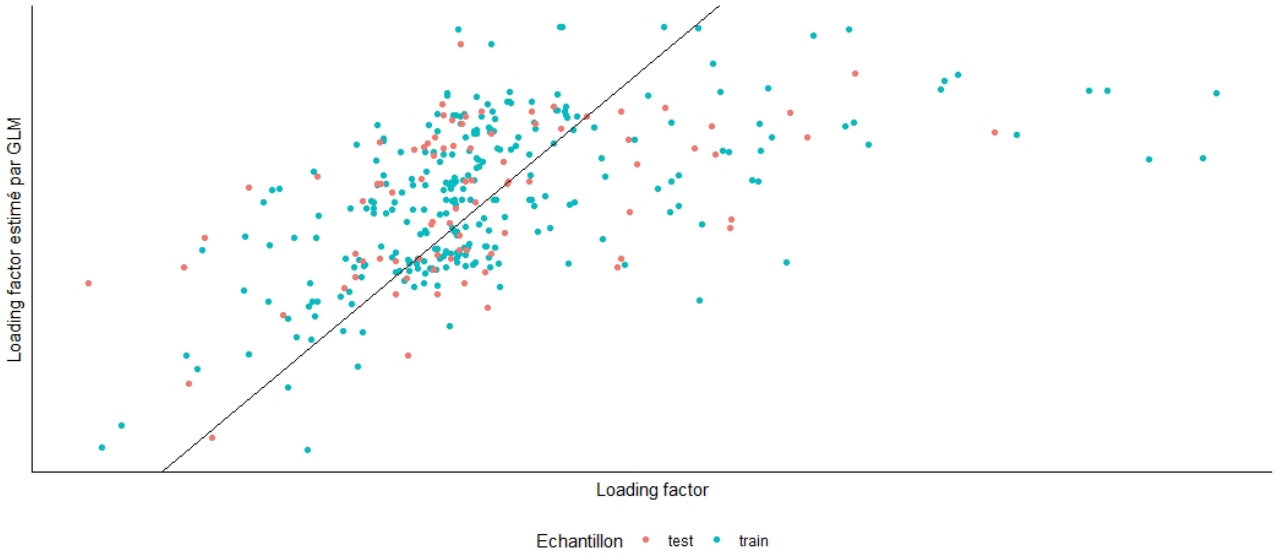


FIGURE 57 – Fitted values du GLM de β

Nous observons une grande variation des prédictions autour de la ligne médiane, ce qui est un indicateur d'une mauvaise performance du modèle. Nous observons de nombreux points à droite de cette ligne médiane. Cela nous informe que nos prédictions sont souvent sous-estimées par rapport au réel coefficient de chargement. Finalement, il n'y a pas d'alignement global des points autour de la droite. Le modèle est donc sous-performant et n'est pas gardé. Un modèle paramétrique gaussien est donc clairement inadapté à notre problématique. L'hypothèse de famille Normale semble être bien trop forte. Il faut donc envisager d'utiliser un modèle sans cette supposition de la loi de $Y \mid \boldsymbol{X}$. Un candidat intéressant est le random forest.

3.6.2 Prédiction de β par forêt aléatoire

L'algorithme des forêts aléatoires est une approche pertinente pour modéliser β. En effet, il ne requiert aucune hypothèse de loi entre Y et $\boldsymbol{X}$ comme le GLM. La seule supposition probabiliste est le caractère indépendant et identiquement distribué de chaque paire $(\boldsymbol{X}_i, Y_i)$. Nous supposons ainsi que $(\boldsymbol{X}_i, Y_i) \underset{\text{iid}}{\sim} P$, avec P une loi inconnue. Ceci signifie que chaque ligne de notre jeu de données est issue du même phénomène (les modélisations des traités) et chacune d'entre elle est générée indépendamment des autres. Cela semble être tout à fait le cas.

Les variables utilisées sont les mêmes que celles du GLM. Le paramètre à optimiser dans une forêt aléatoire est le nombre de variables à sélectionner dans chaque arbre. Communément nommé $mtry$, ce nombre est le même pour tous les arbres. Posons m le nombre total de variables de $\boldsymbol{X}$. Ici, la valeur de m n'est pas le nombre de colonnes de $\boldsymbol{X}$ car nous disposons de descripteurs catégoriels. Ainsi, pour une variable catégorielle disposant de k modalités, il y a k colonnes binaires dans $\boldsymbol{X}$. Estimons donc notre $mtry$ optimal par 10-cross validation. Nous fixons le nombre d'arbres $ntree$ à 1000. Le temps de calcul est bien plus long que dans la modélisation

du GLM. En effet, un nombre d'arbres important associé à de la cross validation demande nécessairement un temps de calcul conséquent. Cette problématique opérationnelle peut être un frein à son utilisation si l'utilisateur du modèle se trouve dans le besoin de l'entraîner souvent. Le résultat obtenu est le suivant :

mtry	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
2	0.2551	0.3828	0.1668	0.0677	0.2291	0.0323
5	0.2318	0.4414	0.1517	0.0521	0.2498	0.0291
9	0.2300	0.4393	0.1500	0.0535	0.2627	0.0297
12	0.2329	0.4219	0.1510	0.0531	0.2649	0.0287
16	0.2356	0.4085	0.1530	0.0555	0.2714	0.0290
19	0.2372	0.4009	0.1540	0.0572	0.2774	0.0289
23	0.2399	0.3859	0.1549	0.0580	0.2723	0.0281
26	0.2384	0.3901	0.1547	0.0576	0.2776	0.0275
30	0.2399	0.3888	0.1550	0.0613	0.2782	0.0292
34	0.2416	0.3812	0.1560	0.0610	0.2776	0.0280

TABLE 33 – Nombre optimal de variables de la forêt aléatoire dans la prédiction de β

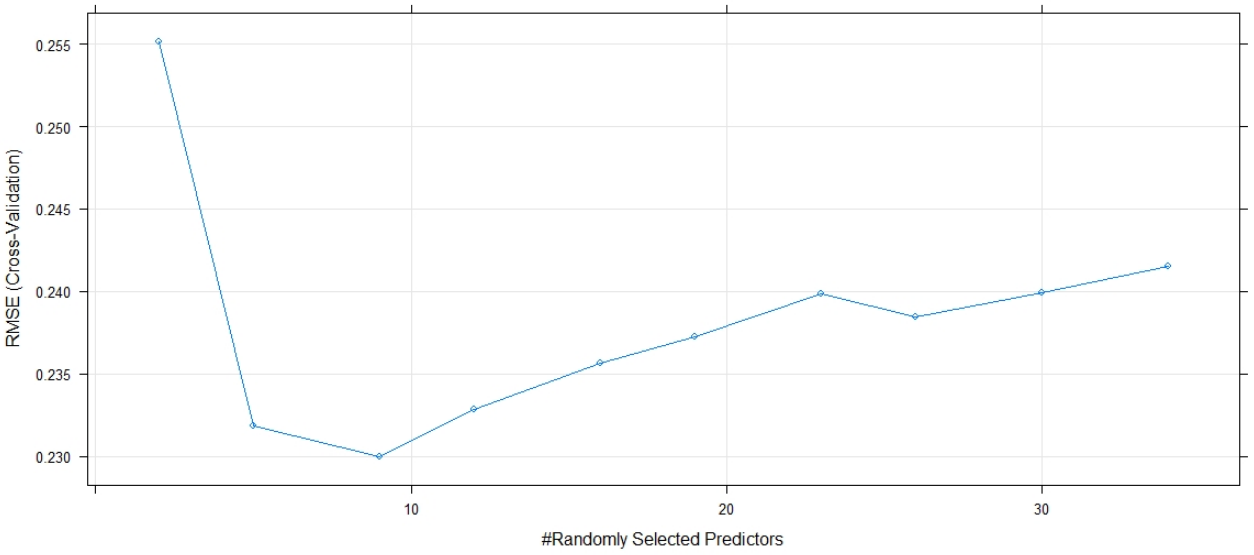


FIGURE 58 – Nombre optimal de prédicteurs du random forest de β

Notre mesure R^2 est aux alentours de 0.4 soit un score bien plus élevé que le GLM. De plus, l'écart type du R^2 en fonction des folds utilisés en cross validation est assez important avec une valeur moyenne de 0.27. Cela indique que le R^2 est volatile et donc que parfois, en fonction des folds, la part de variance expliquée peut fortement varier. Cependant, il faut prendre du recul sur cette mesure. Un R^2 *faible* ne signifie pas nécessairement que le modèle estime mal les données (et inversement) [40]. Un modèle avec un R^2 considéré comme faible (moins de 70 %) peut être tout à fait adapté à une problématique donnée. Par exemple, en sciences sociales, les modèles cherchant à expliquer des comportements humains ont des R^2 inférieurs à 50 %. Leurs conclusions sont tout de même significatives. Simplement, les phénomènes complexes ne peuvent pas être expliqués à 100 % par les descripteurs fournis. De plus, il est assez naturel

40. How To Interpret R-squared in Regression Analysis

de supposer que le coefficient de chargement est lui aussi complexe car il dépend de plusieurs acteurs très différents.

Le nombre de variables optimal est 9. Celui-ci est issu de la figure 58 qui nous donne en abscisse les prédicteurs sélectionnés aléatoirement. Ceux-ci sont les descripteurs de nos données qui, lors de la construction de chaque arbre par bagging, sont tirés aléatoirement dans notre jeu de données initial. Ainsi, nous n'avons pas d'information sur la variable sélectionnée mais nous connaissons le nombre de variables par arbre.

Avec une forêt de 1000 arbres, nous pouvons néanmoins considérer que chaque descripteur est sélectionné au moins une fois dans un arbre. En effet, la probabilité de ne pas tirer la variable i parmi l'ensemble des descripteurs est très faible lorsque le nombre d'arbres est grand, d'autant plus que notre base ne contient pas une importante quantité de variables. L'ordonnée représente le RMSE moyen par fold de cross validation. Cet indicateur est la valeur utilisée comme référence de précision de notre algorithme. Le *mtry* qui minimise le RMSE est donc 9. Ainsi, chaque arbre de la forêt se construit sur 9 variables tirées aléatoirement dans les descripteurs du jeu de données initial. Le RMSE reste tout de même presque constant peu importe la valeur choisie du *mtry*. Explorons plus en détail les performances de notre forêt optimale.

Fold	RMSE	Rsquared	MAE
2	0.2637	0.0180	0.1763
3	0.2216	0.7284	0.1476
7	0.2311	0.4008	0.1509
1	0.2133	0.4558	0.1525
5	0.3139	0.2348	0.1852
6	0.2032	0.2696	0.1258
10	0.2675	0.8783	0.1679
4	0.1837	0.2323	0.1188
8	0.2761	0.5218	0.1808
9	0.1260	0.6527	0.0937
Moyenne	0.2300	0.4393	0.1500

TABLE 34 – Performances de la forêt aléatoire optimale par fold dans la prédiction de β

Comme attendu, la variation des résultats est assez grande. La moyenne est le résultat des folds et est bien celle obtenue dans la sortie de notre modèle pour $mtry = 9$ (34). Il faut néanmoins retenir que le RSME (ou le MAE) par fold est en fait une erreur moyenne de prédiction calculée dix fois sur $1/10^{\text{ème}}$ des données d'apprentissage (pour chaque fold). Effectivement, pour rappel, lors de la cross validation l'algorithme apprend sur 90 % des données de la base d'apprentissage, avec *mtry* fixé. Or, le risque empirique (ou l'erreur de prédiction) est calculé sur la base de validation correspondant aux 10 % restants des données. Le fold le plus performant est le 9 avec un R^2 (calculé sur la base de validation) égal à 0.6527, un RMSE de 0.1260 et un MAE de 0.0937. A contrario, le fold ayant le moins bien performé est le 5 avec un $R^2 = 0.2348$, un $RMSE = 0.3139$ et un $MAE = 0.1852$.

En pratique, la variation ci-dessus est celle espérée. De plus, en fonction du fold choisi, tous les points ne sont pas aussi *facile* à prédire. Il se peut qu'une grande part de valeurs mal estimées se trouvent dans l'échantillon de validation du fold i ce qui conduit à une sous-performance et inversement pour les valeurs bien estimées. Ainsi, cette erreur moyenne n'est donc pas une erreur sur la base d'apprentissage en elle-même. Elle reflète plutôt l'erreur potentielle, ainsi

que sa volatilité, que nous pouvons obtenir en pratique. Elle ne donne cependant en aucun cas la précision de la base d'apprentissage. Pour se faire une idée de cette performance, il est nécessaire de prédire l'ensemble de la base d'apprentissage avec le modèle optimal. Cette forêt est simplement construite sur l'ensemble des données d'entraînement, avec 1000 arbres et 9 variables par arbre. Nous obtenons un RMSE de 0.11191232, un R^2 de 0.92457668 et un MAE de 0.07276166. Ces résultats sont très bons comparés à la performance moyenne de notre cross validation. Ceci nous laisse à penser qu'un risque d'overfitting se présente.

Testons la précision sur la base de test égale à 25 % des individus. La performance obtenue par notre modèle optimal est de 0.1772654 pour le RMSE, 0.4021121 pour le R^2 et 0.1239750 pour le MAE soit une moins bonne performance que celle obtenue sur l'ensemble de la base d'apprentissage. Analysons désormais les variables déterminantes du modèle.

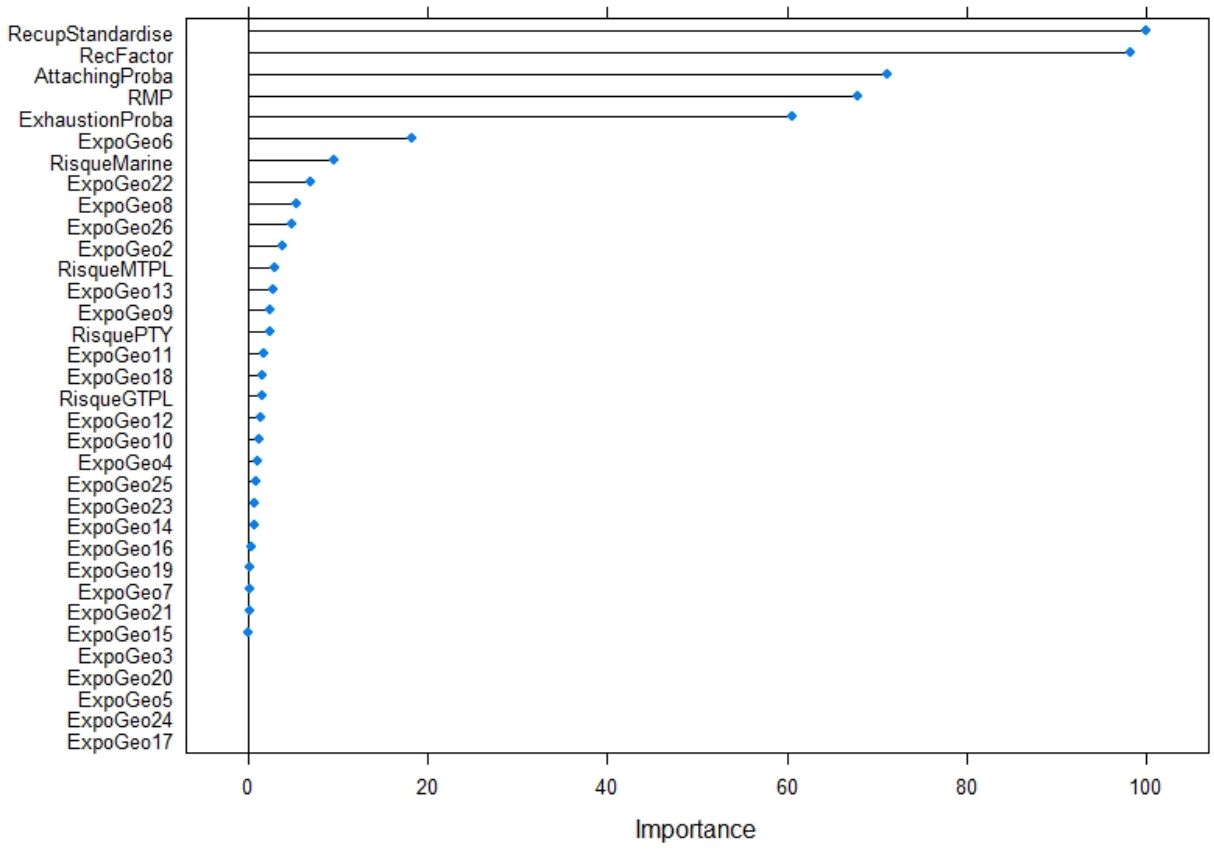


FIGURE 59 – Importance des variables du random forest de β

L'importance d'une variable représente sommairement sa part dans la précision de l'estimation du label par le modèle. Plus celui-ci se base sur une variable pour prédire, plus son importance est proche de 1. L'indicateur de la précision est le RMSE (en régression). Cette importance mesure la diminution de l'erreur (RMSE) lorsqu'une variable est utilisée dans un nœud [41]. La récupération moyenne divisée par son écart type est la variable la plus importante de la forêt suivie de près par le facteur de reconstitution. La probabilité d'attachement, le RMP et la probabilité d'épuisement sont les trois variables suivantes en termes d'importance avec environ 60 % à 75 %. L'importance des autres variables diminue ensuite assez fortement (vers les

41. Plus de détails sur la formule de l'importance ici

20 %). Enfin, certains risques (comme le Marine) ainsi que certaines expositions géographiques (comme les 22, 8, 2 et 26) montrent plus d'importance que leurs autres modalités.

Afin d'obtenir une autre évaluation de la performance générale du modèle utilisons le graphique de comparaison de l'estimation contre les données originales.

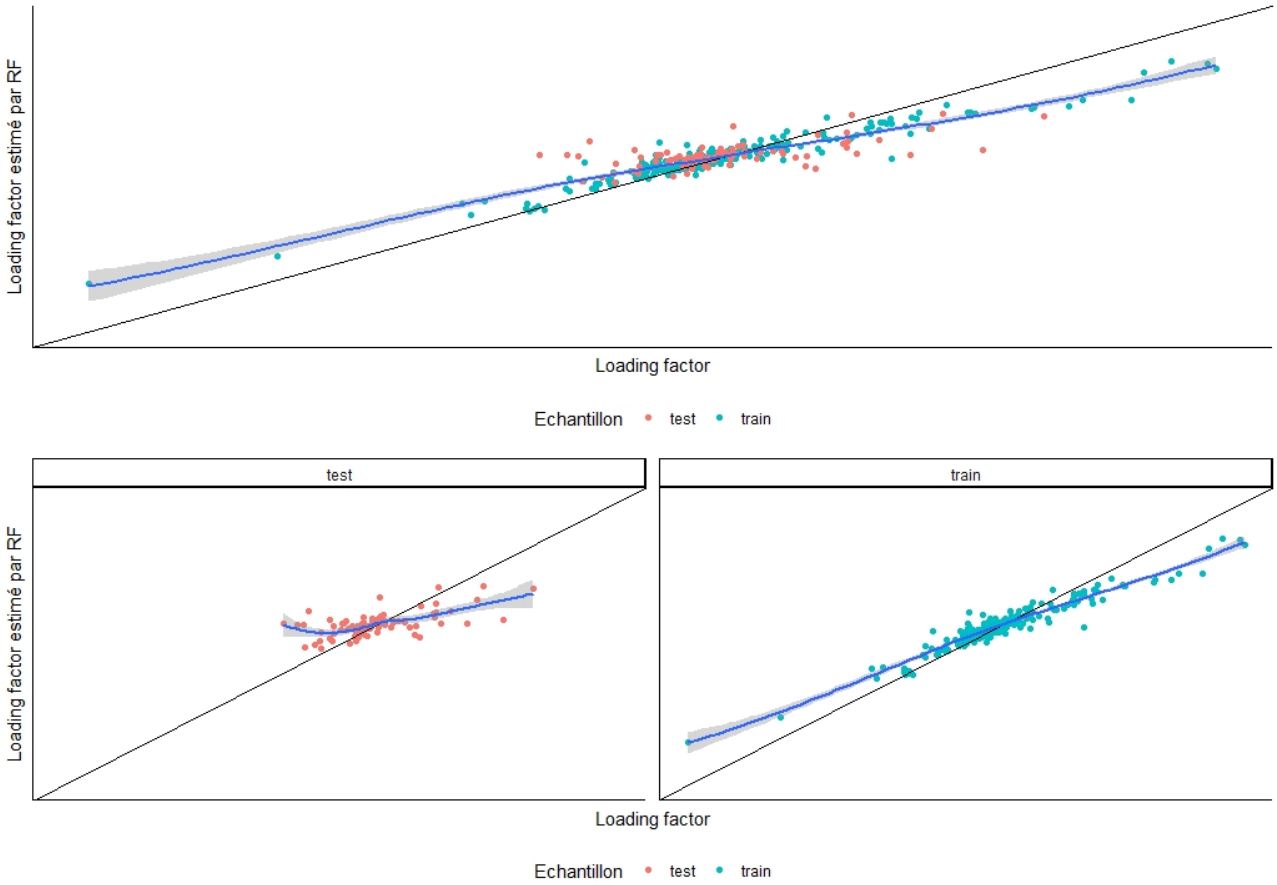


FIGURE 60 – Fitted values du random forest de β

Interprétation du graphique : Les deux graphiques représentent les données originales contre les prédictions. Celui du dessus compile la base de données entière (échantillon d'apprentissage et de test) tandis que celui du dessous affiche la performance en fonction du type de l'échantillon (test ou apprentissage).

Comparé au GLM, nous observons un bien meilleur alignement autour de la ligne médiane $y = x$. Le modèle semble avoir deux caractéristiques principales. Les grands coefficients de chargement sont tout d'abord sous-estimés par le random forest. En effet, nous remarquons un penchant vers la droite des estimations lorsque β augmente. La droite bleue (droite de régression non paramétrique) nous montre bien cette tendance, elle penche aussi vers la droite pour les β élevés. La zone grise est un intervalle de confiance à 95 % de la régression. Les chargements médians, c'est-à-dire ceux standards, sont plutôt bien réunis autour de la droite centrale. Ils sont tout de même légèrement au-dessus de celle-ci soit une surestimation de notre algorithme d'environ 4 % à 7 %. Ceux éloignés sont plus impactés, avec des résidus pouvant aller jusqu'à 25 % mais ils sont assez rares. De plus, la base de test ne contient pas de valeurs aussi extrêmes que celles d'apprentissage dans la partie négative.

La deuxième caractéristique est donc une légère (parfois forte) surestimation des β les plus faibles. Les valeurs les plus faibles correspondent naturellement aux chargements négatifs. Ainsi, surestimer un $\beta < 0$, par exemple prédire -30 % au lieu de -45 %, crée un phénomène de sur-tarification comparé à la vision des réassureurs. Cependant, une différence de 15 % (en pratique plutôt de l'ordre de 10 % à 2 %) peut avoir un impact tout à fait relatif en termes de différence réelle entre prime estimée et prime cotée. En effet, le β estimé charge l'écart type. Or, si celui-ci est faible, un chargement surestimé peut finalement n'avoir qu'un impact modéré sur la prime commerciale estimée. Il faut donc relativiser les valeurs de nos mesures d'erreurs qui ne sont pas des indicateurs sur la tarification en elle-même. Afin d'évaluer la qualité de l'estimation du tarif commercial moyen proposé par les réassureurs, nous estimons la cotation moyenne à l'aide des prédictions de β issues de notre forêt aléatoire.

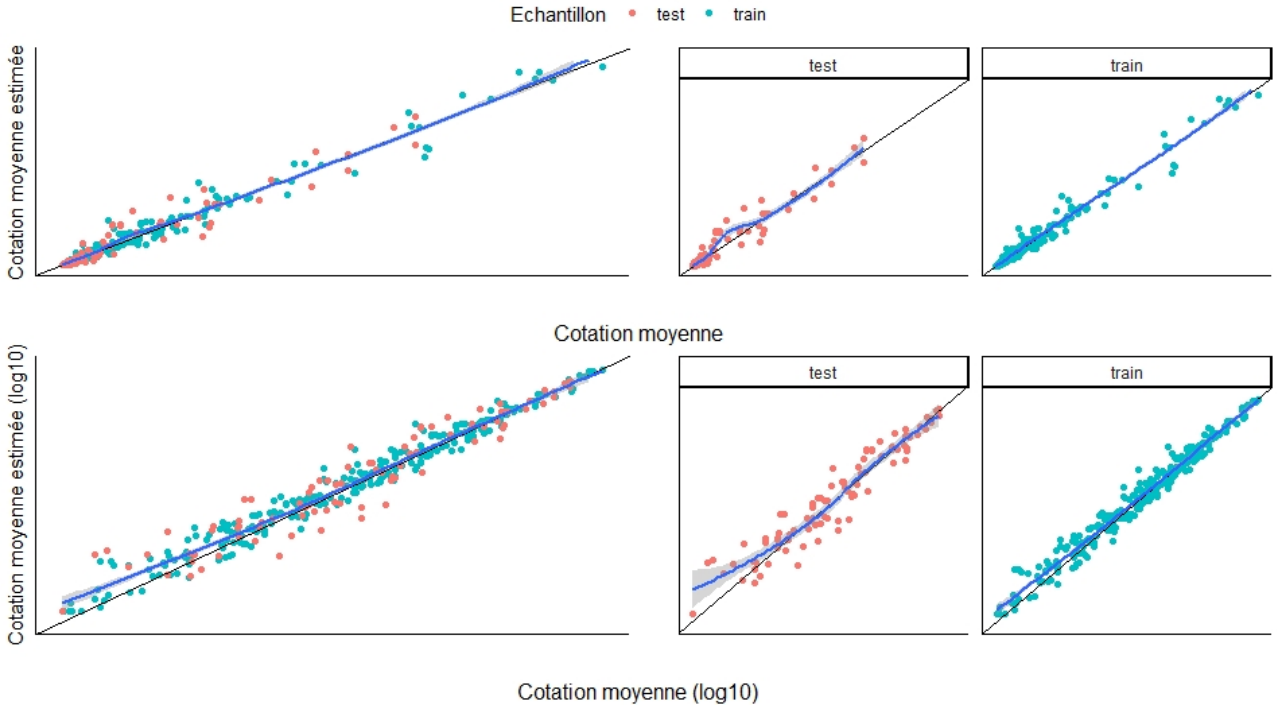


FIGURE 61 – Cotations estimées à partir du random forest de β

Interprétation du graphique : Les deux graphiques représentent les cotations moyennes originales contre leur prédiction. Celui du dessous utilise une échelle en log10 pour les deux axes afin de mettre en valeur les primes les plus faibles. Les graphiques de gauche compilent toutes les données tandis que ceux de droite discriminent par échantillon. La droite bleue est une droite de régression non paramétrique.

Cette figure nous permet d'observer notre modèle final de tarification. Globalement, il y a un bon alignement des prédictions sur la droite médiane. Une tendance à l'augmentation de l'erreur de prédiction semble se révéler lorsque les primes commerciales cotées croissent. Cependant, lorsque nous passons en échelle log10 cette tendance s'efface pour laisser apparaître une constance dans les prédictions. En effet, ce changement d'échelle a un réel intérêt ici étant donnée la forte volatilité des valeurs de primes (de dizaines de milliers à plusieurs centaines de millions). Ainsi, une échelle logarithme nous permet de visualiser la globalité de l'estimation de nos cotations, peu importe leur valeur. Comme attendu, la base d'apprentissage performe mieux que celle de test avec une très bonne estimation des primes. Concernant l'estimation sur la base test, celle-ci se relève légèrement moins correcte mais reste tout de même très convain-

cante. Ainsi, malgré un MAE de 12 % sur l'échantillon de test, celui-ci ne semble pas être si impactant sur notre tarification.

En effet, β étant finalement un coefficient de chargement, son ordre de grandeur semble se révéler tout aussi pertinent que sa valeur réelle. Les prédictions sur la base de test sont néanmoins moins centrées autour de notre ligne médiane. Les petites primes sont celles les moins bien estimées sur notre échelle log. Malgré une performance moins forte, ce qui est en général le résultat attendu, notre modèle reste tout à fait exploitable. Finalement, le plus important est d'estimer un bon ordre de grandeur de la cotation car il est de toute façon impossible d'estimer à la perfection des cotations issues d'acteurs très différents en termes de chiffre d'affaires, de solvabilité ou encore en termes de nombre de traités avec AGRe. Un bon indicateur de la performance de notre modèle est la distribution des résidus. Un résidu est la différence entre une valeur et son estimation. Un bon modèle possède une distribution de résidus centrée en 0. L'ordre de grandeur pouvant fortement varier suivant l'intervalle des valeurs du label, les résidus peuvent être très grands sans pour autant induire une mauvaise estimation.

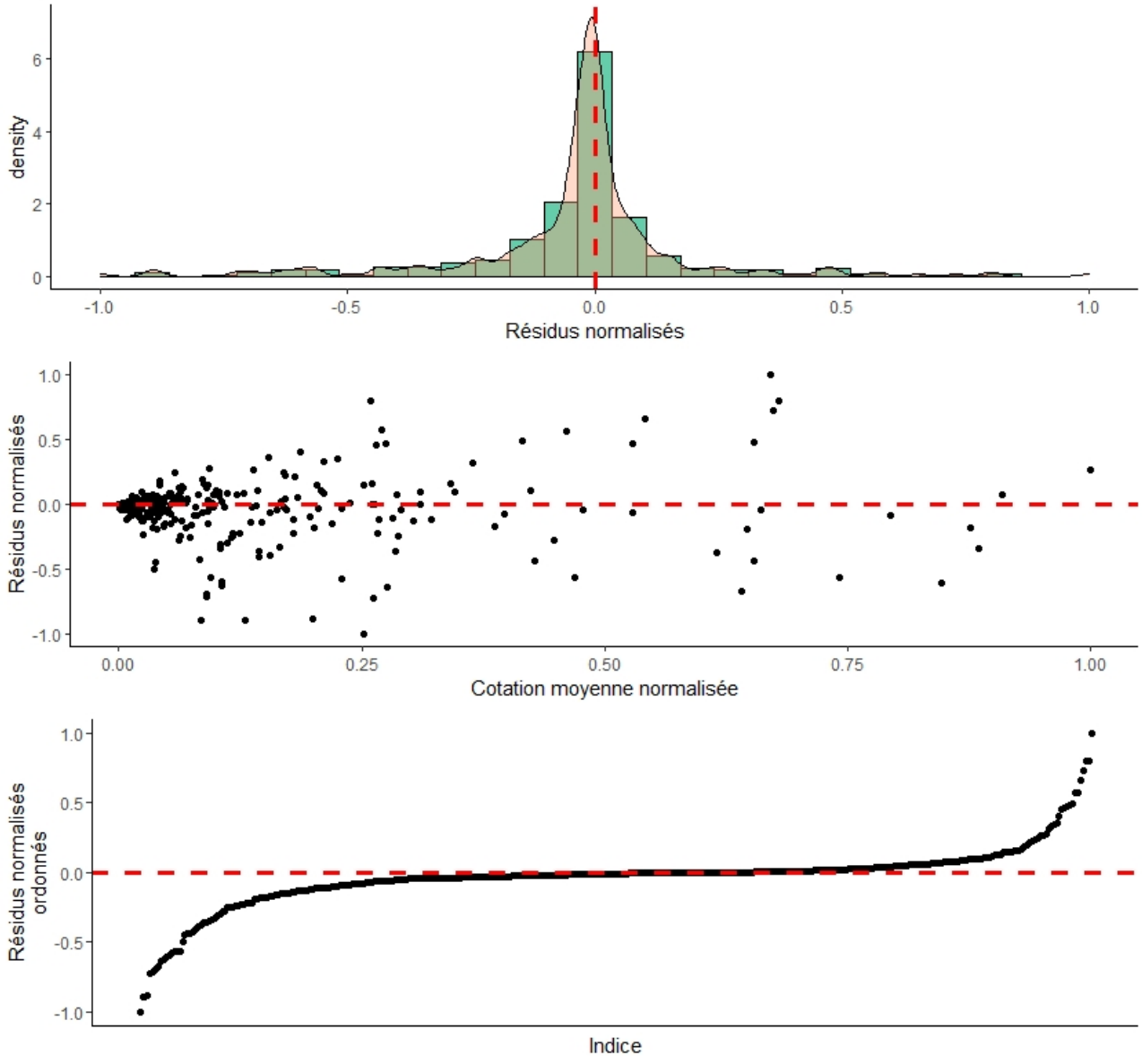


FIGURE 62 – Résidus de l'estimation de la prime commerciale par le random forest de β

Les valeurs affichées sur les axes sont normalisées pour des raisons de confidentialité. L'histogramme nous indique une distribution centrée autour de 0 (ligne rouge) avec une répartition homogène autour de sa moyenne. Nous pouvons observer que les résidus semblent être normalement distribués avec une variance assez faible. Si nous analysons la tendance des résidus en fonction de la cotation moyenne (graphique du milieu) il paraît clair qu'ils augmentent lorsque la cotation augmente. Ceci est cependant tout à fait attendu puisque les primes varient de dizaines de milliers d'euros à plusieurs millions. La différence entre les estimations se révèle donc plus élevée pour les grandes primes bien que cela n'indique pas en soit une mauvaise prédiction.

Enfin, le graphique du bas nous indique, comme pour l'histogramme, qu'il n'y a ni proportion différente entre les résidus positifs et négatifs, ni différence dans les ordres de grandeurs selon le signe. Sommairement, cela signifie qu'il y autant de cas où nous surestimons que de cas où nous sous-estimons la cotation moyenne. Pour valider cet algorithme, le nombre d'arbres reste encore à analyser. Ce paramètre n'est pas à optimiser par cross validation mais plutôt à minimiser dans la même idée que la méthode du coude. En effet, la performance de la forêt étant une fonction décroissante en fonction du nombre d'arbres, il existe une valeur à partir de laquelle le gain de performance est négligeable. Dans notre cas, ce nombre est aux alentours de 600. Rappelons cependant qu'il n'est pas limité et dépend uniquement du temps que l'utilisateur souhaite accorder à l'apprentissage du modèle.

Désormais, utilisons ces cotations pour analyser nos réassureurs.

4 Comparaison des réassureurs

Table des matières

L'outil de tarification développé précédemment est un moyen d'analyse qui se place théoriquement *avant* les cotations. En effet, le but premier est d'obtenir une formule de tarification pour nos traités actuels mais aussi d'estimer la cotation moyenne qui sera donnée sur le marché par les réassureurs. Ainsi, pour un nouveau traité, cet outil permet d'ores et déjà d'obtenir un ordre de grandeur de la prime moyenne. Cette information indique en fait deux sous-informations : la cotation moyenne et le coefficient de chargement β associé au traité. Comme nous l'avons observé, la valeur et le signe de β nous permettent aussi de comparer notre vision interne à celle externe. Maintenant, plaçons-nous dans le cas *post* cotations, c'est-à-dire lorsque tous les traités ont déjà été cotés pour une année.

Le but ici n'est plus de prédire une cotation mais plutôt de la comparer aux autres. En effet, un bon moyen pour une cédante d'en apprendre davantage sur sa réassurance, en plus de savoir tarifer un traité, est d'observer le comportement des réassureurs à travers leurs cotations. Il s'agit notamment de visualiser quel réassureur propose les meilleurs prix, lequel est le plus solvable ou encore celui qui cote le plus de traités. Toute cette analyse nous permettra de comparer nos partenaires de réassurance selon certains critères que nous définirons. Pour ce faire, il est nécessaire de retravailler notre jeu de données. Nous n'avons plus besoin des variables modélisées, seulement des variables contractuelles 6. Nous nous focalisons principalement sur l'année de renouvellement 2021. En effet, certains indicateurs comme le capital engagé par un réassureur ne peut être calculé que par année. Néanmoins, certaines mesures comme le score de précision de notre modèle par réassureur peut être généralisé à nos deux années de cotations. Nous l'indiquerons lorsque ce cas se présentera.

Nous utilisons la même base de données que celle introduite dans la section de description des données. Cependant celle-ci nécessite quelques ajustements et ajouts de données. La finalité est d'avoir une base où chaque individu est un réassureur. Par mesure de confidentialité envers nos partenaires, chaque réassureur sera identifié par un nombre aléatoire lui servant d'identifiant. Pour les comparer le plus justement possible, une information importante est manquante. Pour l'identifier, il faut se poser la question suivante : *un réassureur qui propose les cotations les moins chères est-il nécessairement celui avec qui nous préférons faire affaire ?*

La réponse est évidemment non pour une raison naturelle : la solvabilité. Il paraît clair qu'une cotation de 10 000 € d'un réassureur dont la probabilité de défaut annuelle avoisine les 50 % est moins qualitative qu'une cotation de 20 000 € d'un réassureur ayant une probabilité de défaut annuelle de 1 %. En effet, il est important pour la cédante de s'assurer que son partenaire répondra à ses engagements. Dans le cas contraire, la cédante peut se retrouver dans une situation très complexe en cas de défaut de son réassureur. Pour un traité XS, cela peut conduire tout simplement à une absence de réassurance si le réassureur était le seul à réassurer.

Ceci est donc un risque de crédit se matérialisant par une absence de remboursement de la dette (ici les sinistres) envers le créancier (la cédante). Aussi, afin d'ajouter cette variable à notre base nous choisissons un indicateur reflétant la situation économique d'une entreprise : la notation Standard & Poor's.

4.1 La notation Standard & Poor's

Standard & Poor's (nommé S&P) est une agence de notation américaine créée en 1888 sous le nom de McGraw Hill par James H. McGraw. À l'origine, cette entreprise était une maison d'édition très présente notamment sur le marché scolaire et universitaire. Cependant, en juillet 2015 McGraw-Hill Financial achète SNL Financial, société spécialisée dans l'information financière. En avril 2016, McGraw-Hill devient S&P Global. À ce jour, cette entreprise compte deux agences comme principaux concurrents : Moody's et Fitch Ratings. S&P propose des notations combinant lettre(s) et signe de la forme AA-. Ainsi pour chaque réassureur, nous recherchons sa note S&P la plus récente, généralement celle de 2020. Cette information est souvent disponible sur le site même du réassureur dans la section finances ou rapports annuels. La grille de notation générale se présente sous cette forme :

AAA	SÉCURITÉ OPTIMALE
AA+	
AA	QUALITÉ HAUTE OU BONNE
AA-	
A+	
A	QUALITÉ MOYENNE
A-	
BBB+	
BBB	QUALITÉ MOYENNE INFÉRIEURE
BBB-	
BB+	
BB	SPÉCULATIF
BB-	
B+	
B	HAUTEMENT SPÉCULATIF
B-	
CCC	RISQUE SUBSTANCIEL
CC	EXTREMEMENT SPÉCULATIF
C	PEUT ETRE EN DÉFAUT
D	
	DÉFAUT

FIGURE 63 – Grille de notation Standard & Poor's

Cette note évalue la solvabilité d'une entreprise. Elle peut être obtenue pour donner suite à la demande de l'entreprise ou simplement basée sur des informations publiques sans demande de l'entreprise. Lorsque le réassureur souhaite être noté, il va rémunérer l'agence de notation. Dans ce cas, l'agence ira chercher les informations nécessaires au sein de l'entreprise qui le demande afin de fournir une note plus solide. Généralement, le modèle de notation n'est pas

communiqué afin que l'agence en conserve pleinement le monopole. Cette notation est même souvent associée à une matrice de transition donnant une probabilité de passer d'une note X à Y dans l'année. Ainsi, par projection, il est possible d'estimer la probabilité de défaut sur un horizon de temps assez large.

4.2 Construction des données

Dans nos données, cinq notations sont présentes pour nos réassureurs : A-, A, A+, AA- et AA+. Nous associons chaque réassureur à sa note S&P. De plus, pour chaque traité de notre base, nous calculons plusieurs indicateurs autour des cotations :

— La cotation minimale
— La cotation maximale
— La cotation moyenne
— Le nombre de cotations

Ainsi, pour chaque traité, nous pouvons identifier si un réassureur est celui qui cote le plus ou le moins haut ainsi que son écart à la cotation moyenne (qui est la variable d'intérêt de notre tarificateur). Afin d'agréger toutes ces informations par réassureur pour créer une nouvelle base de données, les variables suivantes sont calculées par réassureur :

— RMP moyen
— Écart type du RMP
— Nombre de cotations
— Part moyenne espérée par traité (pondérée par RMP) [42]
— Nombre de cotations minimales (nombre de fois où le réassureur est le moins cher)
— Proportion des cotations minimales par rapport à toutes les cotations
— Nombre de cotations maximales (nombre de fois où le réassureur est le plus cher)
— Proportion des cotations maximales par rapport à toutes les cotations
— Distance relative moyenne à la cotation moyenne calculée comme

$$\frac{\sum |\text{cotation} - \text{cotation moyenne}|}{\sum \text{cotation moyenne}}$$

— Capacité $= \sum \text{Limite} \times \text{Part espérée}$
— Nombre de cotations du risque i, $i \in \{CAT, GTPL, Marine, MTPL, PTY\}$

L'en-tête de notre nouveau jeu de données se présente donc sous cette forme.

```
> head(reins_data,3)
   RenewalYear  Reinsurer  CreditRating  RMP_Moy       RMP_sd
1:        2021          1            A+  20 340 855   14 260 226
2:        2021          2            A+  22 337 489   29 237 915
3:        2021          3             A  22 833 982   11 183 863

   Nbr_cotations  Part_moyenne_par_RMP  Nbr_min  Pourcent_min  Nbr_max
1:       11               0.08592          1        0.11111        3
2:       4                0.46656          0        0.00000        1
3:       3                0.06496          2        0.66666        0
```

42. La part d'un traité est le pourcentage de primes et sinistres que le réassureur souhaite accepter. Nous parlons de part espérée car c'est la part que le réassureur propose lors de sa cotation et non pas nécessairement la part finale qui lui sera attribuée.

```
   Pourcent_max  Distance_cotation_moy  Capacite  Nbr_Risque_CAT
1:      0.33333                0.08300  16 256 255               0
2:      0.33333                0.12802  23 543 747               0
3:      0.00000                0.12550   5 623 261               2

   Nbr_Risque_GTPL  Nbr_Risque_Marine  Nbr_Risque_MTPL  Nbr_Risque_PTY
1:               0                  0                4               7
2:               0                  3                1               0
3:               0                  0                1               0
```

Cette nouvelle base nous permet d'observer le comportement général des réassureurs. Commençons tout d'abord par évaluer l'appétit au risque des réassureurs en analysant la part espérée moyenne.

4.3 Analyse de la part moyenne par traité

Une façon d'analyser la différence de part par réassureur est de construire un graphique représentant cette part en fonction de la capacité. Celui-ci est présenté ci-dessous.

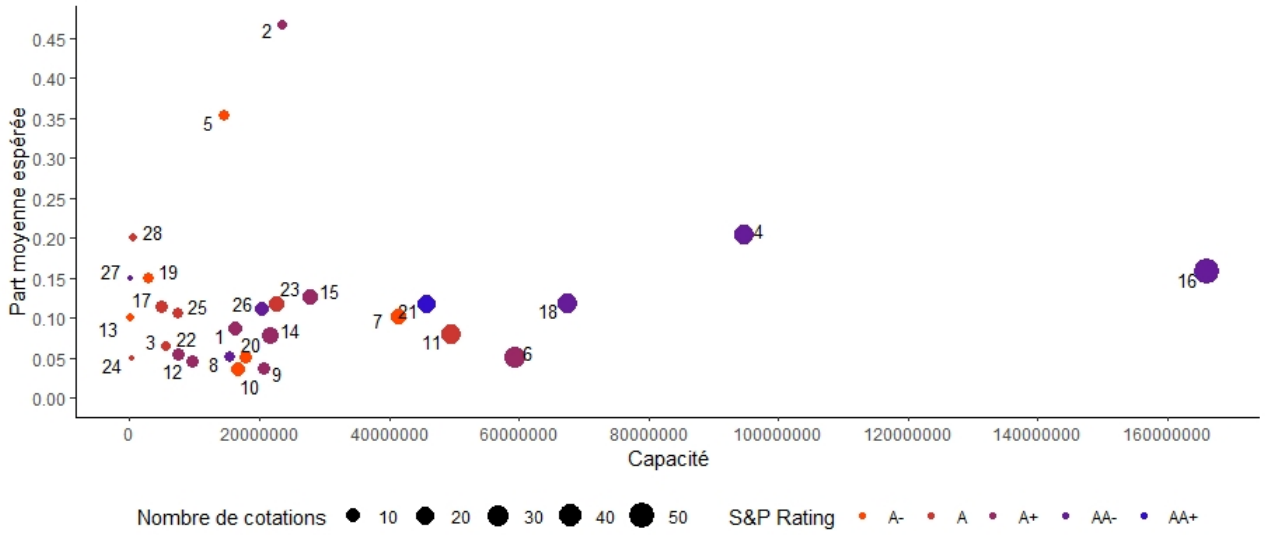


FIGURE 64 – Part espérée moyenne en fonction de la capacité

L'abscisse de ce graphique correspond à la capacité totale égale à la somme des limites multipliées par la part du réassureur. En ordonnée se trouve la part moyenne pondérée par RMP. Cette pondération permet de mieux refléter la part réelle des réassureurs dans le traité qu'ils cotent. Par ailleurs, une simple moyenne sans prise en compte des limites et priorités des traités ne permet pas de refléter avec précision la part souhaitée du réassureur. En effet, si le réassureur 1 propose 40 % sur un traité en 10 XS 10 et 10 % sur un traité 1 000 XS 1 000 nous préférons effectuer une part pondérée par RMP soit $(0.4 \times 15 + 0.1 \times 1500)/1515 = 10.3$ % au lieu de 25 % par moyenne simple. Cet indicateur nous permet en fait de comparer la tendance de la part des réassureurs en fonction des RMP qu'ils cotent. Ainsi si deux réassureurs ont un RMP moyen similaire mais une part moyenne pondérée très différente, cela indique que l'un des deux propose plus de participation sur les RMP hauts. Ceci est donc un bon indicateur pour comparer l'appétit au risque [43] entre nos partenaires. Les numéros sont les identifiants

43. Représenté ici par la part proposée.

des réassureurs. Chaque point est coloré selon la notation S&P du réassureur et sa taille est proportionnelle au nombre de cotations envoyées dans l'année.

Lorsque nous observons ce graphique, tout naturellement nous remarquons les points extrêmes. Le réassureur 16 est celui qui possède la plus grande capacité. Dans notre base, il a la *quantité de risque* la plus importante avec une capacité totale d'environ 160 000 000. Nous pouvons en effet définir la quantité de risque d'un traité comme le coût en capital engagé par les réassureurs. En pratique ce coût est égal à la limite du traité. Ainsi, pour un réassureur, le capital engagé est donc la limite multipliée par sa part qui est donc égale à la capacité. Cependant, il est important de noter que cette vision concerne uniquement notre base. En effet, un réassureur a en réalité plusieurs traités de forme équivalente. S'il possède par exemple 100 traités 10 XS 10 en automobile, sa capacité totale n'est pas de $10 \times 100 = 1000$ car il est très peu probable que tous les traités soient touchés en même temps. Pour mieux évaluer la capacité, nous devons disposer d'une information qu'est le facteur de diversification permettant de mesurer la capacité réelle engagée par un réassureur. Néanmoins, l'obtention d'une telle information pour chaque réassureur est en pratique impossible pour la cédante. Nous supposons donc ici un facteur de diversification de 1 (pas de diversification) qui nous permet tout de même de comparer tous les réassureurs entre eux sur leur capacité à un niveau égal.

Le réassureur 16 est celui qui possède le plus de capacité avec une part moyenne proposée d'environ 16 % et 51 cotations. Nous remarquons qu'en globalité, il existe naturellement une tendance entre le nombre de cotations (la taille du point) et la capacité engagée. Nous observons que plus le réassureur a de capacité, plus il cote de traités. Ce phénomène est tout à fait logique puisque plus un réassureur a de traités, plus sa capacité augmente. Par ailleurs, le nombre de cotations n'explique pas à lui seul la capacité puisque nous remarquons que les réassureurs 6, 18 et 4 ont une taille de points similaire mais une nette différence de capacité totale. Ils ont tous les trois environ 30 cotations. Ainsi, cela nous laisse à penser que pour un même nombre de cotations, le réassureur 4 cote des traités avec un plus grand RMP que les deux autres. Inversement, les réassureurs qui cotent peu possèdent aussi une capacité faible.

La part moyenne espérée est globalement comprise entre 1 % et 25 %. Seuls deux réassureurs (le 5 et le 2) ont une part plus élevée que 25 %. Leur capacité n'est pas faible comparée aux points les plus à gauche du graphique. Cela indique que ces deux partenaires sont moins enclins que leurs homologues à proposer de fortes parts sur les traités. De plus, cette moyenne étant pondérée par le RMP, cela peut tout à fait indiquer que la part proposée est majoritairement calculée sur les RMP hauts. Leur nombre de cotations se révèle être respectivement de 6 pour le réassureur 5 et de 4 pour le 2. Ils cotent donc peu mais la part qu'ils proposent est généralement forte. Si la cédante souhaite donc mettre un traité sur le marché mais qu'elle n'envisage pas de faire intervenir trop de réassureurs, il peut être intéressant pour elle de proposer à ces deux réassureurs de réassurer ce traité.

Par ailleurs, il est intéressant de remarquer que nos trois réassureurs disposant de la plus grande capacité (16, 4 et 18) font partie des mieux notés avec une note égale à AA-. Cela conforte notre idée sur le facteur de diversification. Il paraît légitime de supposer que plus un réassureur *achète* de la quantité de risque plus il a un facteur de diversification important et donc plus il est solvable puisque son potentiel de mutualisation et de diversification est fort. C'est en tout cas l'idée qui semble se dégager de notre base. Le réassureur le mieux noté en AA+ est le 21. Il dispose d'un comportement assez moyen avec une capacité autour de 40 000 000 et une part moyenne d'environ 13 %. Les réassureurs les moins bien notés ici (en A- et A) sont en rouge et en orange, soient les points les plus à gauche du graphique. Ainsi, nous remarquons

clairement une différence d'appétit au risque selon la note des réassureurs. Les moins bien notés sont ceux qui cotent le moins. Ils ont donc moins de capacité et leur solvabilité est plus faible. Ceci renforce une nouvelle fois notre hypothèse de facteur de diversification. Pour s'en convaincre, il est intéressant d'analyser le comportement de la capacité par note. Ci-dessous les boxplots des capacités totales par réassureur en fonction de leur notation S&P.

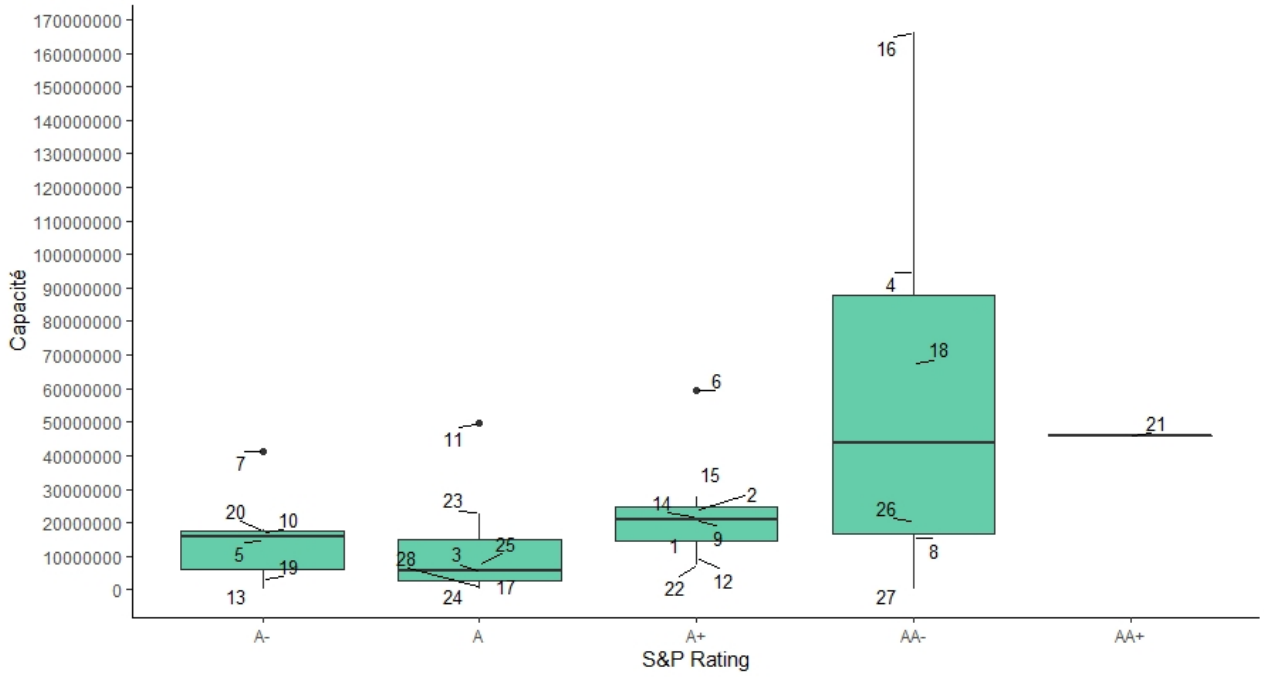


FIGURE 65 – Boxplot de la capacité en fonction du rating S&P

Ces boxplots semblent tout à fait illustrer notre phénomène de diversification et d'appétit au risque. Nous remarquons une capacité bien plus importante dans les ratings hauts (AA- et AA+) comparés aux plus faibles tandis que le nombre de réassureurs par rating est le même (de 6 à 8). Comme constaté sur le graphique précédent, le réassureur 16 est celui qui propose la plus grande capacité. Pour les réassureurs AA- et AA+, la capacité médiane est d'environ 40 000 000 tandis que pour les notes plus basses, elle se trouve autour de 10 000 000 à 20 000 000 soit 2 à 4 fois moins. Il existe cependant des réassureurs notés A qui ont une capacité plus élevée que la médiane des AA : les réassureurs 11 et 6. Ils ont donc un comportement différent de leurs homologues ayant la même note. Inversement, certains réassureurs AA- ont une capacité assez faible d'environ 10 000 000, comparés aux réassureurs de leur groupe. Ceci peut s'expliquer par le nombre de cotations envoyées par chaque réassureur. En effet, plus un réassureur a de traités, plus sa capacité est grande. Ainsi, une autre indication intéressante nous permettant de mesurer l'appétit au risque par rating est la capacité moyenne.

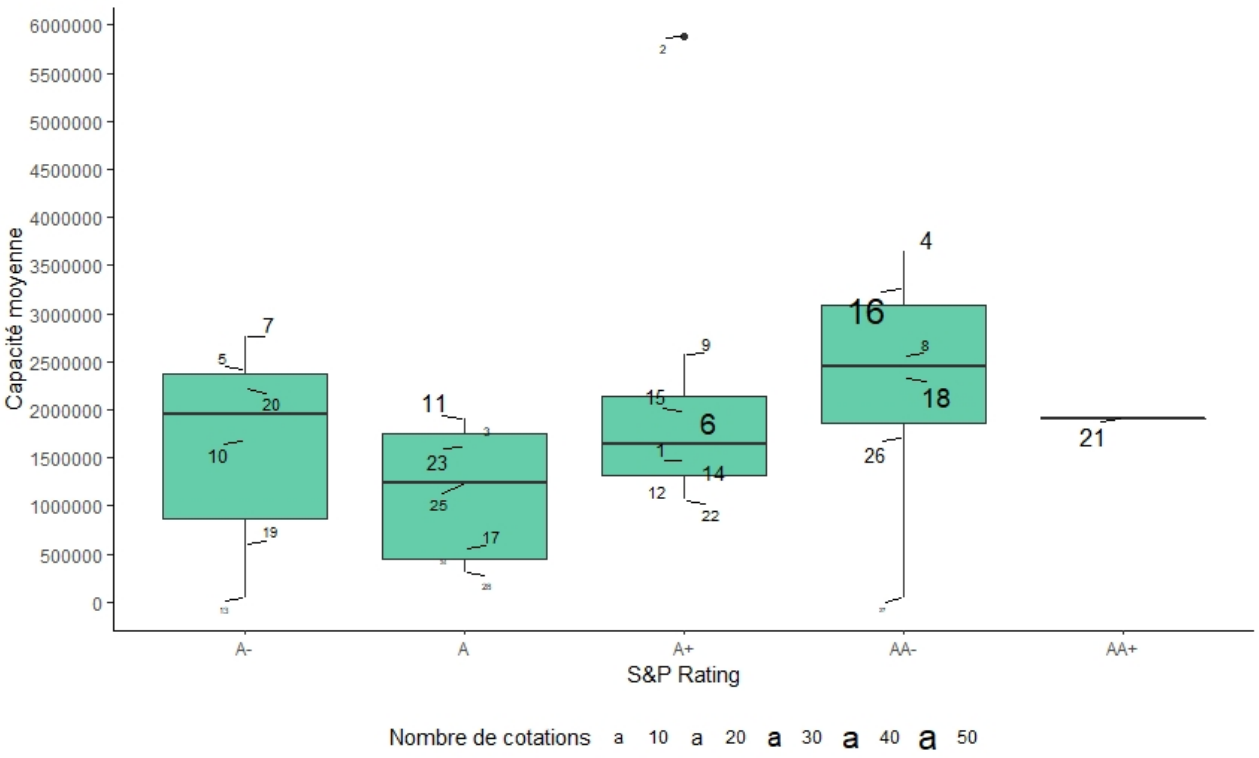


FIGURE 66 – Boxplot de la capacité moyenne en fonction du rating S&P

Cette nouvelle mesure, qu'est la capacité moyenne, change la disposition du boxplot précédent. L'impact est principalement visible sur les réassureurs ayant de nombreuses de cotations. Ils sont visibles par la taille de leur identifiant qui est proportionnelle à leur nombre de cotations. Certains réassureurs disposant d'un très faible nombre de cotations ne sont même plus identifiables. Parmi eux, nous retrouvons le 27 qui est le minimum du boxplot du rating AA-. Celui-ci n'a en réalité qu'une seule cotation sur un traité et il n'est donc pas très représentatif de la tendance de la capacité moyenne de son groupe. Néanmoins, la tendance croissante des boxplots observée précédemment semble toutefois se maintenir avec la capacité moyenne sauf pour le rating A-. Le changement majeur se trouve dans la distance interquartiles des boxplots, qui est réduite par rapport aux cas précédents. Le point le plus extrême de ce graphique est le réassureur 2, qui est le point tout en haut du boxplot du rating A+. Celui-ci détient aussi un faible nombre de cotations mais elles sont toutes associées à des traités ayant des limites hautes, ce qui explique la valeur de sa capacité moyenne.

Toutefois, nous ne parvenons pas à distinguer de différences significatives entre le nombre de cotations par rating. Pour chaque note, le nombre moyen de cotations (qui est en fait la taille moyenne des identifiants) est plutôt semblable. Les réassureurs 6, 16 et 18 se distinguent tout de même par leur nombre de cotations, comme sur le graphique introduit en premier lieu dans cette section. Le réassureur 16 n'est plus en tête du boxplot mais est placé ici sur le 3$^{\text{ème}}$ quartile. Ainsi, sa capacité totale est de 160 000 000 tandis que sa capacité moyenne est de 30 000 000. Il est désormais bien plus proche de son homologue, le réassureur 18, ayant une capacité totale d'environ 70 000 000 et une capacité moyenne de 18 000 000. Ceci nous indique que le réassureur 16 a une disparité importante dans les ordres de grandeurs des limites des traités qu'il cote par rapport au réassureur 18.

Cette section a donc permis d'analyser sommairement les tendances probables entre rating

S&P, capacité et part espérée par réassureur. Ceci nous a donné un moyen d'analyser en quelque sorte la quantité de risque par réassureur. Désormais, plaçons-nous dans un contexte différent, celui de l'analyse des prix proposés.

4.4 Analyse des cotations minimales et maximales

Cette section vise à analyser *les réassureurs proposant les meilleurs tarifs*. La notion de *meilleur* est ici intéressante à définir. Une façon naïve de définir le réassureur qui propose le meilleur prix est de le définir simplement comme celui qui cote en moyenne le plus bas le plus souvent. Or une cotation basse n'est pas nécessairement synonyme de bon prix. Nous préférons nécessairement une prime de 100 d'un réassureur AAA qu'une prime de 80 d'un réassureur BB. De plus, il est aussi possible que le meilleur réassureur (celui qui cote le plus bas le plus souvent) soit aussi surreprésenté dans les hautes cotations. Cela indique donc que ce réassureur est aussi souvent celui qui propose les cotations les plus chères, ce qui suggère un comportement extrême dans ses cotations. Il peut donc être à la fois celui qui cote le plus bas sur certains traités et le plus haut sur d'autres. Ainsi, afin de juger du meilleur tarif par réassureur nous nous basons sur le graphique suivant.

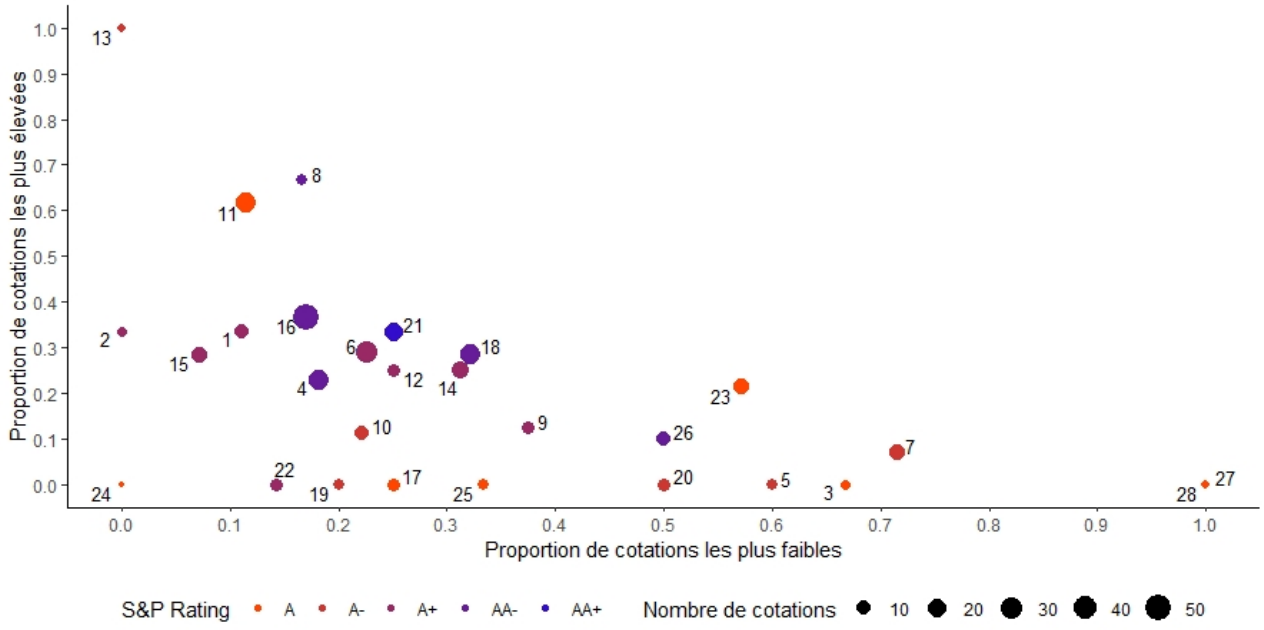


FIGURE 67 – Proportion des cotations les plus faibles et élevées

Ce graphique représente en abscisse le nombre de cas en pourcentage où le réassureur est celui qui cote la prime la plus basse (et la plus haute en ordonnée). Il est important de noter que cet indicateur n'est calculé que sur les traités où il existe au moins 2 réassureurs. Si un réassureur i est seul à coter il est à la fois le plus cher et le moins cher. Ce cas n'apporte donc pas d'information dans ce contexte. Nous observons en premier lieu des points extrêmes. Les réassureurs 27 et 28 sont toujours ceux cotant la prime la plus faible mais leur nombre de cotations [44] est égal à 1. De plus, leur notation est une des plus basses avec un rating de A. Ainsi, nous sommes clairement dans l'exemple d'illustration introduit ci-avant où ce sont bien ces deux réassureurs qui proposent les meilleurs prix mais finalement ils cotent peu et sont les moins solvables. Ils ne sont donc naturellement pas les meilleurs partenaires avec qui il semble

44. Nombre de cotations où ils ne sont pas seuls à coter.

légitime de céder beaucoup de traités. Dans tous les cas, ils ne proposent pas un grand nombre de cotations ce qui nous empêche de se fier à eux pour coter plusieurs traités. Le réassureur 13 est celui qui est toujours le plus cher mais son nombre de cotations est aussi de 1. Il fait partie de ceux les plus mal notés avec un rating de A-. Ainsi, ce réassureur est toujours le plus cher (à mettre en perspective avec son nombre de cotations) et un des moins solvables. Procédons alors à l'analyse des réassureurs importants en matière de nombre de cotations.

Nos réassureurs majoritaires sur le plan du nombre de cotations sont les 18, 16, 6, 4, 11 et 21. Ils ont tous plus de 25 cotations sur l'année avec un maximum de 51 pour le 16 (le plus gros point sur le graphique). Le réassureur 16 est le moins cher dans 17 % de ses cotations et le plus cher dans 36 % des cas. Le réassureur 18 quant à lui possède une proportion de cotations maximales à 28 % et minimales à 12 %. De plus, ces deux réassureurs ont la même notation S&P. Ils sont donc comparables en termes de solvabilité et de nombre de cotations. Le réassureur 18 est donc en général moins cher que le réassureur 16. Ils ont tous deux une proportion de cotations maximales proche. Ainsi, dans environ 30 % des cas ce sont eux les réassureurs qui cotent le plus haut. Alors, si nous souhaitons comparer ces deux réassureurs, il paraît vraisemblable de dire que le 18 est plus avantageux que le 16. Par ailleurs, il faut noter que ces deux réassureurs ne réassurent pas forcément les mêmes types de traités (même branche et même RMP) bien que leur capacité moyenne soit proche. Ainsi, en fonction du type de traité à coter il est probable que seulement l'un des deux propose une cotation.

Nous remarquons que les réassureurs les moins bien notés sont pour la quasi-totalité d'entre eux présents sur le bas de l'axe des ordonnées. Il ne sont pas nécessairement les moins chers mais ils sont néanmoins très peu souvent les plus chers. Le réassureur 7 est un de ceux qui semble détenir le meilleur rapport cotation maximale sur cotation minimale avec une proportion des plus faibles tarifs de 71 % et celle des plus forts de 7 %, avec un nombre de cotations de 15. Il cote donc un nombre de traités non négligeable avec un tarif souvent attractif. Cependant, sa solvabilité n'est pas aussi bonne que celle de nos autres partenaires avec un rating de A- comparé par exemple au réassureur 26 noté AA-, avec 14 cotations et un pourcentage de cotations minimales égal à 50 %. Ils sont donc tous deux comparables cependant l'un est le moins cher dans 70 % des cas et noté A- tandis que le second est le moins cher dans 50 % des cas mais est mieux noté avec un rating de AA-. Le classement du meilleur réassureur (basé sur les cotations) entre ces deux réassureurs repose donc sur un compromis entre prix et notations.

Enfin, les réassureurs notés A+ (en violet-bleu) semblent avoir pour la majorité d'entre eux une proportion de cotations les plus élevées entre 25 % et 35 % avec un nombre de cotations homogène (5 à 10 en moyenne). Toutefois, une grande disparité dans leur proportion à être les moins chers est observée. Le réassureur 2 par exemple n'est jamais celui qui cote le plus bas et est le plus cher dans un tiers des cas. Inversement, le réassureur 9 est lui le moins cher dans 37.5 % des cas et celui qui cote le plus haut dans 12.5 % des cas avec le double de cotations du réassureur 2 (8 contre 4). Le réassureur 9 semble donc être plus intéressant que le 2 si nous nous fions uniquement aux données ci-dessus. Il reste cependant à comparer les différents types de traités cotés par ces deux réassureurs.

Une tendance générale semble tout de même se dégager, les réassureurs les mieux notés ne sont pas les moins chers en général mais ils proposent plus de cotations. Inversement, les réassureurs les moins solvables sont souvent ceux qui cotent le moins cher mais ils possèdent un nombre de cotations plus faible. Il existe des cas particuliers comme celui du réassureur 11 qui a 26 cotations, une proportion de cotation minimale de 11 % et maximale de 61 %. De plus, il a la note la plus basse qui est A. Ce réassureur semble être donc un des moins avantageux pour

la cédante puisqu'il est rarement celui qui propose le meilleur tarif et souvent celui qui cote le plus haut avec une solvabilité moindre comparée à nos autres partenaires de réassurance. Il montre tout de même l'avantage de participer à la cotation de nombreux traités. Analysons désormais la potentielle performance de notre outil de tarification par réassureur.

4.5 Analyse des cotations moyennes

La cotation moyenne étant le label prédit par nos modèles, il paraît naturel de regarder l'écart général des cotations des réassureurs avec cette cotation moyenne. Ceci peut d'ores et déjà nous indiquer quel réassureur sera le mieux estimé par notre modèle. De plus, cette analyse peut aussi nous montrer quel réassureur est le plus loin de la moyenne. Ci-dessous le graphique illustrant cette distance à la cotation moyenne.

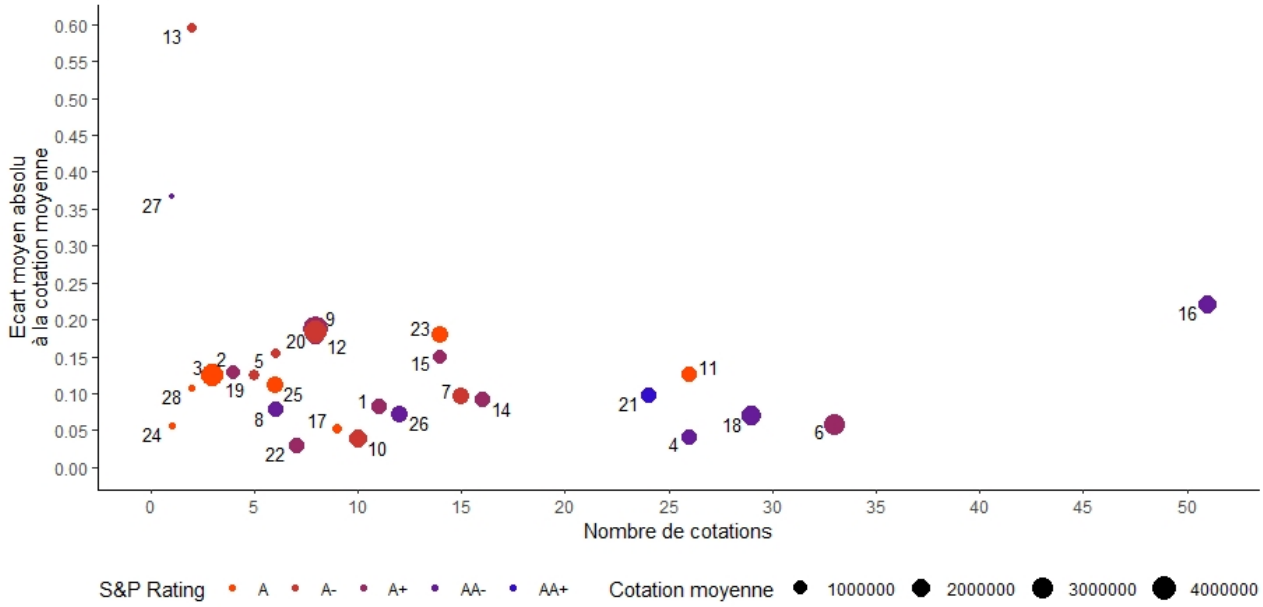


FIGURE 68 – Écart des cotations à la cotation moyenne

Les réassureurs 16 et 13 semblent une nouvelle fois se distinguer. Le 16 se démarque par son grand nombre de cotations et une distance moyenne de 22 %. Le 13 quant à lui montre un très fort écart de 60 % à la cotation moyenne pour deux cotations. Notons que nos analyses précédentes ont aussi montré que ce réassureur est celui qui a coté le plus haut pour ses deux cotations. Il est donc attendu qu'il soit loin de la cotation moyenne s'il est le plus cher dans 100 % des cas.

Globalement, l'écart à la cotation moyenne est entre 5 % et 20 %. Les réassureurs les plus proches de l'axe des abscisses sont donc ceux qui sont potentiellement les mieux estimés dans leurs cotations par notre modèle. Ainsi, lorsque le réassureur 22, 10, 4 ou 24 est sollicité pour une cotation, nous nous attendons à ce que nos modèles estiment correctement leur prime commerciale. A contrario, les réassureurs 9, 20, 12 et 23 par exemple ont un écart moyen de 20 % environ. Pour ces cas-là nous supposons donc qu'en général, notre modèle prédira une cotation assez différente de la leur. À noter que pour ces derniers, leur cotation est assez élevée et donc cet écart peut être relativisé car il est davantage pondéré par les cotations élevées que celles faibles.

Ainsi, un réassureur avec une cotation moyenne et un écart à la cotation moyenne élevé n'est pas nécessairement mal estimé par le modèle. En effet, les grandes primes ont plus de poids dans ce calcul de l'écart moyen. Elles peuvent donc être surreprésentées dans cette valeur de la distance à la cotation moyenne ce qui n'implique donc pas nécessairement une mauvaise estimation pour les primes faibles. Nous nous retrouvons ici dans un cas semblable à l'estimation de β par erreur relative ou absolue. Étant donné que nous calculons un écart absolu, si l'ordre de grandeur des primes cotées pour un même réassureur n'est pas le même, la valeur de l'écart moyen sera calculée en majeure partie sur ces cotations importantes. Cet indicateur révèle tout de même une information non négligeable sur la proximité entre la cotation du réassureur et la cotation moyenne.

Ainsi, une façon différente de visualiser la distance avec la cotation moyenne sans réelle prise en compte des ordres de grandeurs des cotations est de la calculer comme suit :

$$\frac{1}{n}\sum_{i=1}^{n}\frac{|\text{cotation}_i - \text{cotation moyenne}_i|}{\text{cotation moyenne}_i}$$

Cette valeur est l'erreur relative moyenne. Si nous observons une nouvelle fois, comme sur le graphique précédent, notre écart en remplaçant son calcul par notre formule ci-dessus le résultat est le suivant :

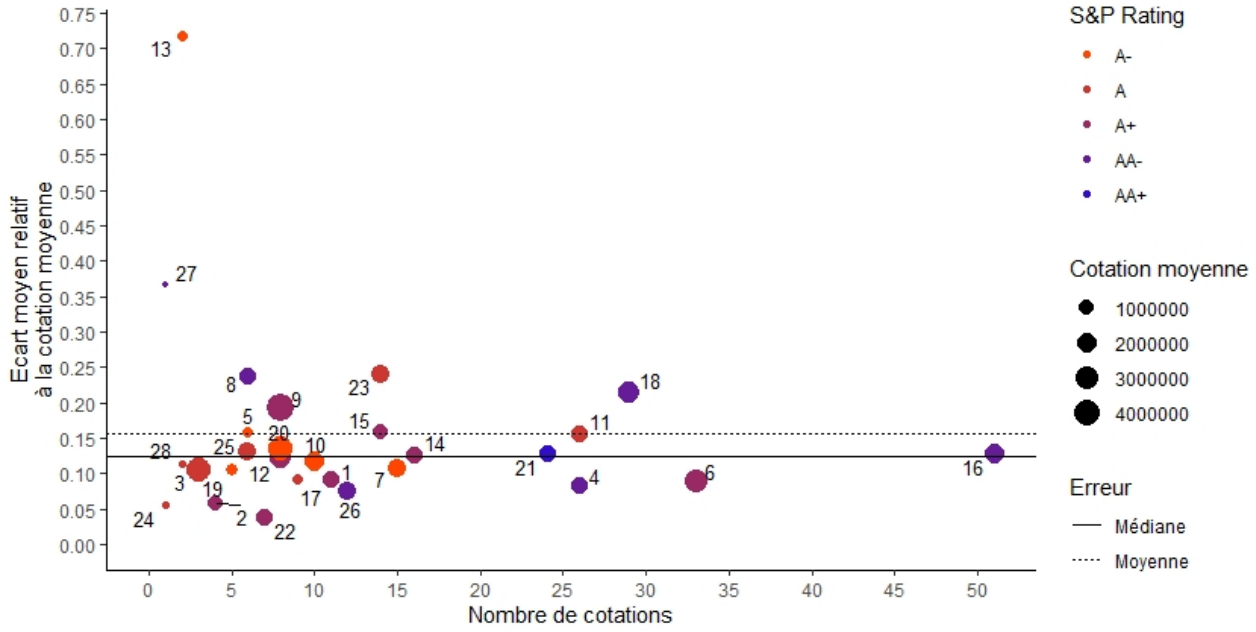


FIGURE 69 – Écarts des cotations à la cotation moyenne

Ce graphique représente donc l'écart moyen à la cotation moyenne avec notre nouvelle formule. À cela, nous ajoutons l'écart moyen (ligne en pointillés) de 16.4 % et l'écart médian (ligne en trait plein) de 12.3 %. Nous remarquons un léger changement par rapport à la formule précédente. Ainsi, avec cette nouvelle mesure nous pouvons espérer que notre modèle calculera une cotation plus ou moins différente de 16 % en moyenne des cotations reçues des réassureurs [45]. Concernant les différences par réassureur sur ce nouveau graphique, le réassureur 13 passe d'un écart de 60 % à 72 %. Ceci indique que son écart est plus élevé sur les cotations faibles que sur

45. En supposant une prédiction *parfaite* de la cotation moyenne de notre modèle, ce qui n'est pas nécessairement le cas en pratique.

celles élevées étant donné que notre ancienne formule pondérait l'écart par la cotation moyenne. Ce réassureur est donc celui le moins bien estimé par notre modèle. A contrario, certains réassureurs comme le 12 et 20 voient leur écart diminuer d'environ 5 %. Cela nous permet de constater que ces réassureurs sont donc plus éloignés de la cotation moyenne lorsque celle-ci est importante comparée à leurs autres cotations.

Les réassureurs disposant des plus grandes cotations moyennes (les plus gros points) ne sont pas tous aussi éloignés de la cotation moyenne du traité qu'ils cotent. Plus clairement, si un réassureur propose en moyenne des tarifs élevés, il ne semble pas que son écart à la cotation moyenne soit explicable par cet indicateur. Par exemple, les réassureurs 9, 20, 12, 2, 6 et 16 ont une cotation moyenne assez semblable (même taille de point) mais leur écart est approximativement identique avec d'autres réassureurs proposant des tarifs plus faibles en moyenne. Il semble ainsi complexe d'évaluer une tendance de l'écart à la cotation moyenne de nos réassureurs avec uniquement cette information. Pour affiner notre analyse, il convient de regarder aussi la différence traité par traité.

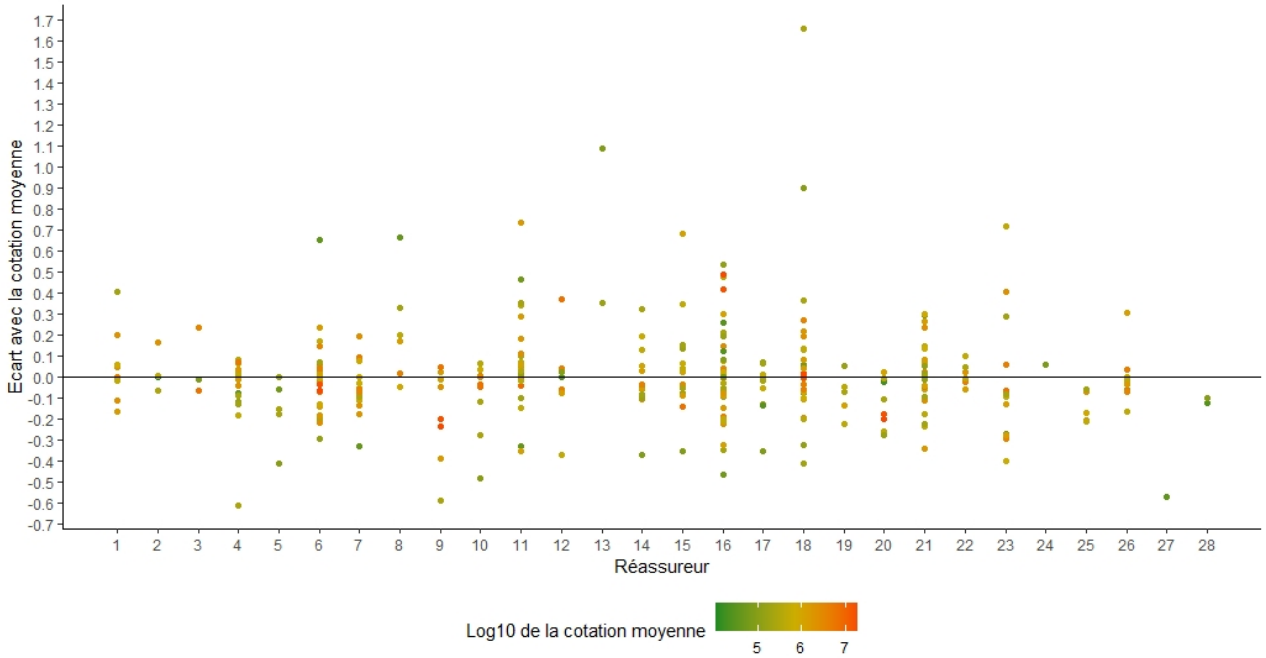


FIGURE 70 – Écart ligne à ligne des cotations à la cotation moyenne

Sur ce graphique, chaque point représente une cotation pour un traité XS en particulier. En abscisse, nous retrouvons chaque réassureur tandis qu'en ordonnée est affiché l'écart relatif de la cotation à la cotation moyenne du traité coté [46]. La couleur du point dépend de l'ordre de grandeur de la cotation moyenne du traité coté. Cette échelle est logarithmique afin de mieux différencier les très faibles cotations de celles plus *standard*. L'idée est donc d'analyser si un écart important est lié à l'ordre de grandeur de la cotation moyenne. Ceci se traduirait par des points rouges pour les grands écarts et verts pour ceux les plus faibles. Étant donné les résultats apportés par ce graphique, il semble qu'aucune tendance nous permettant d'affirmer un lien entre montant de la cotation moyenne et écart avec celle-ci ne semble se dégager. Par exemple, le réassureur 6 est proche (écart de -10 %) de la cotation moyenne lorsqu'elle est élevée, est éloigné ($\pm$ 20 %) lorsqu'elle est moyenne et même très éloigné lorsqu'elle est faible (-30 % et 70 %). Nous observons donc une diminution de l'écart lorsque la cotation augmente.

46. $\frac{\text{cotation}-\text{cotation moyenne}}{\text{cotation moyenne}}$

Ceci n'est pas le cas par exemple pour le réassureur 16 qui cote environ 50 % au-dessus de la cotation moyenne lorsqu'elle est importante et ± 10 % lorsqu'elle est moyenne ou faible, soit l'effet inverse du réassureur 6.

Le réassureur 13 est celui présentant l'écart moyen le plus élevé sur le graphique précédent avec environ 72.5 % pour un total de 2 cotations. Ici, ses écarts lignes à lignes nous suggèrent qu'il cote des primes basses et qu'il est tout de même assez éloigné des cotations moyennes. Ainsi, il apparaît comme un point extrême sur notre précédent graphique et sur celui-ci. Inversement, le réassureur 18 qui était un point tout à fait standard avec environ 22.5 % d'écart est ici un des plus extrêmes car il cote très au-dessus de certaines cotations moyennes, notamment lorsqu'elles sont faibles. Inversement, lorsque la cotation est élevée ou standard, il est plutôt proche de la moyenne : il est donc probable que notre modèle estime tout particulièrement bien les cotations de ce réassureur lorsqu'elles sont élevées. Cela nous informe donc que les écarts moyens globaux ne sont pas suffisants pour généraliser le comportement du réassureur vis-à-vis de la cotation moyenne. Il est donc judicieux de les étudier de paire avec les écarts relatifs lignes à lignes. Cela nous permet d'observer quels réassureurs sont les mieux modélisés par notre outil de tarification. Ici, ce sont donc ceux qui possèdent les points les plus centrés autour de zéro. En calculant l'erreur médiane et moyenne, avec prise en compte de la volatilité des erreurs, nous pouvons créer un classement. Le choix de l'indicateur de la précision est subjectif, il peut être par exemple égal à l'écart moyen ou médian. Un choix intéressant peut être de considérer l'erreur moyenne des points compris dans un intervalle interquartile. Nous nommons cette mesure la précision espérée PE.

4.5.1 Score de fiabilité PE du modèle de tarification par réassureur

Comme indiqué en début de section, il est tout à fait possible de généraliser PE sur nos deux années de cotations afin de prendre en compte la volatilité annuelle des cotations de nos réassureurs. Aussi, afin d'observer les différences annuelles, ce score peut tout à fait être calculé année par année. Ici nous choisissons cependant de le calculer avec toutes les cotations de notre base puisque notre modèle est entraîné sur ces deux années. Nous posons :

- $e = e_1, \dots, e_n$ le vecteur des erreurs relatives calculées pour le réassureur R avec $e_i = |\text{cotation}_i - \text{cotation moyenne}_i|/\text{cotation moyenne}_i$.
- l'intervalle interquartile $IQR = [e_{(0.1)}, e_{(0.9)}]$ [47] avec $e_{(0.1)}$ (respectivement $e_{(0.9)}$) le quantile à 10 % (respectivement à 90 %) de e.
- $m = \sum_{i=1}^{n} \mathbb{1}\{e_{(0.1)} \leq e_i \leq e_{(0.9)}\}$ le nombre de valeurs de e comprises dans IQR.

Alors, la précision espérée du réassureur R se calcule comme suit :

$$PE = \frac{1}{m} \sum_{i=1}^{n} e_i \times \mathbb{1}\{e_i \in IQR\}$$

Cette mesure nous permet d'obtenir l'écart moyen *général* (dans 80 % des cas) que le réassureur R a avec la cotation moyenne. Poser la borne supérieur de IQR égale à $e_{(0.9)}$ nous permet donc de supprimer les cas rares (potentiellement extrêmes) pouvant faire dévier notre mesure moyenne espérée. Cependant, cela améliore nécessairement fortement l'écart moyen de nos réassureurs de façon *fictive*, le rendant potentiellement bien meilleur que celui observé en pratique. Afin de compenser cet effet trop *optimiste*, nous supprimons 10 % des meilleurs écarts

47. La valeur de $e_{(x)}$ est déterminée par la fonction de répartition empirique. Dans notre cas, celle-ci est une fonction constante par morceaux. Ainsi, si $e_{(x)}$ n'existe pas en tant que vraie valeur dans e, la statistique d'ordre la plus proche de $e_{(x)}$ dans e est choisie.

(souvent proches de zéro) en posant la borne inférieure de IQR égale à $e_{(0.1)}$. De plus, si un réassureur cote quelques fois très proche de la cotation moyenne et plusieurs fois assez loin de celle-ci, la valeur moyenne de son écart sera fortement impactée par les faibles valeurs, rendant ainsi sa précision moyenne *meilleure* que celle observée en pratique. Pour ces raisons, ce choix de bornes de IQR nous paraît légitime. Finalement, cet indicateur de précision PE de notre modèle est calculé pour chaque réassureur.

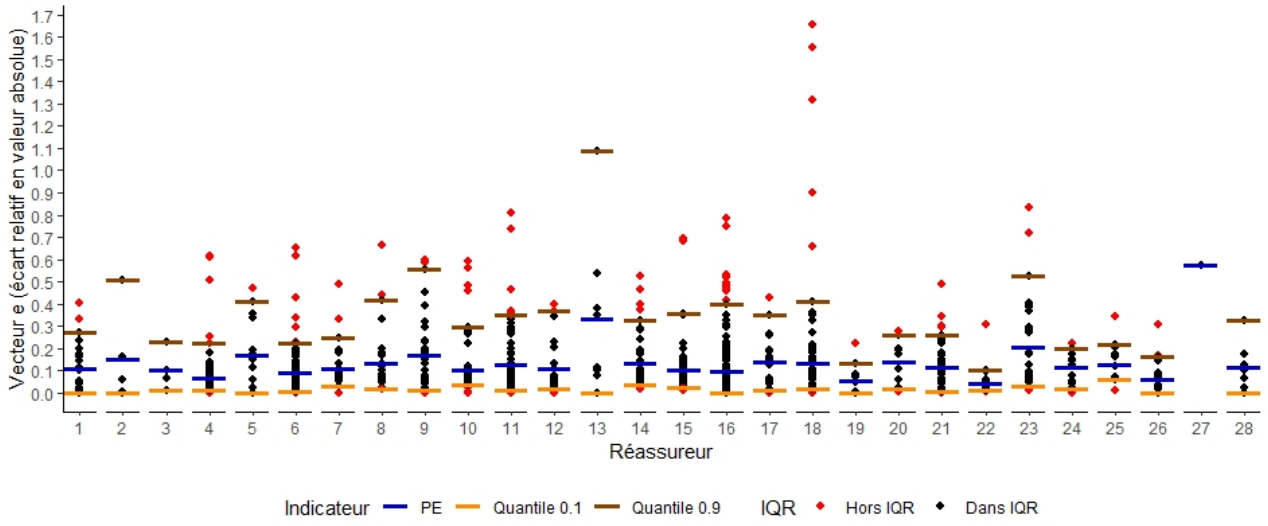


FIGURE 71 – Calcul de PE par réassureur pour 2020 et 2021

Interprétation du graphique : Les points représentent le vecteur e, pour chaque réassureur. Ils sont colorés en fonction de leur appartenance à IQR. Si $e_i \in IQR$ le point est noir tandis qu'il est rouge dans le cas inverse. Le quantile 0.1 est $e_{(0.1)}$ et $e_{(0.9)}$ est le quantile 0.9. La précision espérée PE est représentée par la ligne bleue, par réassureur.

Cet indicateur nous permet donc de calculer une précision espérée sans prise en compte de valeurs extrêmes. Le choix de IQR reste tout de même subjectif. Si l'on veut une précision plutôt pessimiste, il peut être intéressant de par exemple ne pas définir de borne supérieure afin de prendre un maximum de valeurs importantes. Pour une précision plutôt optimiste, l'absence de borne inférieure dans IQR est une façon d'améliorer la précision.

Nous calculons donc un *score de fiabilité* de notre modèle par réassureur représenté par PE. Cela nous permet ainsi de créer un classement de nos réassureurs. Nous observons que le réassureur 22 arrive en tête avec un score $PE = 0.039$. Alors, lors de nos prochaines tarifications, si le réassureur 22 intervient dans les cotations, il est fortement probable que notre modèle estime une valeur assez proche de sa cotation. Le second de ce classement est le réassureur 19 avec $PE = 0.051$ soit 0.012 de plus que le réassureur 22. Nous remarquons que pour ces deux partenaires, un seul point est au-dessus de IQR (point rouge). Celui du réassureur 22 est plus important que celui du 19 (0.4 contre 0.3 environ). Cette observation montre bien l'effet du choix de IQR sur notre score PE. En effet, si nous avions opté pour un choix pessimiste de PE en incluant les grandes valeurs extrêmes, le réassureur 19 aurait été mieux placé que le 22. Ceci montre l'importance du choix des bornes de IQR dans la cohérence des résultats obtenus. Cet impact est notamment très marqué pour le réassureur 18 disposant d'écarts extrêmes mais un $PE = 0.130$ qui est finalement peu élevé. Il arrive en $18^{\text{ème}}$ position sur 28. Sans ce choix de borne supérieure, il serait vraisemblablement celui qui aurait le moins bon score. Le réassureur arrivant en dernière position est le 27 avec un score $PE = 0.573$. Ce réassureur ne présentant

qu'une seule cotation au total, cet indicateur n'a que très peu de valeur. Il est donc préférable d'observer celui placé juste avant, le réassureur 13. Avec un $PE = 0.331$, il dispose d'écarts à la cotation moyenne très dispersés expliquant ce mauvais score. L'ensemble du classement est disponible en annexe 35. Penchons-nous désormais sur le type de traités que les réassureurs préfèrent coter.

4.6 Analyse des types de traités

Pour analyser les types de traités cotés par nos réassureurs, nous nous basons sur deux métriques qui sont le risque et le RMP. Pour rappel, le RMP est le Relative Median Point égal à 0.5× limite + priorité. Cet indicateur marque le point milieu de la tranche et nous permet de représenter la structure du traité avec une seule valeur (au lieu de deux). Ci-dessous notre graphique d'intérêt.

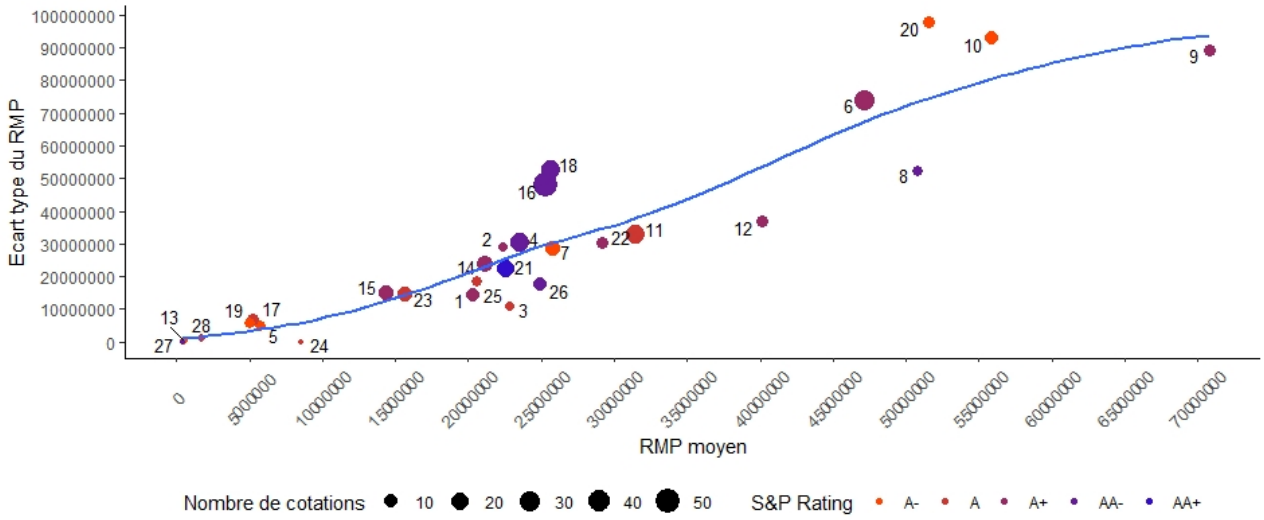


FIGURE 72 – RMP par réassureur

En ordonnée se trouve l'écart type du RMP et sa moyenne est en abscisse. Une nouvelle fois, la taille du point est proportionnelle au nombre de cotations. La droite bleue est une droite de régression non paramétrique (loess). À noter que pour les réassureurs n'ayant coté qu'un seul traité nous fixons l'écart type à 0. Nous observons une tendance de croissance linéaire du RMP moyen en fonction de sa volatilité. En effet, plus le RMP moyen augmente plus l'écart type du RMP est important, et ce, peu importe le nombre de cotations. Cela semble indiquer que lorsqu'un réassureur cote des traités dits *hauts* en moyenne c'est-à-dire avec une limite et une priorité élevés, il cote en réalité des traités assez différents. Cette différence s'accentue à mesure que le RMP moyen est grand.

L'exemple typique est celui du réassureur 9 qui possède un RMP moyen égal à 70 798 101 et un écart type de 89 285 492 pour un nombre total de cotations de 8. Ce réassureur cote donc des traités hauts mais aussi d'autres plus bas puisque l'écart type est élevé. Ainsi, il semble fortement diversifier les structures de traités qu'il cote. Nous pouvons donc le qualifier de réassureur *polyvalent* au sens où il cote des traités très différents. Il est donc un partenaire idéal pour coter des traités n'ayant à priori pas du tout les mêmes expositions au risque. Néanmoins, pour rappel, sa part moyenne par traité (pondérée par RMP) n'était que de 4 %. Ainsi, et surtout lorsque le RMP est haut, ce réassureur ne propose jamais une grande part dans les traités qu'il cote. Ainsi, la cédante ne peut pas se reposer sur ce réassureur pour réassurer de grandes parts

de ses traités et surtout lorsque ceux-ci ont des RMP hauts. Par ailleurs, ce réassureur est le moins cher dans 37.5 % de ses cotations et le plus cher dans seulement 12.5 % des cas avec un rating A+.

D'autres réassureurs lui sont assez similaires comme le 6 et 8 par exemple qui ont tous les deux des RMP moyens importants (entre 45 000 000 et 50 000 000). L'écart type du 6 équivaut dans son ordre de grandeur à celui du 9 avec une valeur de 74 032 302 tandis que celui du réassureur 8 vaut quant à lui 52 307 592. Ils cotent tous deux des traités divers. Le réassureur 8 est par ailleurs mieux noté que le 6 avec un AA- contre A+. Cependant, le réassureur 6 propose plus de cotations que ce dernier. De plus, tous deux ont une part moyenne de 5 %. Le réassureur 8 est moins souvent celui qui cote le plus bas comparé au réassureur 6 (16 % contre 22 %).

Nos deux réassureurs comptant le plus de traités (le 16 et 18) disposent eux d'un profil très similaire avec un nombre de cotations semblable ainsi qu'un RMP moyen et écart type très proches. Leur RMP moyen est quasi identique et vaut environ 25 000 000. L'écart type du RMP du réassureur 18 est légèrement supérieur à celui du 16 avec approximativement 55 000 000 contre 50 000 000. Leur rating est équivalent avec une note de AA-. Ainsi, si l'on considère uniquement les ordres de grandeur des limites et priorités des traités cotés sans prise en compte du risque, leur profil de cotation semble être tout à fait équivalent. Leur part moyenne est cependant elle légèrement différente avec 16 % pour le 16 et 12 % pour le 18. Le réassureur le mieux noté, le 21, est lui moins diversifié que les deux derniers dans ses cotations avec un RMP moyen de 22 533 035 et un écart type du RMP de 22 502 710. Sa moyenne est donc presque égale à l'écart type.

Les réassureurs 20 et 10, qui font partie des moins bien notés avec un rating de A-, sont cependant les plus hauts dans l'écart type de leur RMP qui vaut environ 97 000 000 pour le 20 et 93 000 000 pour le 10. De plus, la moyenne de leur RMP est aux alentours de 50 000 000 (respectivement 55 000 000) pour le 20 (respectivement pour le 10). Leur nombre de cotations est environ 9 pour les deux (8 et 10). Ainsi, ils présentent un profil similaire avec des traités cotés dans les mêmes ordres de grandeur. De plus, ils semblent bien plus diversifiés dans leur cotation que les autres avec une grande volatilité de RMP coté. Leur RMP moyen étant en plus important, ces réassureurs cotent donc des tranches hautes voir très hautes mais aussi plus basses. Ils sont donc tout aussi polyvalents que le réassureur 9 mais possèdent une note plus faible. Par ailleurs, sur 8 cotations, le réassureur 20 est le moins cher dans 50 % des cas tandis que son homologue, le réassureur 10, l'est dans seulement 22 % des cas pour 10 cotations. De plus, le réassureur 20 n'est jamais le plus cher tandis que le 10 l'est dans 10 % des cas. Ainsi, pour un profil de risque presque identique, le réassureur 20 semble préférable au 10 avec une part moyenne de 5 % contre 3 % pour le 10. Il faut cependant encore analyser les risques qu'ils cotent afin de s'assurer de leur similarité sur ce point.

Finalement, analysons les types de traités cotés par les réassureurs en fonction du risque.

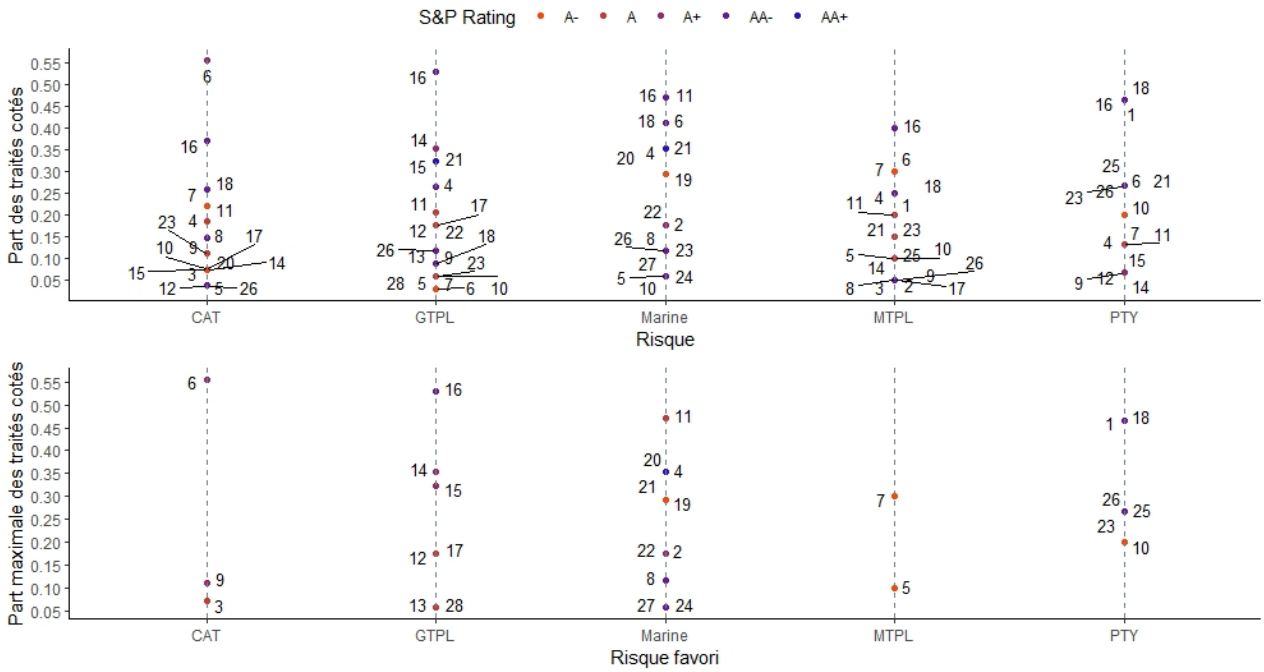


FIGURE 73 – Risques cotés par réassureur

Cette figure, composée de deux graphiques, offre une analyse complète des risques cotés par les réassureurs. De haut en bas, les graphiques représentent :

— La part des traités cotés de chaque risque. Par exemple, s'il existe n traités CAT et que le réassureur cote k traités CAT, la valeur affichée en ordonnée pour ce risque et ce réassureur est k/n.
— Le risque favori par réassureur défini comme le risque ayant la part maximale de cotations. Ceci correspond au point le plus élevé, par réassureur, sur le graphique du dessus. Ainsi, nous supposons donc qu'un réassureur préfère coter le risque i s'il cote en majorité les traités du risque i.

La façon de définir le risque favori est ici assez subjective. Nous pouvons supposer que le risque favori d'un réassureur est celui ayant le plus de cotations. Par exemple, si le réassureur cote 10 traités Marine et 5 traités CAT, son risque préféré est le Marine. Néanmoins, nous ne prenons pas en compte le nombre de traités à coter. En effet, supposons qu'en réalité il n'y ait que 5 traités à coter en CAT et 20 en Marine. Alors, ce réassureur est intervenu sur 100 % des traités CAT et 50 % des traités Marine. Dans ce cas, son risque favori, au sens de celui ayant été le plus coté en proportion, est le CAT. Cette dernière façon de définir le risque favori est celle retenue ici dans le second graphique.

Tout d'abord, nous remarquons qu'au maximum, un réassureur intervient sur près de la moitié des traités à coter. Cela se produit chez les réassureurs 6 et 16. Le 6 cote 55 % des traités CAT tandis que le 16 cote entre 35 % et 50 % environ des traités de chaque risque. Logiquement, les réassureurs ayant le plus de cotations apparaissent en haut de ce graphique. Le réassureur 16 possédant 51 cotations, il apparaît comme un des points les plus élevés pour tous les risques.

Il est intéressant de remarquer que ce réassureur ne dispose pas de parts différentes en fonction des risques mais qu'au contraire elles semblent assez stables. Ceci indique une volonté de

diversification homogène des traités proposés par AXA Global Re. En effet, lorsque n traités sont émis sur le marché par AGRe, le réassureur 16 cote environ 45 % des n traités. Ainsi, bien que son risque favori soit le GTPL coté à 52 %, il semble assez indifférent aux risques et préfère au contraire homogénéiser ses parts.

A contrario, le réassureur 6 qui est le deuxième réassureur cotant le plus de traités, ne semble pas suivre le comportement du réassureur 16. En effet, il cote 55 % des traités CAT (son risque favori) et seulement 3 % des traités GTPL. Le nombre de traités pour ces deux risques étant similaire, il semble que ce réassureur soit assez enclin à réassurer les traités AGRe en GTPL par rapport aux CAT.

Le MTPL se démarque des autres risques car ce sont uniquement deux réassureurs notés A- qui l'ont en risque favori. Le réassureur 7 (respectivement 5) cote environ 32 % (respectivement 13 %) des traités de ce risque.

Enfin, sur le graphique du dessus, la répartition des ratings par risque est hétérogène. Ainsi, aucun risque n'est coté uniquement par des réassureurs bien ou mal notés. Une exception semble tout de même se dégager sur le risque Marine où uniquement un réassureur A- est présent contre une majorité de AA- et AA+.

Conclusion

L'objectif de ce mémoire était donc, dans un premier temps, d'essayer de créer un lien entre les cotations envoyées par les réassureurs et notre vision interne de la réassurance. Pour se faire, nous nous sommes basés sur une formule de tarification par écart type où le but principal était d'estimer le coefficient de chargement de l'écart type β. Pour y parvenir, plusieurs méthodes ont été proposées avec des améliorations apportées au fur et à mesure de cette section. Nous distinguons donc deux grandes méthodes : une par chargement constant par groupe (ou cluster) et une autre par un chargement ligne à ligne. Le processus de tarification dépend donc de la méthode que nous voulons utiliser. Il peut se résumer à la procédure ci-dessous.

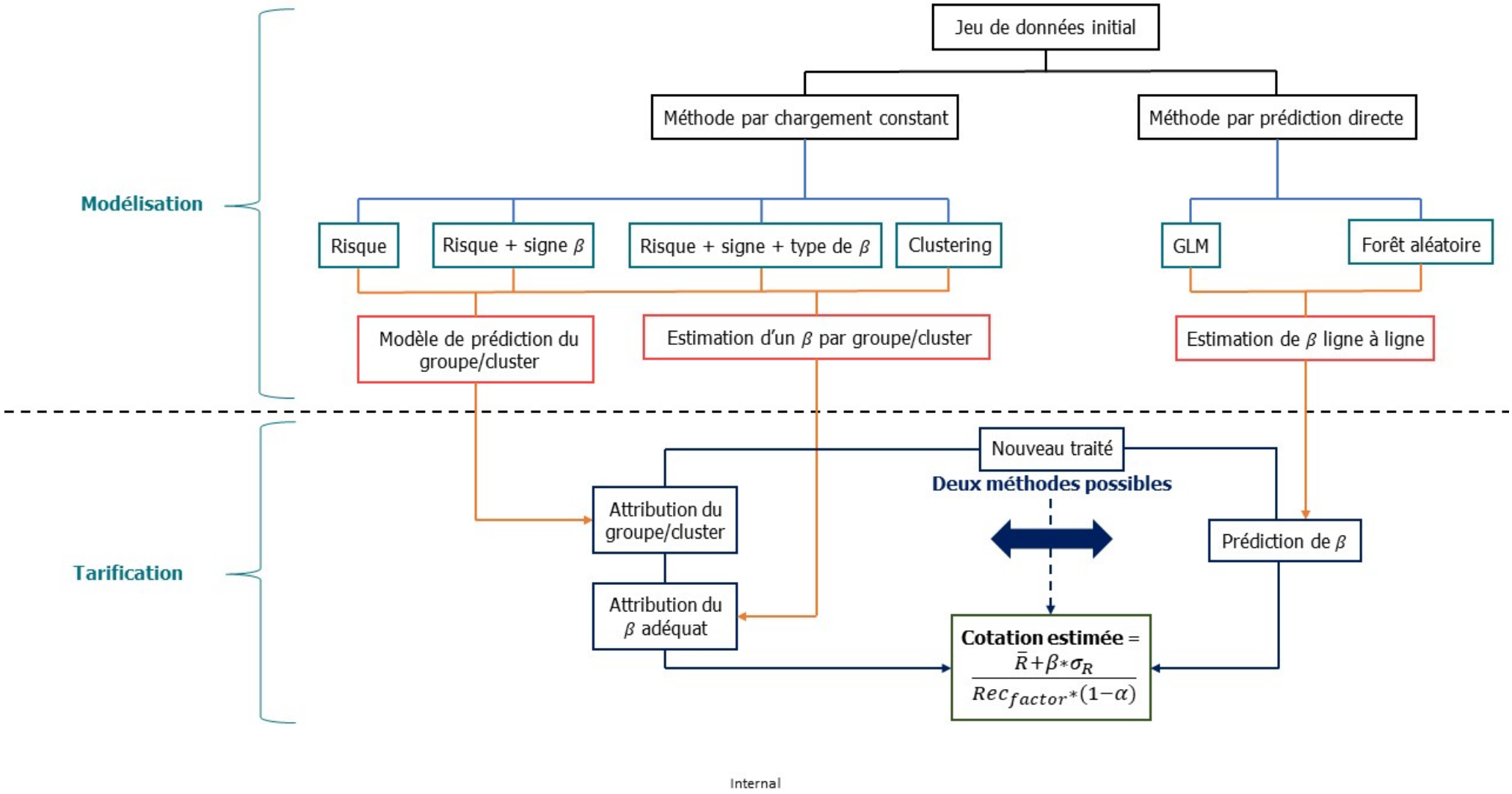


FIGURE 74 – Modélisation de la tarification par chargement sur écart type des traités XS

Nous avons donc modélisé cinq sous-modèles pour la méthode par chargement constant et deux pour la prédiction directe soit sept sous-modèles au total. Chacun offre ses avantages et ses inconvénients mais il existe un constat général simple : **il n'existe aucune solution de tarification répondant conjointement aux critères de simplicité, transparence, efficacité et rapidité**. De plus, ils dépendent tous en intégralité de la qualité des données fournies en amont. Notre base étant construite elle-même sur des modélisations d'assurance, le travail nécessaire pour aboutir à un tel jeu de données est très important. Plusieurs étapes comme la récolte de données, la modélisation des sinistres d'assurance puis la modélisation des traités de réassurance et enfin la jointure avec les cotations des réassureurs représentent une majorité du temps passé à l'élaboration de ces modèles. AXA ayant l'avantage d'être un assureur mondial implanté dans différents pays, cela nous permet de constituer une solide base de données avec un nombre important d'exemples de cotations. Ceci n'est pas le cas pour tous les assureurs (ou réassureurs en rétro-cession). En effet, une cédante disposant de peu de traités peut difficilement établir un modèle de ce type. Finalement, le choix entre les sept sous-modèles dépend de plusieurs critères dont principalement deux à savoir la performance et la transparence.

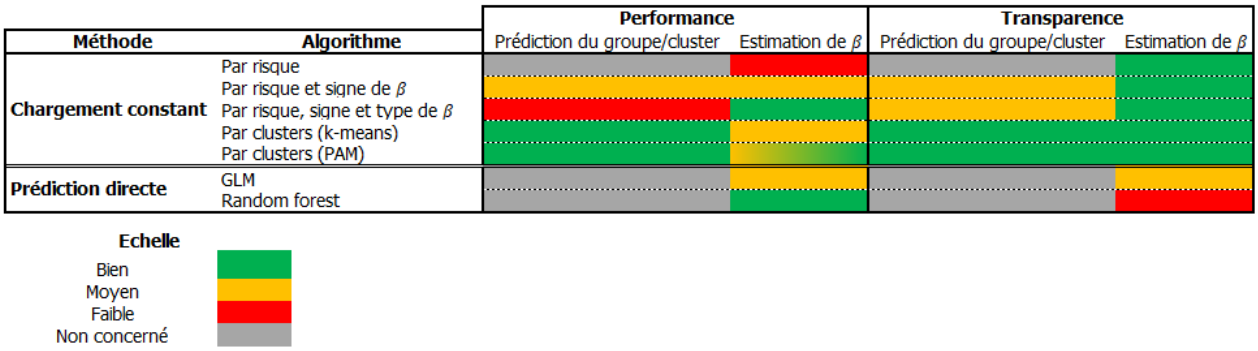


FIGURE 75 – Comparaison des différents modèles de tarification

Analysons chaque modèle un à un. Le modèle par chargement constant segmenté par risque est le plus *naïf* des modèles utilisés. Il ne prédit pas un bon coefficient de chargement à cause de la forte disparité des β. Il est donc très simple mais sous-performant. Cet inconvénient lui permet néanmoins une grande transparence puisque la formule de tarification se trouve être la même par risque. De plus, il est très simple de comprendre le choix de β puisqu'il est simplement issu de quatre courbes convexes dont le minimum est le coefficient de chargement choisi. Il informe néanmoins la cédante sur les différences de vision entre les réassureurs et les modélisations internes des traités.

Le modèle par chargement constant segmenté par risque et par signe de β est quant à lui plus performant que le dernier sur la prédiction de β. Les risques assez peu volatiles en terme de chargement comme le risque 1 sont très adaptés à cette approche. Nous savons donc déjà à l'avance quel risque peut bénéficier d'une formule de tarification constante, ce qui est une information non négligeable. L'inconvénient majeur est qu'un nouveau traité à tarifer doit en amont être attribué au bon groupe. Il est nécessaire de prédire le signe de β (pour rappel, groupe 1 pour $\beta < 0$ et groupe 2 pour $\beta \geq 0$). Or, notre modèle de prédiction par CART binaire ne présente pas une excellente performance. Il est possible d'utiliser un modèle de prédiction du groupe plus complexe mais cela nuirait à l'avantage prépondérant de la méthode par chargement constant qui est sa simplicité. De plus, le CART possède l'avantage considérable de proposer une prédiction visuelle par arbre ce qui, d'un point de vue opérationnel, peut se

révéler très utile. Néanmoins si l'on souhaite tout de même adopter cette approche, il peut être intéressant d'utiliser des algorithmes plus performants que le CART mais la transparence serait entachée. Ce modèle de prédiction a tout de même l'avantage d'offrir une vue des seuils choisis par l'arbre, indiquant quelles variables potentielles sont les plus déterminantes dans la différence de vision (mesurée par le signe de β) entre cédante et réassureurs. Certains risques n'étant pas assez stables dans la valeur de β, une autre modélisation atypique est entreprise.

Celle-ci est le modèle par chargement constant segmenté par risque et type de β. Nous modélisons un seuil atypique pour les risques en ayant besoin. Cependant, un tel seuil réduit le nombre de données par groupe. Ainsi, étant donné le faible nombre de données de β négatifs (et de surcroît les β négatifs atypiques) il n'est pas possible de créer un groupe atypique pour ces cas. Néanmoins, nous le faisons pour les β positifs des risques 2, 4 et 5. L'analyse de ceux-ci nous a permis de construire un seuil commun pour les risques 2 et 4 et un seuil unique pour le 5. Nous avons donc créé un troisième groupe. Les β ainsi obtenus sont bien plus représentatifs de la réalité des risques. Il nous permettent de tarifer à la fois les traités en vision pessimiste ($\beta < 0$), standard ($\beta \geq 0$ et attritionnel) et atypique ($\beta \geq 0$ et grand). Il existe une formule de tarification par écart type par groupe soit un total de dix formules. Malheureusement cet ajout de groupes nuit à la qualité de prédiction de notre CART qui est désormais un arbre de classification multi-classes.

Afin de résoudre en partie ce problème d'attribution du bon groupe, nous avons modélisé des clusters sans la prise en compte de β. Nous avons fait l'hypothèse que ceux-ci constituaient des groupes homogènes de β ce qui n'est malheureusement pas toujours vérifié. Un modèle de clustering simple par k-means crée en premier lieu n'est pas concluant. La fréquence de données par clusters est assez mal répartie. Ainsi, certains groupes ont trop de données et sont donc sous-performants lorsque nous estimons leur β par notre algorithme de minimisation. Par ailleurs, la heat map permet une bonne visualisation des différences entre clusters. Ce graphique nous indique en quelque sorte la *carte d'identité* de chaque cluster créé. Cette méthode permet, pour un nouveau traité, de lui attribuer son bon groupe sans prédiction. En effet, les clusters étant construits par distance euclidienne, il est très simple de trouver le groupe d'un nouvel individu : c'est simplement celui le plus proche. Malheureusement ce calcul de distance ne permet pas la prise en compte de variables quantitatives qui se montrent être assez déterminantes dans nos calculs.

Nous avons donc choisi de changer notre façon de créer nos clusters. La distance euclidienne n'était pas adaptée à notre objectif d'ajout de variables catégorielles comme le risque ou la région du traité. Nous avons donc introduit la distance de Gower couplée à l'algorithme de clustering nommé PAM qui se montre être plus robuste et explicite que les k-means. Le temps de calcul est malheureusement plus élevé mais nécessaire car le gain d'informations n'est pas négligeable. Le nombre optimal de clusters passe de 7 à 14 soit deux plus fois plus de clusters qu'auparavant. La fréquence de données entre clusters est donc bien mieux répartie bien que certains restent majoritaires en terme de quantité d'individus. L'estimation de β se trouve donc être plus fiable car la distribution des β intra-cluster est plus stable. Il reste cependant certains clusters où le choix final du β pour tarifer n'est pas pertinent, souvent dû à un problème de volatilité nous empêchant de fixer un coefficient de chargement constant.

Finalement, l'imposition d'une formule avec un chargement constant pour un grand nombre de traités nous permet de bénéficier d'une grande simplicité et de transparence dans notre modèle. Le coût de ce choix est une potentielle mauvaise estimation de la cotation moyenne. Une solution intéressante est donc de prédire un coefficient de chargement ligne à ligne plu-

tôt qu'un chargement fixe pour un nombre k de lignes. Ceci implique cependant une formule unique de tarification par individu. Ainsi, le gain d'estimation se paie par un impact direct sur la complexité du modèle. Tout dépend donc du critère de préférence de l'utilisateur dans le choix des méthodes. Pour obtenir cette estimation, nous nous sommes basés tout d'abord sur un algorithme paramétrique bien connu : le GLM. Cette méthode nous a donné des résultats peu convaincants, l'hypothèse de normalité étant trop forte.

Nous avons donc entrepris un changement : le choix d'un algorithme non paramétrique. Le candidat idéal pour nos besoins est l'algorithme du random forest. Il possède l'avantage d'être efficace, assez peu coûteux en termes de temps de calcul comparé à certains de ses homologues et enfin d'être tout de même assez explicite, du moins dans son processus de construction. La prédiction de β obtenue par cet algorithme est très satisfaisante. La prime cotée estimée à partir de ces coefficients de chargements prédits est tout à fait proche de la réalité pour la majeure partie de nos individus (test et apprentissage). La performance de tarification est donc très convaincante mais peu explicite sur le réel calcul sous-jacent de β. Nous ne savons pas expliquer avec précision les calculs internes conduisant aux prédictions de la forêt.

Chaque méthode apporte donc ses avantages et ses inconvénients. Le choix de l'une d'entre elles dépend donc des objectifs et des besoins de l'utilisateur. Une façon de s'assurer de la pertinence de notre tarification est d'utiliser nos trois méthodes les plus performantes : celle par prédiction direct avec forêt aléatoire, celle par application de l'algorithme de minimisation avec nos trois groupes de β (négatif, standard et atypique) et enfin celle par PAM. En fonction des résultats et avec nos connaissances sur la fiabilité et cohérence des estimations du coefficient de chargement, il est possible de trancher sur la prime qui semble la plus fiable.

Dans un second temps, nous voulions analyser plus en détail nos différents réassureurs. Pour ce faire, nous avons créée une base de données compilant un maximum d'informations sur les cotations et les réassureurs. Parmi elles, nous retrouvons par exemple le rating fait par l'agence de notation Standards & Poor's (communément appelé le rating S&P), le nombre de cas où un réassureur est celui qui cote le tarif le plus faible (et le plus élevé) ainsi que son écart avec la cotation moyenne. Aussi, d'autres indicateurs comme le RMP moyen et son écart type ont été ajouté. Enfin, nous avons analysé les différences de cotations par risque. Tout cela nous a permis d'identifier des tendances comme par exemple le fait que plus un réassureur cote un RMP moyen haut, plus la volatilité du RMP coté est élevé. Ainsi, sommairement, plus un réassureur cote des traités hauts (limite et priorité élevées) plus il cote des traités différents en termes de RMP. Cela peut donc être assimilé à une forme de mutualisation et de diversification. Aussi, à travers les ratings, nous avons pu identifier certains comportement comme des tendances de nombre de cotations ou encore des capacités engagées par les réassureurs. Plus un réassureur est noté haut, plus il engage une forte capacité (définie par limite $\times$ part du réassureur dans le traité) car il a tendance à plus coter. Enfin, un score de fiabilité noté PE permettant de juger la précision de nos modèles de tarification par réassureur a été créé. Tout ceci permet donc à la cédante de caractériser, selon ses préférences, ses réassureurs.

Ce mémoire met donc en perspective deux façons d'analyser la réassurance : une pré-cotation et une autre post-cotation. Celle en pré-cotation consiste à modéliser des méthodes de tarification des traités XS par principe de prime par écart type. Ces modèles permettent donc d'avoir une idée de la cotation moyenne d'un traité sur le marché de la réassurance avant même sa cotation par les réassureurs. Cependant, en fonction du modèle choisit, les résultats peuvent être fortement différents. Globalement, plus la méthode est simple et transparente plus nos tarifs sont approximatifs, voir faux dans certains cas. La méthode donnant un résultat très

satisfaisant utilise des forêts aléatoires qui sont donc par essence moins explicites que notre modèle par minimisation. La limite de ce modèle est donc sa transparence. Le modèle le plus explicite présentant les meilleurs résultats est celui par minimisation sur des clusters par l'algorithme PAM. Il ne nécessite aucune prédiction et les formules de tarifications sont désormais connues. Cependant, certains clusters sont moins bien estimés mais ils sont connus. Ainsi, un choix possible est tout simplement de ne pas les considérer.

Cette tarification a été développée en entreprise à travers un outil R permettant aux collaborateurs de tarifer leurs traités. Ainsi, lors de chaque année de renouvellement, cet outil a vocation à être utilisé pour anticiper les cotations des réassureurs. De plus, un programme de mise à jour a été produit conjointement à l'outil de tarification afin de permettre aux modèles de se perfectionner d'année en année. Les collaborateurs peuvent alors enrichir les modèles par les données qu'ils produiront chaque année. Ces méthodes de tarification ont donc vocation à être perfectionnées dans le temps.

Pour conclure, plusieurs améliorations peuvent être apportées. En tarification, la prédiction du groupe lorsque nous segmentons par signe et type de β peut se faire avec des algorithmes plus performants que des CART bien qu'ils soient bien moins transparents et simples à utiliser. Nous pouvons par exemple utiliser un ensemble de prédicteurs de plusieurs algorithmes et utiliser la prédiction majoritaire. De plus, il peut être aussi intéressant de construire un modèle prédisant à la fois la cotation maximale, moyenne et minimale. Enfin, une façon supplémentaire d'analyser les réassureurs serait donc d'essayer de modéliser le facteur de diversification et de l'appliquer dans le calcul de la capacité totale afin d'observer la capacité réelle engagée par réassureur. De plus, afin d'étudier l'ensemble des années de cotations dans un même temps, il est possible de moyenner certains de nos indicateurs annuels.

Note de synthèse

Contexte

La réassurance est un domaine assez différent de l'assurance. Il se caractérise en particulier par la façon dont est créé un produit de réassurance (un traité ou un contrat). En assurance, l'assureur élabore seul son produit. De sa modélisation à sa mise sur le marché, l'assureur est le seul acteur dans sa réalisation. En réassurance, nous assistons à une inversion des rôles. L'assureur est placé en tant que cédante (il achète de la réassurance). Il émet les garanties qu'il souhaite dans son traité (priorité, portée ...) mais il ne le tarifie pas. En effet, les tarifs sont élaborés par les réassureurs lorsque le traité est mis sur le marché. L'estimation de la prime commerciale du traité par un réassureur se nomme une cotation. Cette donnée est le point central des études menées dans ce mémoire.

En réassurance, la cédante est dans une position similaire à celle de l'assuré bien qu'elle soit capable en pratique de juger l'ordre de grandeur de ses primes. Pour tous ses traités, elle reçoit différentes primes mais elle n'a que peu de vision sur le modèle de tarification sous-jacent. La cédante fait donc face à un manque de transparence sur l'explicabilité des cotations de ses traités. Ce mémoire s'attache donc à répondre à la problématique suivante :

Comment une cédante peut-elle exploiter les cotations de ses réassureurs ?

Objectifs

Les études réalisées dans ce mémoire portent sur deux échelles de temps : pré-cotation et post-cotation. Lorsqu'elle est en mesure de modéliser les récupérations de ses traités, c'est-à-dire le montant moyen annuel de sinistre à charge d'un traité en réassurance XS, la cédante peut utiliser les cotations pour construire un modèle de tarification interne. Ceci se place dans le cadre pré-cotation puisque que nous cherchons à anticiper les cotations des réassureurs avant qu'elles ne soient communiquées. En disposant des récupérations moyennes et de la volatilité de celles-ci, il est possible de tarifer un traité en se basant sur le principe de prime par écart type :

$$PC = \frac{\bar{R} + \beta \times \sigma_R}{Rec_{factor} \times (1 - \alpha)}$$

Avec PC la prime commerciale du traité (ici la cotation moyenne des réassureurs), $\bar{R}$ la récupération moyenne, σ_R l'écart type des récupérations, β le coefficient de chargement, Rec_{factor} le facteur de reconstitution (part des reconstitutions dans la prime) et α le taux de frais. Nous disposons d'ores et déjà de PC, $\bar{R}$, σ_R et Rec_{factor}. La valeur de α est fixée après une étude des frais de courtage d'AXA Global Re et des frais de gestion moyens des réassureurs. Ainsi, il reste à déterminer la variable β définie par

$$\beta = \frac{PC \times Rec_{factor} \times (1 - \alpha) - \bar{R}}{\sigma_R}$$

Cette variable β représente le lien entre les modélisations internes de la cédante (AXA Global Re) et les réassureurs. L'objectif principal de cette tarification par écart type est donc de prédire, pour un nouveau traité, son coefficient de chargement β associé afin d'estimer directement en interne sa cotation moyenne. Pour ce faire, deux grandes classes de modèles sont utilisées :

— **Modèles par β constant :** pour un ensemble de traités, nous estimons β par un algorithme de minimisation qui fixe une valeur optimale $\hat{\beta}^*$. Il permet d'obtenir une formule de tarification unique pour un ensemble de traités. Ainsi, lorsqu'un nouveau traité doit être tarifé, nous lui associons un $\hat{\beta}^*$ basé sur les différents ensembles de traités utilisés préalablement. La prime commerciale peut alors être estimée.
— **Modèles par prédiction de β :** nous entraînons un modèle dans le but de prédire le β d'un nouveau traité. Il permet d'obtenir une formule de tarification unique pour chaque nouveau traité à tarifer.

Ces modèles réalisés, la cédante peut donc anticiper, avec une certaine précision en fonction de la méthode, les cotations des réassureurs. De plus, elle peut aussi exploiter les cotations pour établir des comparaisons entre ses partenaires de réassurance. Cette analyse se place dans la phase post-cotations.

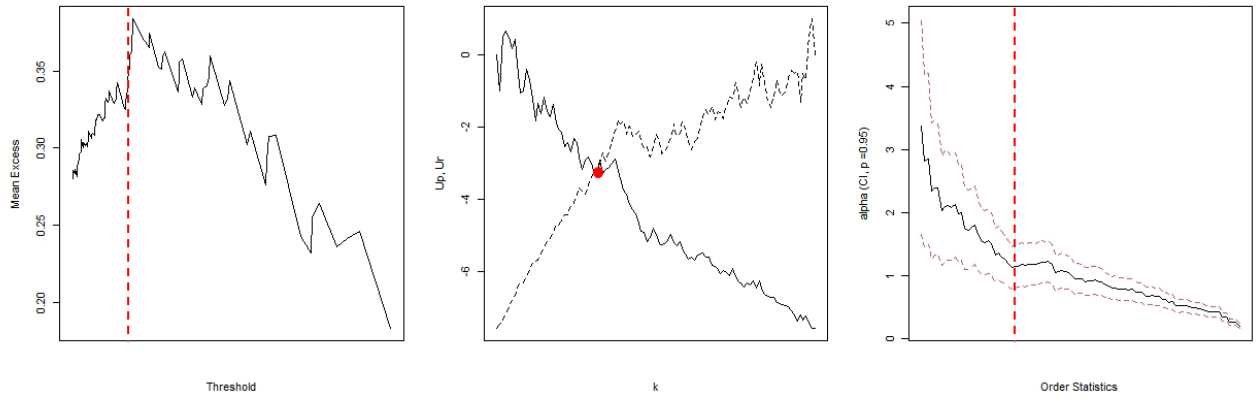
Modèles de tarifications

Plusieurs modèles de tarifications très différents sont développés dans ce mémoire. Chaque méthode présente des avantages et des inconvénients. Celle par β constant utilise des modèles boîtes blanches se montrant transparents mais peu efficaces dans certains cas. La première méthode développée propose de créer plusieurs groupes en fonction de la valeur *réelle* de β estimée pour chaque traité :

— **Groupe 1 :** $\beta < 0$
— **Groupe 2 :** $\beta \geq 0$ et attritionnel
— **Groupe 3 :** $\beta \geq 0$ et atypique

À première vue, une valeur négative de β ne semble pas naturelle. Cependant, cela est tout à fait sensé dans notre cas. Deux raisons principales peuvent expliquer ce phénomène. Premièrement, une vision pessimiste des modélisations de nos traités comparée à la vision des réassureurs peut engendrer des cas où la prime pure modélisée en interne est supérieure à la cotation des réassureurs. Deuxièmement, il convient de rappeler que les réassureurs disposent d'un pouvoir de diversification et de mutualisation. Or, nos traités sont modélisés indépendamment un à un. Ainsi, il n'y a aucun phénomène de compensation modélisé dans les récupérations de nos traités. Cependant, les réassureurs possédant d'ores et déjà un grand nombre de traités divers, ils peuvent proposer un tarif bien plus faible que s'ils réassuraient uniquement le traité coté.

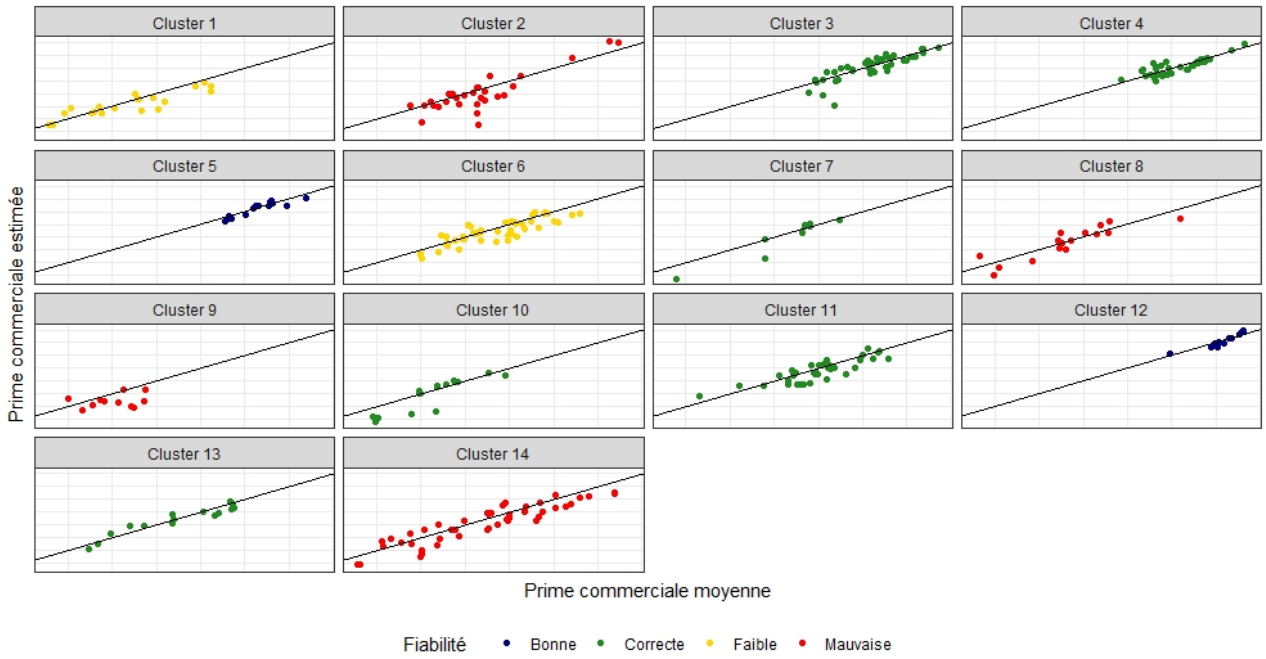
Chaque groupe est déterminé après le calcul ligne à ligne de β. Le seuil atypique de $\beta \geq 0$ est choisi après étude de trois méthodes de détermination de seuil extrême.



Détermination du seuil atypique de β des traités MTPL et GTPL

Une fois ce seuil fixé, nous sommes en capacité d'affilier chaque traité à son groupe en fonction de sa valeur de β. Pour chaque groupe, nous appliquons alors un algorithme de minimisation permettant de trouver une valeur optimale de β notée $\hat{\beta}^*$. Ainsi, nous disposons pour chaque groupe d'une formule de tarification fixe identique. La cédante dispose donc d'une méthode simple de tarification lui permettant de tarifer plusieurs groupes de traités communs. Lorsqu'un nouveau traité doit être tarifé, nous prédisons son groupe à l'aide d'algorithmes de machine learning. Cette prédiction s'avère peu précise notamment pour les groupes 1 et 3. Pour y remédier, nous nous proposons de garder la même idée de tarification par groupe en imposant une façon déterministe d'affilier un traité à son groupe.

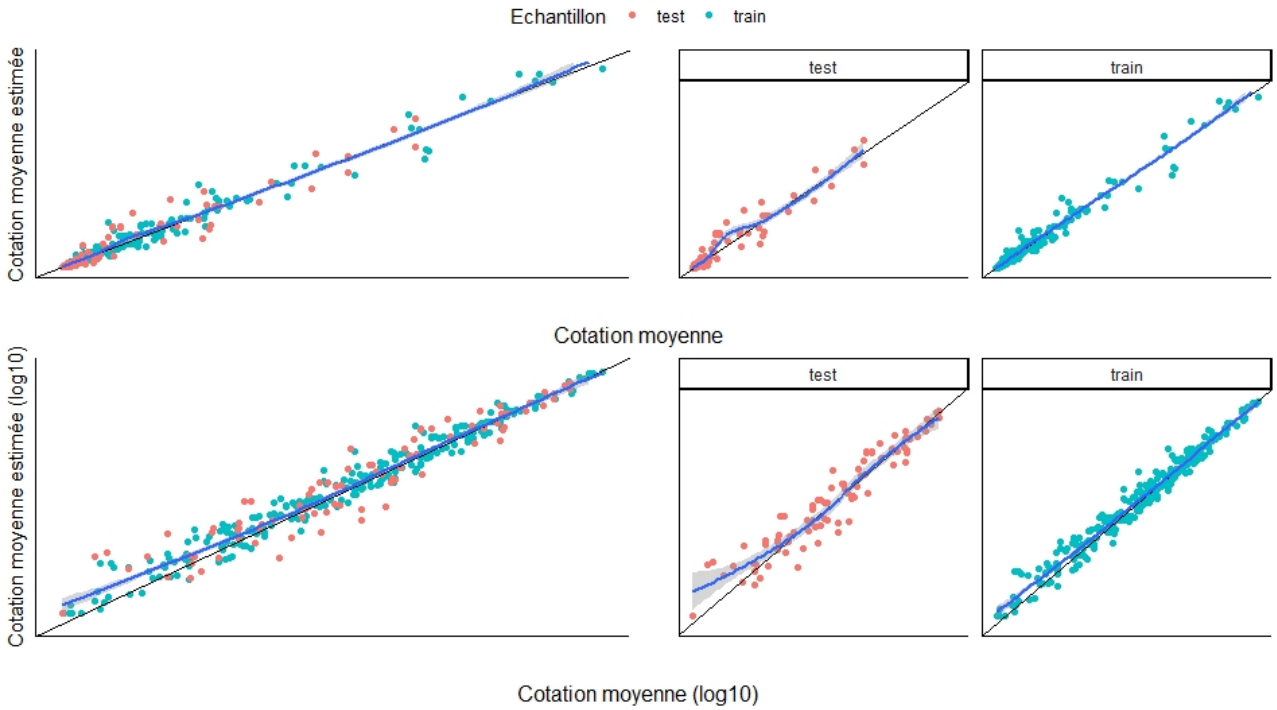
Une solution efficace pour répondre à cette problématique est l'utilisation de l'algorithme PAM pour Partitioning Around Medoids. Basée sur la distance de Gower, cette méthode permet de créer des clusters homogènes de traités. Pour chaque cluster, un médoïde est choisi comme individu de référence. Ainsi, pour affilier un nouveau traité à son cluster, il convient simplement de choisir le médoïde le plus proche de celui-ci (au sens de la distance de Gower). Alors, nous estimons un $\hat{\beta}^*$ par cluster avec notre algorithme de minimisation. Cette méthode crée 14 clusters où chacun possède un unique $\hat{\beta}^*$. Pour chaque traité de chaque cluster, il est donc possible d'estimer la cotation moyenne et de la comparer avec la cotation moyenne réelle proposée par les réassureurs.



Estimations des cotations par cluster après application du PAM

L'observation de ces estimations met en évidence certains clusters qui disposent de traités ayant une valeur de β très proche. Ainsi, l'algorithme de minimisation calcule une valeur optimale $\hat{\beta}^*$ nous permettant d'estimer avec précision la cotation des réassureurs (clusters 5 et 12). A contrario, certains clusters sont hétérogènes dans leurs valeurs de β. Son estimation se montre donc moins précise. Néanmoins, la façon de déterminer le cluster d'un nouveau traité se fait sans aléa (contrairement à la méthode précédente) mais en fonction du cluster, l'estimation de la prime commerciale peut se révéler peu précise. Cette méthode possède tout de même l'avantage non négligeable d'être relativement simple et transparente dans sa façon de tarifier et classer les traités.

Pour palier au manque de performance de certains clusters, nous développons une méthode par prédiction directe de β. Pour ce faire, nous utilisons notamment l'algorithme de prédiction des forêts aléatoires présentant l'avantage d'être performant mais l'inconvénient d'être peu explicable. Ainsi, il fait partie des modèles dits boîtes noires car il souffre d'un manque d'explicabilité comparé aux deux méthodes précédentes. Après une optimisation des paramètres de cet algorithme, les prédictions de β semblent être tout à fait satisfaisantes. Ainsi, en fonction de la base d'apprentissage et de test, nous estimons les primes commerciales en nous basant sur les valeurs prédites des β.



Cotations estimées à partir de l'algorithme des forêts aléatoires de β

Les performances sont satisfaisantes. Nous remarquons par ailleurs une moins bonne précision sur l'échantillon de test indiquant un léger effet d'overfitting. Avec cet algorithme, la cédante peut alors tarifer un nouveau traité en prédisant directement sa valeur de β pour estimer la prime commerciale.

De plus, chaque méthode ayant ses avantages et ses inconvénients, la cédante a le choix parmi un panel de plusieurs méthodes.

		Performance		Transparence	
Méthode	**Algorithme**	Prédiction du groupe/cluster	Estimation de β	Prédiction du groupe/cluster	Estimation de β
Chargement constant	Par risque	Non concerné	Faible	Non concerné	Bien
	Par risque et signe de β	Non concerné	Faible	Non concerné	Bien
	Par risque, signe et type de β	Faible	Bien	Moyen	Bien
	Par clusters (k-means)	Bien	Moyen	Bien	Bien
	Par clusters (PAM)	Bien	Moyen	Bien	Bien
Prédiction directe	GLM	Non concerné	Moyen	Non concerné	Moyen
	Random forest	Non concerné	Bien	Non concerné	Faible

Echelle
Bien
Moyen
Faible
Non concerné

Comparaison des différents modèles de tarification

Ainsi, en fonction de ses besoins de transparence et de performance, la cédante est à même de choisir le modèle lui semblant le plus adapté à ses besoins. Elle peut aussi tout à fait utiliser tous ces modèles afin de comparer les différentes primes estimées et ainsi choisir une valeur finale potentiellement plus pertinente que si elle n'utilisait qu'un seul algorithme. Passons désormais à la partie post-cotation consistant à analyser les réassureurs.

Comparaison des réassureurs

Les cotations des réassureurs nous révèlent en réalité plusieurs informations sur leurs comportements lorsqu'ils cotent un traité d'AXA Global Re. Nous pouvons donc tenter de les comparer sur plusieurs critères comme leur notation S&P, le nombre de traités cotés, la répartition de leurs cotations par risque ou encore la capacité totale définie par la somme des portées des traités cotés. De plus, nous procédons à la création d'un score de précision par réassureur. Il mesure l'écart entre sa cotation et la cotation moyenne pour un même traité. Nos modèles estimant la cotation moyenne, la mesure de cet écart par réassureur nous permet d'apprécier la qualité de nos estimations pour chaque réassureur.

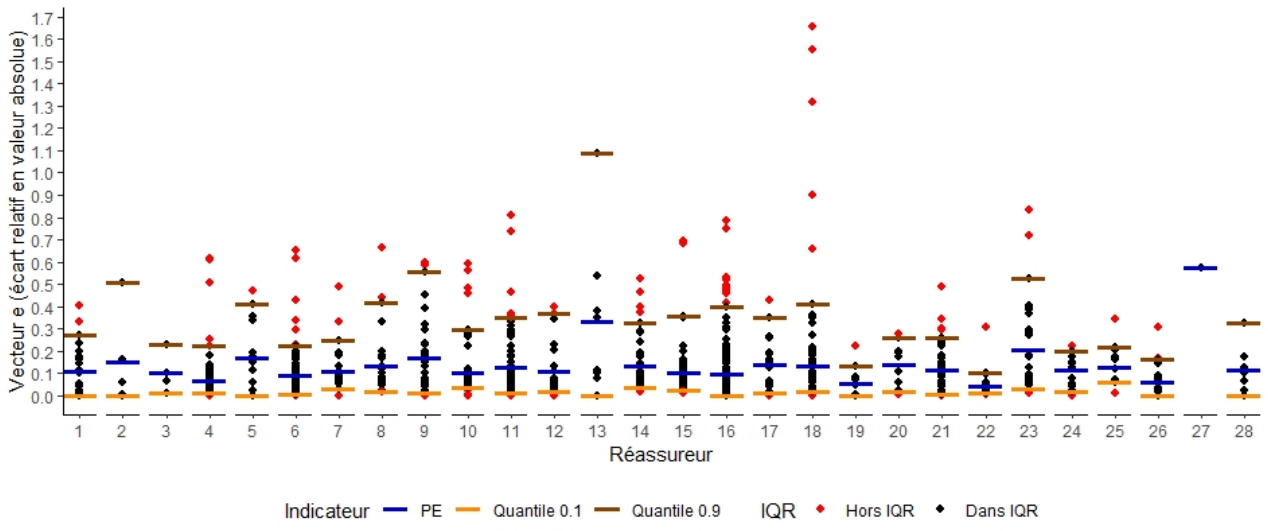
Lorsqu'un réassureur cote généralement plus ou moins haut que la cotation moyenne, nos modèles ne parviennent pas à correctement estimer ses cotations. Inversement, si ce réassureur est proche de la cotation moyenne, nos modèles peuvent correctement estimer ses futures cotations. Pour agréger en un seul indicateur statistique les écarts entre cotations et cotations moyennes par réassureur, nous créons le score PE pour *précision estimée* :

$$PE = \frac{1}{m} \sum_{i=1}^{n} e_i \times \mathbb{1}\{e_i \in IQR\}$$

Ainsi, pour chaque réassureur R, nous calculons son score de précision avec :

- $e = e_1, \ldots, e_n$ le vecteur des erreurs relatives calculées pour le réassureur R avec $e_i = |\text{cotation}_i - \text{cotation moyenne}_i|/\text{cotation moyenne}_i$.
- l'intervalle interquartile $IQR = [e_{(0.1)}, e_{(0.9)}]$ avec $e_{(0.1)}$ (respectivement $e_{(0.9)}$) le quantile à 10 % (respectivement à 90 %) de e.
- $m = \sum_{i=1}^{n} \mathbb{1}\{e_{(0.1)} \leq e_i \leq e_{(0.9)}\}$ le nombre de valeurs de e comprises dans IQR.

Le résultat de ces calculs, par réassureur, est le suivant :



Calcul de PE par réassureur

Sur ce graphique, le score par réassureur est représenté par le trait bleu. Plus celui-ci est proche de 0, plus nous pouvons espérer que nos modèles de tarification soient adaptés à ce réassureur. Un classement par réassureur en fonction de sa valeur de PE est alors établi.

Executive summary

Context

Reinsurance is a rather different field from insurance. It is characterized in particular by the way in which a reinsurance product (a treaty or a contract) is created. In insurance, the insurer develops his product alone. From its modeling to its marketing, the insurer is the only agent in the realization of its product. In reinsurance, we are observing a inversion of roles. The insurer is placed as a cedant (he buys reinsurance). It provides the guarantees that it wishes in its treaty (priority, limit...) but it does not price it. Indeed, the prices are made by the reinsurers when the treaty is put on the market. The estimate of the commercial premium of the treaty by a reinsurer is called a quotation. This data is the focus of the studies conducted in this paper.

In reinsurance, the ceding company is in a similar position to the insured, although in practice it is able to judge the size of its premiums. For all its treaties, it receives different premiums on which it has limited knowledge of the underlying pricing model. The ceding company is therefore faced with a lack of transparency on the explicability of its treaty quotes. This paper aims to answer the following question :

How can a ceding company exploit the quotations of its reinsurers ?

Objectives

The studies conducted in this paper focus on two time scales : pre-quotation and post-quotation. When the cedant is able to model the recoveries of its treaties, i.e. the average annual loss amount for an XS reinsurance treaty, it can use the quotations to build an internal pricing model. This is in the pre-quotation context since we are trying to anticipate the reinsurers' quotations before they are made. Given the average recoveries and their volatility, it is possible to price a treaty based on the following standard deviation premium principle :

$$PC = \frac{\bar{R} + \beta \times \sigma_R}{Rec_{factor} \times (1 - \alpha)}$$

With PC the commercial premium of the treaty (here the average quotation of the reinsurers), $\bar{R}$ the average recovery, σ_R the standard deviation of the recoveries, β the loading factor, Rec_{factor} the reconstitution factor (share of the reconstitutions in the premium) and α the expense ratio. We already have PC, $\bar{R}$, σ_R and Rec_{factor}. The value of α is determined after a study of AXA Global Re's brokerage fees and the average management fees of reinsurers. Thus, it remains to determine the variable β defined by

$$\beta = \frac{PC \times Rec_{factor} \times (1 - \alpha) - \bar{R}}{\sigma_R}$$

This β variable represents the link between the internal modelling of the ceding company (AXA Global Re) and the reinsurers. The main objective of this standard deviation pricing is therefore to predict, for a new treaty, its associated loading factor β in order to directly estimate internally its average quotation. To do this, two main classes of models are used :

— **Models by constant β :** for a set of treaties, we estimate β by a minimization algorithm that sets an optimal value of β called $\hat{\beta}^*$. This allows us to obtain a pricing formula for a set of treaties. Thus, when a new treaty needs to be priced, we associate a $\hat{\beta}^*$ to it based on the different sets of treaties used previously. The commercial premium can then be estimated.

— **Models by prediction of β :** we train a model to predict the β of a new treaty. This provides a unique pricing formula for each new treaty to be priced.

Once these models have been completed, the ceding company can anticipate the reinsurers' quotations with a certain degree of accuracy, depending on the method used. Moreover, it is also possible to use the quotations to establish comparisons between our reinsurance partners. This analysis takes place in the post-quotation phase.

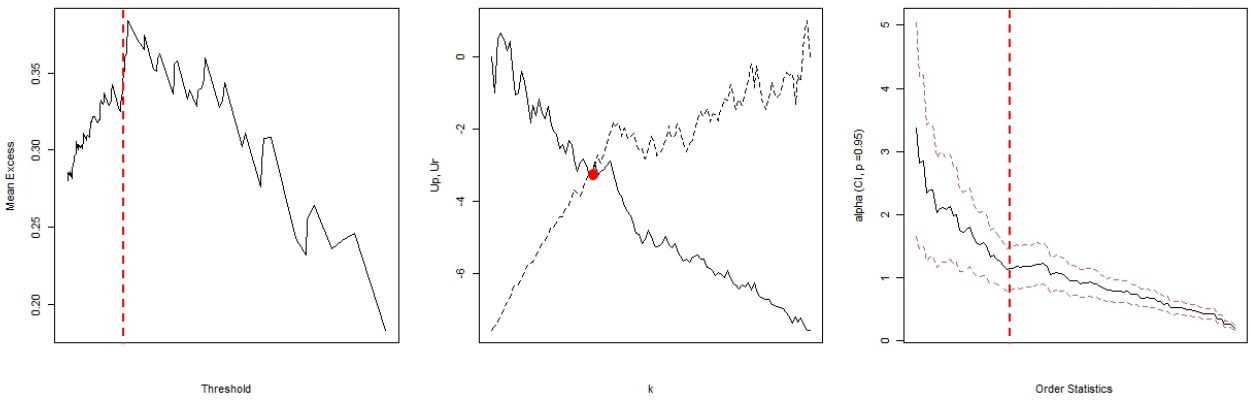
Pricing models

Several very different pricing models are developed in this paper. Each method has its own advantages and disadvantages. The constant β method uses white box models which are transparent but not very efficient in some cases. The first method developed proposes to create several groups according to the *real* value of β calculated for each treaty :

— **Group 1 :** $\beta < 0$
— **Group 2 :** $\beta \geq 0$ and attritional
— **Group 3 :** $\beta \geq 0$ and atypical

At first sight, a negative value of β does not seem natural. However, it makes sense in our case. There are two main reasons for this. First, a pessimistic view of our treaty models compared to the reinsurers' view can lead to cases where the internally modeled pure premium is higher than the reinsurers' quote. Secondly, it should be remembered that reinsurers have the power of diversification and mutualization. However, our treaties are modeled independently of each other. Thus, there is no compensation phenomenon modeled in the recoveries of our treaties. So, when reinsurers already own a large number of diverse treaties, they can quote a much lower price than if they were only reinsuring the quoted treaty.

Each group is determined after a line-by-line calculation of β. The atypical threshold of $\beta \geq 0$ is chosen after studying three extreme threshold determination methods.

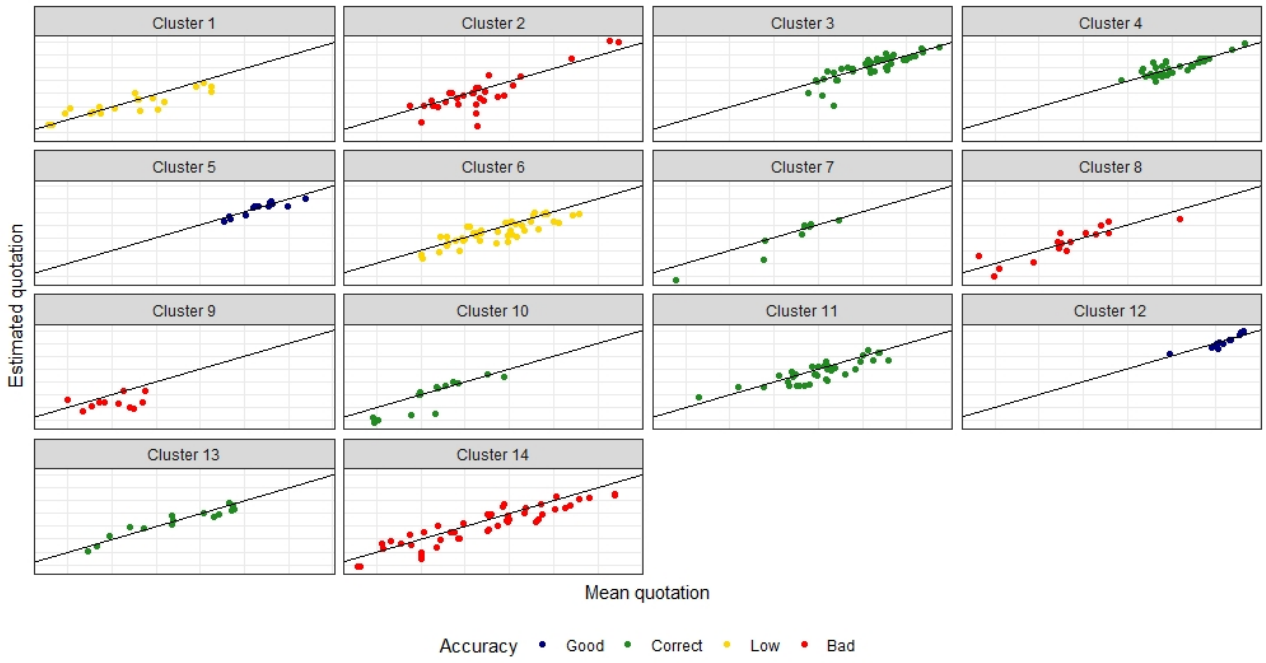


Atypical β threshold determination of MTPL and GTPL treaties

Once this threshold is set, we are able to assign each treaty to its group according to its β value. For each group, we then apply a minimization algorithm to find an optimal value of β. Thus, we have an identical fixed pricing formula for each group. This provides the cedant with a simple pricing method that gives it a way to price multiple groups of common treaties. When a new treaty needs to be priced, we predict its group using machine learning algorithms. This prediction is not very accurate, especially for groups 1 and 3. To solve this problem, we propose

to keep the same idea of pricing by groups by imposing a deterministic way of associating a treaty to its group.

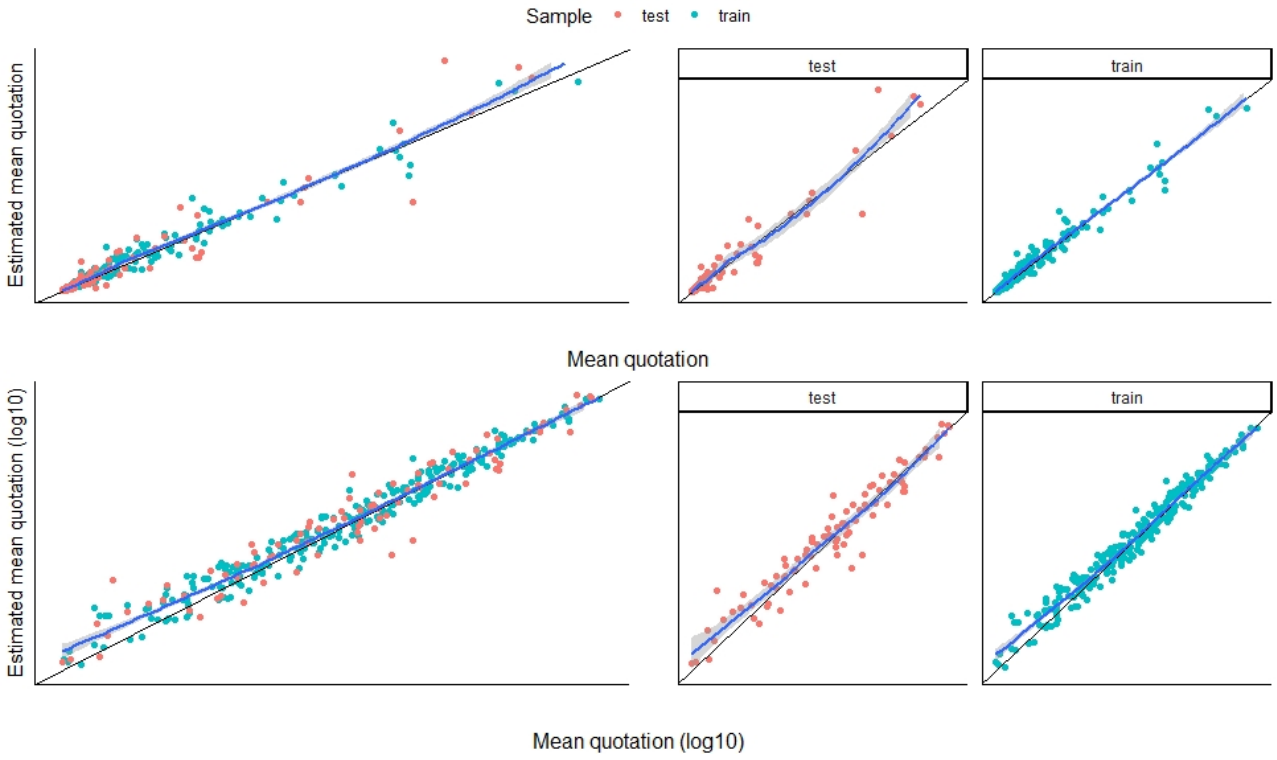
An efficient solution to this problem is to use the PAM algorithm for Partitioning Around Medoids. Based on the Gower distance, this method allows to create homogeneous clusters of treaties. For each cluster, a medoid is chosen as reference individual. Thus, to affiliate a new treaty to its cluster, it is only necessary to choose the medoid closest to it (in the sense of the Gower distance). Then, we estimate with our minimization algorithm a β per cluster. This method creates 14 clusters where each has a unique β. For each treaty of each cluster, it is thus possible to estimate the average quotation and to compare it with the real average quotation proposed by the reinsurers.



Estimates of quotations by cluster after application of the PAM

As we observe, some clusters have treaties with a very close value of β. Thus, the minimization algorithm computes an optimal $\hat{\beta}^*$ value allowing us to estimate with precision the quotation of the reinsurers (clusters 5 and 12). On the opposite, some clusters are heterogeneous in their β values. Its estimation is consequently less precise. Nevertheless, the way of determining the cluster of a new treaty is done without any randomness (contrary to the previous method) but depending on the cluster, the estimation of the commercial premium can be not very precise. This method still has the significant advantage of being relatively simple and transparent in the way it prices and classifies treaties.

To overcome this lack of performance of some clusters, we develop a method by direct prediction of β. To do this, we use the random forest prediction algorithm which has the advantage of being efficient but not very explainable. Thus, it is part of the so-called black box models because it suffers from a lack of explicability compared to the two previous methods. After an optimization of the algorithm's parameters, the predictions of β seem to be quite satisfactory. So, according to the learning and testing base, we estimate the commercial premiums based on the predicted values of β.



Quotations estimated from the random forest algorithm of β

The performances are quite satisfactory. However, we notice a lower accuracy on the test sample showing a slight overfitting effect. With this algorithm, the cedant can then price a new treaty by directly predicting its β value in order to estimate the commercial premium.

Moreover, each method has its advantages and disadvantages, so the cedant has a choice of several methods.

Method	Algorithm	Performance: Group/cluster prediction	Performance: Estimation of β	Transparency: Group/cluster prediction	Transparency: Estimation of β
Constant loading	By risk	Not concerned	Poor	Not concerned	Well
	By risk and sign of β	Average	Average	Average	Well
	By risk, sign and type of β	Poor	Well	Average	Well
	By clusters (k-means)	Well	Average	Well	Well
	By clusters (PAM)	Well	Average to Well	Well	Well
Direct prediction	GLM	Not concerned	Average	Not concerned	Average
	Random forest	Not concerned	Well	Not concerned	Poor

Scale
Well
Average
Poor
Not concerned

Comparison of different pricing models

Thus, depending on its needs for transparency and performance, the cedant is able to choose the model that seems to be the most adapted to its needs. It can also use all these various models in order to compare the different estimated premiums and thus choose a final value that is potentially more relevant than if it were to use only one algorithm. Let's now move on to the post-quotation part, which consists of analyzing the reinsurers.

Reinsurer Comparison

The reinsurers' ratings actually reveal several details about their behavior when they quote an AXA Global Re treaty. We can therefore attempt to compare them on several criteria such as their S&P rating, the number of treaties quoted, the distribution of their quotes by risk or the total capacity defined by the sum of the limits of the treaties quoted. In addition, we create an accuracy score for each reinsurer. It measures the difference between its quotation and the average quotation for the same treaty. Since our models estimate the average quotation, measuring this deviation by reinsurer allows us to know if the quotations of a reinsurer are well modeled by our models.

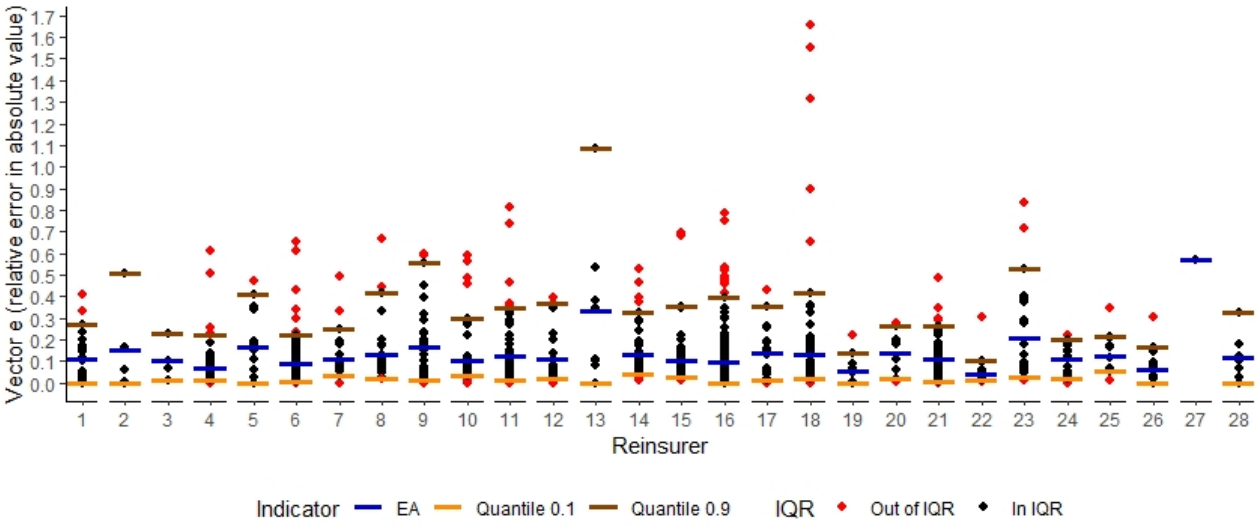
When a reinsurer quotes generally higher or lower than the average quotation, our models fails to correctly estimate its quotations. Conversely, if this reinsurer is close to the average quotation, our models can correctly estimate its future quotations. To aggregate the differences between quotations and average quotations by reinsurer into a single statistical indicator, we create the score EA for *estimated accuracy* :

$$EA = \frac{1}{m} \sum_{i=1}^{n} e_i \times \mathbb{1}\{e_i \in IQR\}$$

Thus, for each reinsurer R, we calculate its accuracy score with :

— $e = e_1, \ldots, e_n$ the vector of relative errors calculated for reinsurer R with $e_i = |\text{quotation}_i - \text{average quotation}_i| / \text{average quotation}_i$.
— the interquartile range $IQR = [e_{(0.1)}, e_{(0.9)}]$ with $e_{(0.1)}$ (respectively $e_{(0.9)}$) the quantile at 10 % (respectively at 90 %) of e.
— $m = \sum_{i=1}^{n} \mathbb{1}\{e_{(0.1)} \leq e_i \leq e_{(0.9)}\}$ the number of values of e included in IQR.

The result of these calculations, by reinsurer, is as follows :



Calculation of EA by reinsurer

On this graph, the score per reinsurer is represented by the blue line. The closer it is to 0, the more we can expect our pricing models to be adapted to this reinsurer. A ranking by reinsurer according to its EA value is then established.

Références

[1] Institut des Actuaires / Armelle Guillou et Alexandre You. *Introduction à la théorie des valeurs extrêmes : Applications en actuariat*. URL : https://www.institutdesactuaires.com/global/gene/link.php?doc_id=657&fg=1.

[2] Emmanuel DUBREUIL - AON Benfield. *Quels risques transférer à un réassureur ?* URL : http://www.ressources-actuarielles.net/ext/ia/sitesepia.nsf/0/e23826985e0b5791c12576e100742cac/%24file/sepia20100706_ed.pdf.

[3] Cristina BUTUCEA. *Introduction à l'Apprentissage Statistique*. ENSAE, 2020.

[4] Arthur CHARPENTIER. *Actuariat IARD - ACT2040 - Partie 4 - modèles linéaires généralisés*. URL : http://freakonometrics.free.fr/slides-2040-4.pdf.

[5] Christophe DUTANG. *Actuariat de l'Assurance Non-Vie*. ENSAE, 2020.

[6] Conseil de l'union européenne. *Règlement général sur la protection des données*. URL : https://www.consilium.europa.eu/fr/policies/data-protection-reform/data-protection-regulation/.

[7] Serap Gönülal. *Motor Third-Party Liability Insurance*. URL : https://openknowledge.worldbank.org/handle/10986/27732?show=full.

[8] Fédération Française de l'Assurance. *Les chiffres de l'assurance en 2020*. URL : https://www.ffa-assurance.fr/etudes-et-chiffres-cles/les-chiffres-de-assurance-en-2020.

[9] Pierre LACOSTE. *Risk management et Réassurance*. ENSAE, 2020.

[10] Daniel P. Martin. *Clustering Mixed Data Types in R*. URL : https://dpmartin42.github.io/posts/r/cluster-mixed-types.

[11] Arnak DALALYAN et Mohamed HEBIRI. *Apprentissage Statistique Appliqué*. ENSAE, 2020.

[12] Christian Y. ROBERT. *Théorie du risque*. ENSAE, 2020.

[13] Statista. *Le big bang du big data*. URL : https://fr.statista.com/infographie/17800/big-data-evolution-quantite-donnees-numeriques-creees-dans-le-monde/.

[14] Pierre-E. THEROND. *Mesures et comparaison de risques*. ISFA, 2005.

[15] Datascientest / Thibault V. *K-means : Focus sur cet algorithme de Clustering et Machine Learning*. URL : https://datascientest.com/algorithme-des-k-means.

[16] Ingrid Hobæk Haffa Yinzhi Wanga et Arne Husebya. *Modelling extreme claims via composite models and threshold selection methods*. URL : https://arxiv.org/pdf/1911.02418.pdf.

Annexes

List of Algorithms

Table des figures

Liste des tableaux

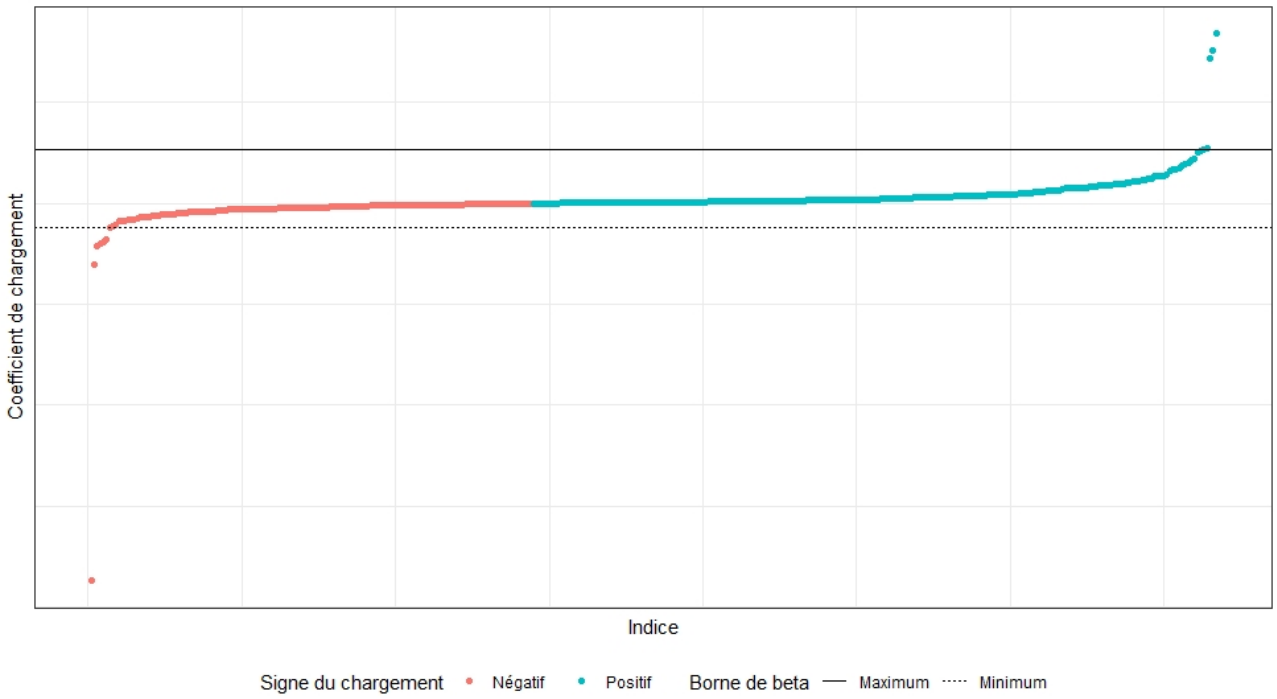


FIGURE 76 – Choix des bornes supérieures et inférieures aberrantes de β

	Reference	
Predicted	Event	No Event
Event	A	B
No Event	C	D

The formulas used here are:

$$Sensitivity = \frac{A}{A+C}$$

$$Specificity = \frac{D}{B+D}$$

$$Prevalence = \frac{A+C}{A+B+C+D}$$

$$PPV = \frac{sensitivity \times prevalence}{((sensitivity \times prevalence) + ((1 - specificity) \times (1 - prevalence))}$$

$$NPV = \frac{specificity \times (1 - prevalence)}{((1 - sensitivity) \times prevalence) + ((specificity) \times (1 - prevalence))}$$

$$Detection\ Rate = \frac{A}{A+B+C+D}$$

$$Detection\ Prevalence = \frac{A+B}{A+B+C+D}$$

$$Balanced\ Accuracy = (sensitivity + specificity)/2$$

$$Precision = \frac{A}{A+B}$$

$$Recall = \frac{A}{A+C}$$

$$F1 = \frac{(1+\beta^2) \times precision \times recall}{(\beta^2 \times precision) + recall}$$

FIGURE 77 – Liste des indicateurs associés à la matrice de confusion en classification binaire

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.57431  -0.13057  -0.03864   0.07787   1.15954

Coefficients:
                  Estimate Std. Error t value             Pr(>|t|)
(Intercept)        0.11993    0.02313   5.184 0.000000259551925784 ***
RisqueGTPL         0.31651    0.02198  14.403 < 0.0000000000000002 ***
RisqueMarine       0.10766    0.02994   3.595             0.000339 ***
RisqueMTPL         0.27518    0.02730  10.079 < 0.0000000000000002 ***
RisquePTY          0.12980    0.02439   5.322 0.000000125479742960 ***
ExpoGeo2          -0.14432    0.03409  -4.233 0.000025024187516816 ***
ExpoGeo6          -0.09660    0.04621  -2.091             0.036791 *
ExpoGeo8          -0.10672    0.02931  -3.641             0.000285 ***
ExpoGeo9          -0.07471    0.02995  -2.494             0.012779 *
ExpoGeo10         -0.11459    0.04999  -2.292             0.022088 *
ExpoGeo11          0.10212    0.04649   2.197             0.028260 *
ExpoGeo12          0.21325    0.05563   3.833             0.000134 ***
ExpoGeo13         -0.17088    0.03114  -5.487 0.000000051211390016 ***
ExpoGeo14         -0.36463    0.08287  -4.400 0.000011929444680245 ***
ExpoGeo15         -0.30711    0.09942  -3.089             0.002061 **
ExpoGeo16         -0.33502    0.05366  -6.243 0.000000000618676795 ***
ExpoGeo18         -0.24771    0.03854  -6.427 0.000000000196307464 ***
ExpoGeo19         -0.31923    0.03906  -8.172 0.000000000000000858 ***
ExpoGeo21         -0.09815    0.05000  -1.963             0.049894 *
ExpoGeo22         -0.19049    0.03090  -6.164 0.000000001009026932 ***
ExpoGeo23         -0.15829    0.03262  -4.853 0.000001398937708173 ***
ExpoGeo26         -0.38610    0.12012  -3.214             0.001347 **
ExhaustionProba   -3.19077    0.73960  -4.314 0.000017517842548891 ***
RecupStandardise  -0.21043    0.02905  -7.243 0.000000000000845492 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.05442475)

    Null deviance: 88.707  on 1081  degrees of freedom
Residual deviance: 57.581  on 1058  degrees of freedom
AIC: -53.32

Number of Fisher Scoring iterations: 2
```

FIGURE 78 – GLM de l'estimation directe de β

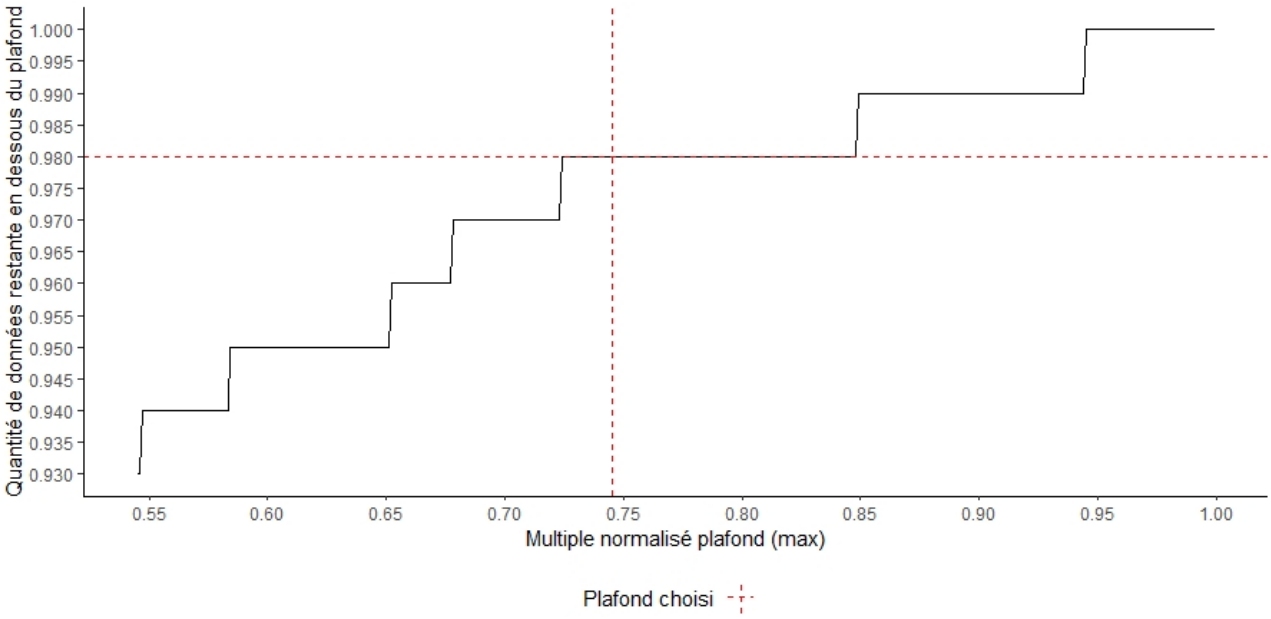


FIGURE 79 – Sensibilité de la quantité de données en fonction du multiple aberrant (plafond)

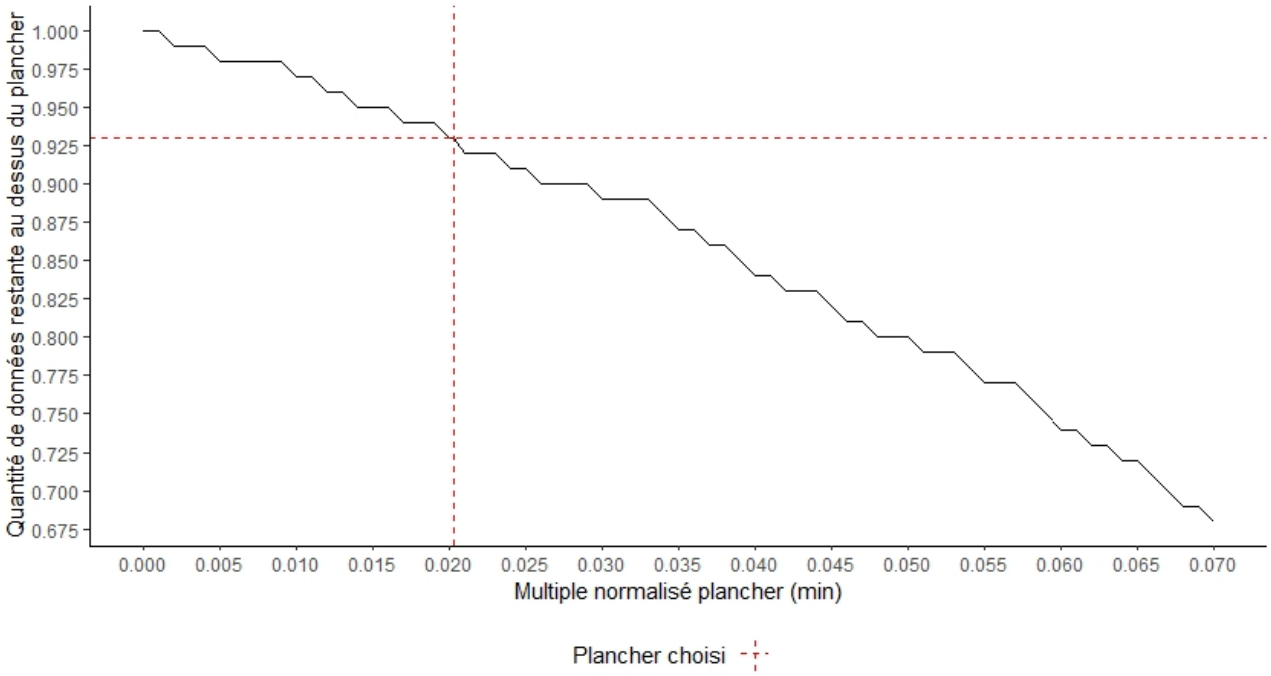


FIGURE 80 – Sensibilité de la quantité de données en fonction du multiple aberrant (plancher)

Réassureur	PE
22	0.0394
19	0.0518
26	0.0623
4	0.0687
6	0.0890
16	0.0984
10	0.1023
3	0.1034
15	0.1041
7	0.1074
1	0.1099
12	0.1104
24	0.1125
21	0.1135
28	0.1175
25	0.1251
11	0.1270
8	0.1300
18	0.1309
14	0.1315
20	0.1374
17	0.1401
2	0.1484
9	0.1673
5	0.1673
23	0.2077
13	0.3315
27	0.5731

TABLE 35 – Classement par réassureur en fonction du score PE de fiabilité du modèle

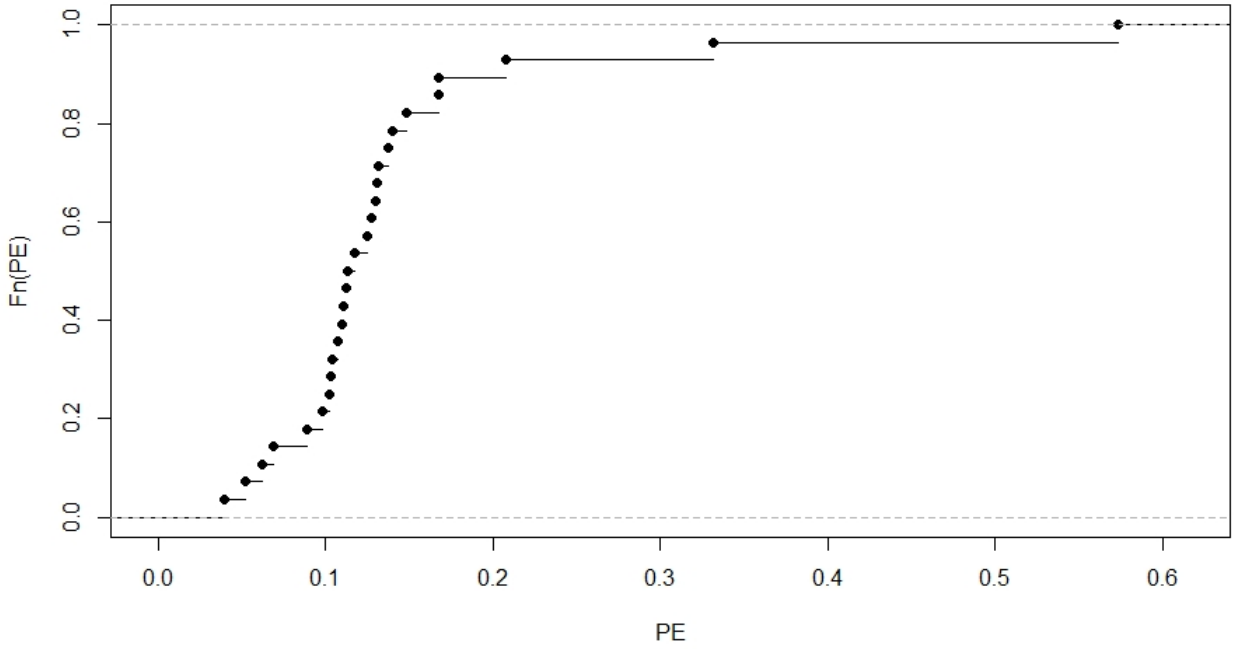


FIGURE 81 – Fonction de répartition empirique de PE

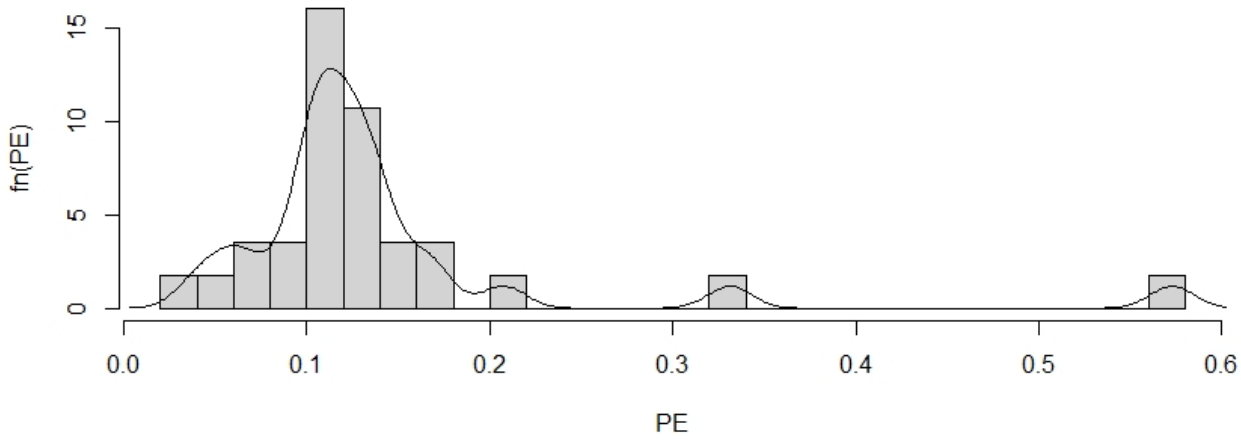


FIGURE 82 – Densité empirique de PE

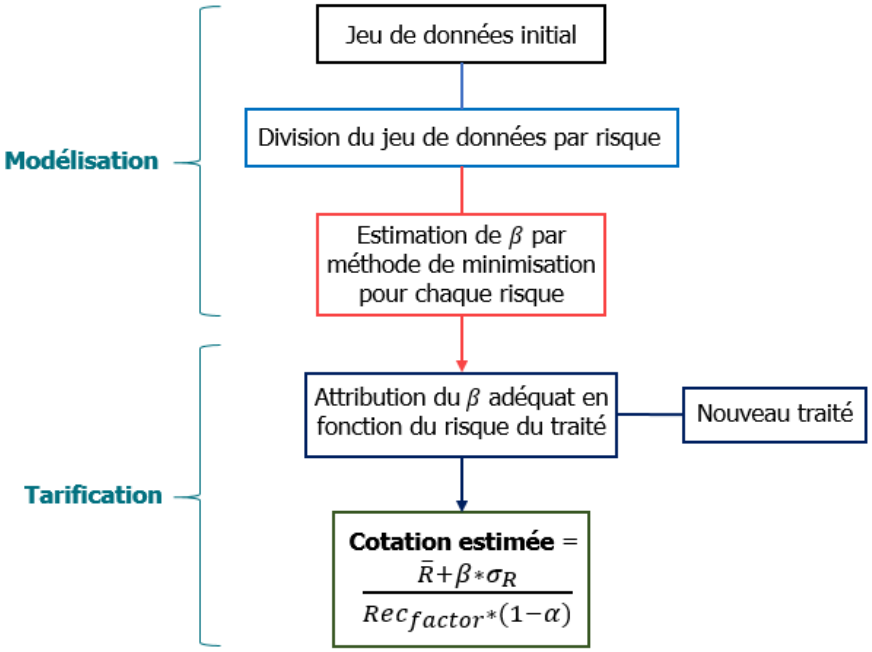


FIGURE 83 – Méthode de modélisation naïve de β

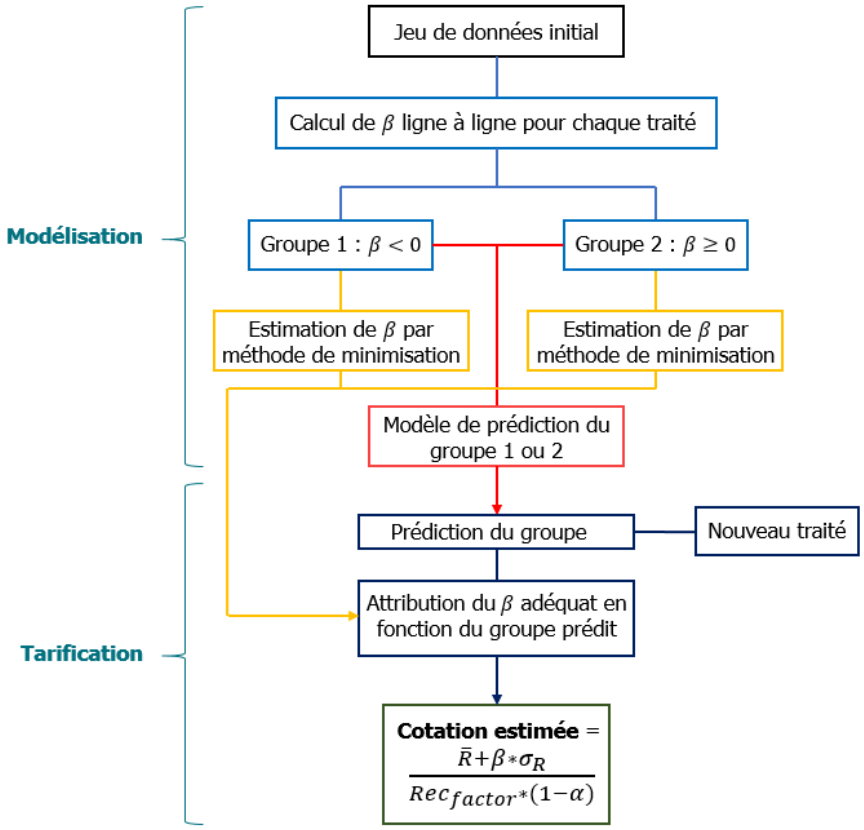


FIGURE 84 – Méthode de modélisation par segmentation par signe de β

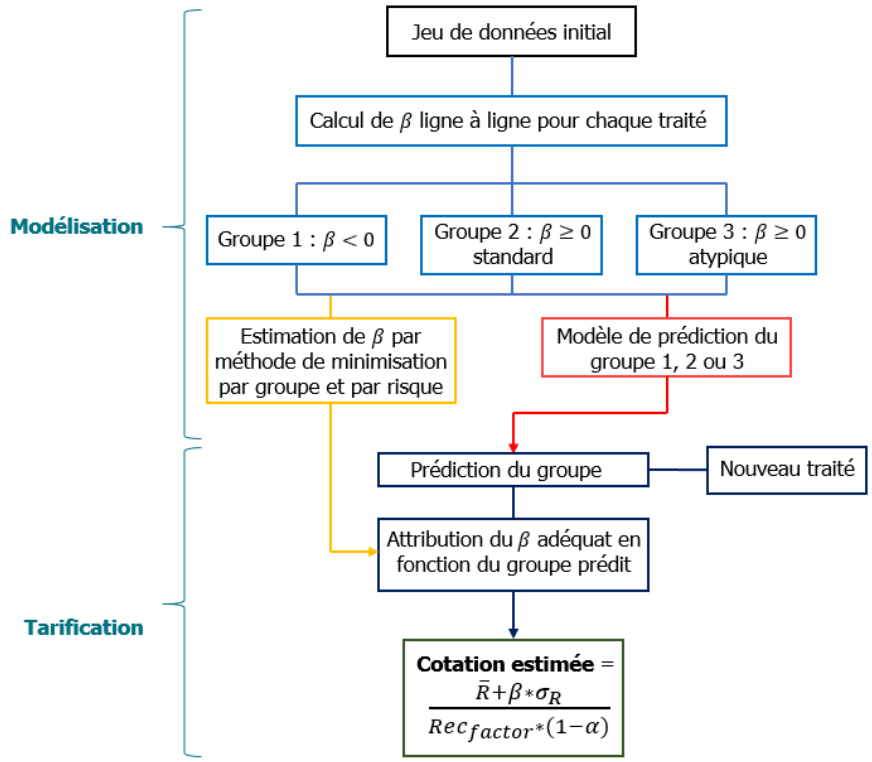


FIGURE 85 – Méthode de modélisation par segmentation par signe et type de β

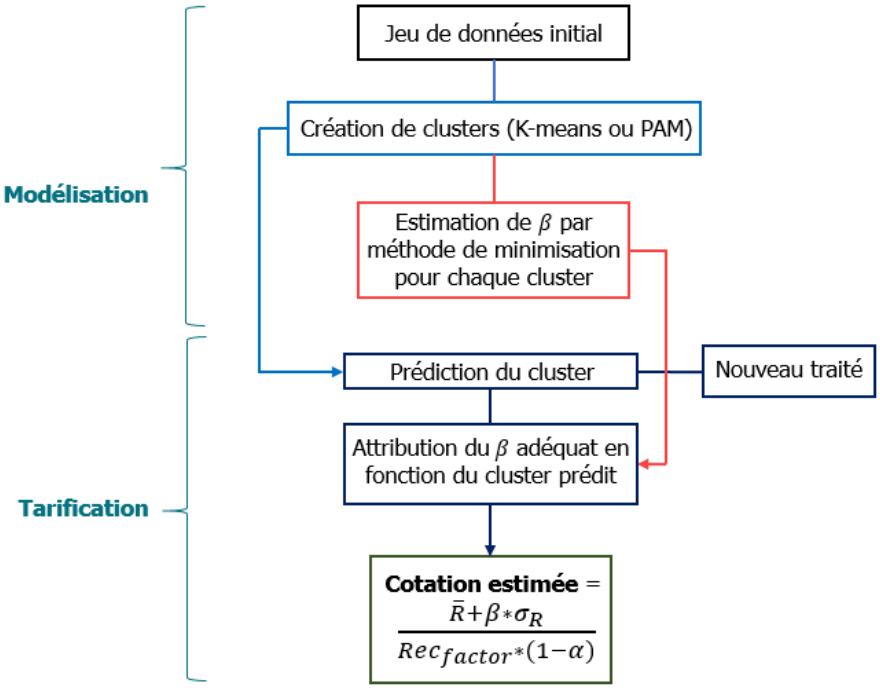


FIGURE 86 – Méthode de modélisation par clusters de β

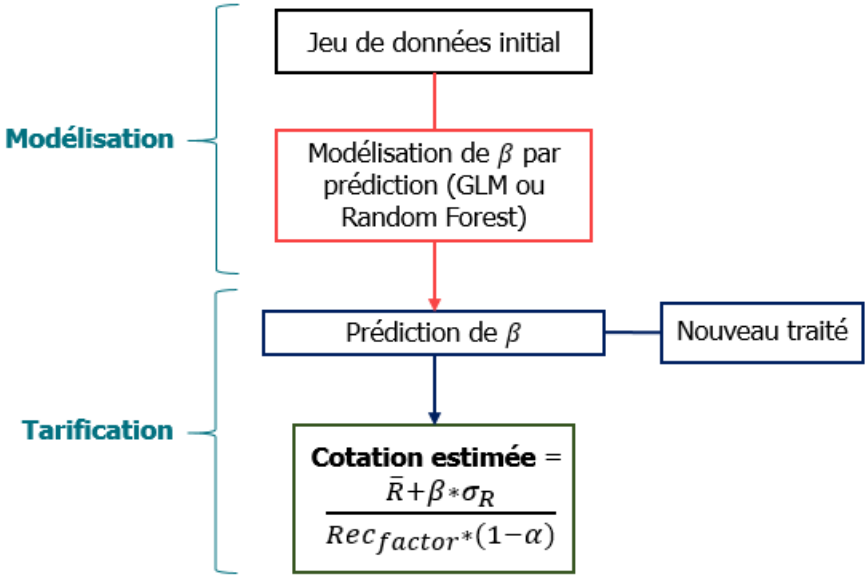


FIGURE 87 – Méthode de modélisation par prédiction directe de β

Propriétés des principes de prime

Un principe de prime se doit de respecter certaines propriétés. Nous nous intéressons à quatre types de propriétés :

1. Propriétés de rationalité : propriétés que doit suivre un principe de prime qui paraissent raisonnables pour la majorité des personnes.
2. Propriétés d'additivité et homogénéité : propriétés caractérisant le comportement de somme de risques ou lorsqu'ils sont multipliés par une constante.
3. Propriétés de comparaison : propriétés permettant de comparer (au sens du plus ou moins risqué) les risques entre eux.
4. Propriétés techniques ou mathématiques : propriétés utilisées pour définir des familles de mesures.

Rationalité

1. **Pas de chargements inutiles** :

$$P(X) \leq \max(X) = \sup\{x : P(X \leq x) < 1\}$$

   Cette propriété traduit le fait qu'il n'est pas possible de demander plus que le sinistre maximum comme prime.

2. **Profitabilité, chargements positifs** :

$$P(X) \geq \mathbb{E}[X]$$

   La prime doit être supérieure à la prime pure si l'assureur souhaite faire du bénéfice et ne pas faire faillite. En effet, dans le cas où $P(X) = \mathbb{E}[X]$, le résultat de l'assureur sera égal à zéro en moyenne. De plus, la probabilité de ruine à long terme de l'assureur convergera vers 1 (la vitesse de convergence dépendant du nombre de risques souscrits) car il sera fortement probable que des sinistres supérieurs à la prime pure surviennent.

3. **Pas de chargements injustifiés** :
   Pour toute constance c,

$$P(c) = c$$

   Si le risque est constant et donc déterministe, le montant en cas de perte est connu. Ainsi, la prime doit être égale au risque car il est certain.

4. **Objectivité** :
   Pour tout risque X et Y de même loi,

$$P(X) = P(Y)$$

   Le principe de prime ne dépend que de la loi du risque.

5. **Translation** :
   Pour toute constance c,

$$P(X + c) = P(X) + c$$

   Les frais fixes s'ajoutent au montant de la prime.

Additivité et homogénéité

6. **Sous-additivité** :
   Pour tous risques X, Y définis sur le même espace de probabilité,

$$P(X+Y) \leq P(X) + P(Y)$$

   Cette propriété est celle de la diversification. La prime de deux risques est plus faible que la somme des primes de chaque risque individuel. Le risque est ainsi réduit par un procédé de diversification. En particulier, si $X \perp\!\!\!\perp Y$ et $P(X+Y) = P(X) + P(Y)$, P satisfait la propriété d'additivité pour les risques indépendants.

7. **Additivité pour les risques comonotoniques** :
   Pour toutes fonctions non décroissantes h et g,

$$P(h(X) + g(X)) = P(h(X)) + P(g(X))$$

   La prime à demander pour deux risques liés de manière croissante avec un risque unique doit être égale à la somme des primes individuelles car il n'y a pas de diversification possible.

   Remarque sur les risques comonotoniques

   Définition
   Un ensemble $\mathcal{E} \subset \mathbb{E}^2$ est dit comonotonique si $\forall\ (x_1, x_2), (y_1, y_2) \in \mathcal{E}$,

$$x_i < y_i \Rightarrow x_j \leq y_j \text{ pour } i \neq j$$

   Propositions

   1. Un vecteur aléatoire (X, Y) est comonotonique si et seulement si X et Y peuvent être écrits comme des fonctions non-décroissantes d'une seule variable aléatoire réelle.
   2. Un vecteur aléatoire (X, Y) est comonotonique si et seulement si

$$\forall x, y \in \mathbb{R}, P(X \leq x, Y \leq y) = \min\{P(X \leq x), P(Y \leq y)\}$$

   3. Un vecteur aléatoire (X, Y) est comonotonique si et seulement si

$$(X, Y) \overset{\mathrm{d}}{=} (F_X^{-1}(U), F_Y^{-1}(U))$$

      où $F_X^{-1}(U)$ est l'inverse de la fonction de répartition de X (appelée fonction quantile de X) et U une variable aléatoire de loi uniforme sur $[0, 1]$.

8. **Homogénéité positive** :
   Pour toute constance positive c,

$$P(cX) = cP(X)$$

   La prime pour assurer une valeur cX (c risques identiques) est simplement la prime $P(X)$ multipliée par c.

Comparaison des risques

9. **Monotonie** :
   Pour tous risques X et Y définis sur le même espace de probabilité,

$$P(X \leq Y) = 1 \quad \Rightarrow \quad P(X) \leq P(Y)$$

   Si le montant du sinistre X est toujours inférieur au montant du sinistre Y alors nécessairement la prime associée au risque X est plus faible que la prime associée au risque Y. Aussi, cette propriété se traduit par le fait que X est moins dangereux que Y.

   Remarque sur les relations d'ordre

   Soit F_X, F_Y, F_Z les fonctions de répartitions respectives de X, Y, Z. La relation binaire $\preceq$ est une relation d'ordre partielle si :

   1. Transitivité : si $F_X \preceq F_Y$ et $F_Y \preceq F_Z$ alors $F_X \preceq F_Z$.
   2. Réflexivité : $F_X \preceq F_X$.
   3. Antisymétrie : si $F_X \preceq F_Y$ et $F_Y \preceq F_X$ alors $F_X = F_Y$.

   À noter qu'il est important de différencier la relation de comparaison *presque sure* définie par $X \leq Y$ p.s où ce sont les réalisations des variables aléatoires (et non les fonctions de répartition) qui sont comparées.

10. **Invariance de type 1** :
    Le risque Y domine stochastiquement le risque X à l'ordre 1, noté $Y \succeq_{DS1} X$ si :

$$P(X > d) \leq P(Y < d) \quad \forall d \in \mathbb{R}$$

    Ainsi, la propriété d'invariance de type 1 est satisfaite si pour tous risques X et Y,

$$X \preceq_{DS1} Y \quad \Rightarrow \quad P(X) \leq P(Y)$$

    Ceci se traduit par le fait que si, pour tout niveau d, la probabilité d'observer un sinistre supérieur à d pour Y est plus grande que pour X alors la prime de Y est plus élevée que celle de X.

11. **Invariance de type 2** :
    Le risque Y domine stochastiquement le risque X à l'ordre 2, noté $X \preceq_{DS2} Y$, si :

$$\mathbb{E}[(X - d)_+] \leq \mathbb{E}[(Y - d)_+] \quad \forall d \in \mathbb{R}$$

    La propriété d'invariance de type 2 est satisfaite si pour tous risques X, Y :

$$X \preceq_{DS2} Y \quad \Rightarrow \quad P(X) \leq P(Y)$$

    Ce principe se traduit par le fait que, pour tout niveau d, si le montant moyen du sinistre au-delà du niveau d est plus grand pour le risque Y que pour le risque X alors la prime associée à Y est plus élevée que la prime associée à X.

Propriétés techniques ou mathématiques

12. **Convexité** :
    Pour tous risques X et Y définis sur le même espace de probabilité, pour toute constante $\alpha \in [0,1]$,
    $$P(\alpha X + (1-\alpha Y)) \leq \alpha P(X) + (1-\alpha)P(Y).$$
13. **Itérativité** :
    Pour tous risques X et Y définis sur le même espace de probabilité,
    $$P(X) = P(P(X|Y)).$$
14. **Convergence en loi** :
    Si $(X_n)_{n\in\mathbb{N}}$ converge en loi vers X et si $\max(X_n) \xrightarrow[n\to\infty]{} \max(X)$, alors
    $$P(X_n) \xrightarrow[n\to\infty]{} P(X).$$
15. **Stabilité par mélange** :
    Soient X', X_1 et X_2 des risques et $p \in [0,1]$. Si $P(X_1) = P(X_2)$, alors :
    $$P(pF_{X_1} + (1-p)F_{X'}) = P(pF_{X_2} + (1-p)F_{X'}).$$

Qualification des principes de prime

Nous classifions les principes de prime selon plusieurs groupes généraux combinant les propriétés introduites précédemment.

1. Un principe de calcul de prime est dit monétaire s'il satisfait les propriétés de translation [5] et de monotonie [9].
2. Un principe de calcul de prime est dit convexe (ou faiblement cohérent) s'il satisfait les propriétés de translation [5], monotonie [9] et de convexité [12].
3. Un principe de calcul de prime est dit cohérent s'il satisfait les propriétés de translation [5], sous-additivité [6], homogénéité positive [8] et de monotonie [9].

Propositions et preuves

Proposition 1. ***Invariance de type 1 [10]***

1. *$X \preceq_{DS1} Y$ si et seulement s'il existe des variables aléatoires $X' \overset{d}{=} X$ et $Y' \overset{d}{=} Y$ telles que $P(X' \leq Y') = 1$.*
   (Remarquons ici le lien entre la dominance stochastique à l'ordre 1 et la propriété de monotonie précédemment introduite.)
2. *$X \preceq_{DS1} Y$ si et seulement si :*
   $$\mathbb{E}[u(-Y)] \leq \mathbb{E}[u(-X)]$$
   pour toute fonction u croissante.
3. *Si X et Y admettent des densités et s'il existe une constance c telle que :*
   $$f_X(d) \geq f_Y(d) \quad \text{pour } d \in]-\infty, c[$$
   $$f_X(d) \leq f_Y(d) \quad \text{pour } d \in]c, \infty[$$
   alors $X \preceq_{DS1} Y$.

Preuve de la proposition 1.

1. La preuve sera uniquement donnée dans le cas $F_Y^{-1}(F_Y(d)) = d$ et F_X est continue. Dans ce cas, $F_X(X) \overset{\text{d}}{=} U$ où U est une variable aléatoire de loi uniforme sur [0,1]. De plus, $F_Y^{-1}(U) \overset{\text{d}}{=} Y$ car $F_Y(Y) \overset{\text{d}}{=} U$ et $F_Y^{-1}(F_Y(Y)) = Y$ par hypothèse. Ainsi, en posant $X' = X$ et $Y' = F_Y^{-1}(F_X(X))$, nécessairement

$$X' \leq Y' \text{ p.s.}$$

2. Posons $v(x) = -u(-x)$, v est alors une fonction croissante. Ainsi,

$$\mathbb{E}[u(-X)] \geq \mathbb{E}[u(-Y)]$$

est équivalent à la condition

$$\mathbb{E}[v(X)] \leq \mathbb{E}[v(Y)]$$

pour toute fonction v non décroissante.
Sens $\Leftarrow$:
Ce sens est trivial car $1 - F_X(t) = \bar{F}_X(t) = \mathbb{E}[\mathbb{1}_{\{X>t\}}]$ et la fonction $x \to \mathbb{1}_{\{x>t\}}$ est non-décroissante pour tout t.
Sens $\Rightarrow$:
En utilisant la proposition 1 :

$$\mathbb{E}[v(X)] = \mathbb{E}[v(X')] \leq \mathbb{E}[v(Y')] = \mathbb{E}[v(Y)].$$

3. Pour $x < c$ nous avons

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\ du \geq \int_{-\infty}^{x} f_Y(u)\ du = F_Y(x)$$

Pour $x > c$:

$$F_X(x) = 1 - \int_{x}^{\infty} f_X(u)\ du \geq 1 - \int_{x}^{\infty} f_Y(u)\ du = F_Y(x)$$

□

□

Proposition 2. ***Invariance de type 2 [11]***

1. $X \preceq_{DS2} Y$ si et seulement si :

$$\mathbb{E}[u(-X)] \geq \mathbb{E}[u(-Y)]$$

pour toute fonction u croissante et concave.

2. $X \preceq_{DS2} Y$ si et seulement s'il existe une variable aléatoire D telle que :

$$X + D \overset{d}{=} Y \text{ et } \mathbb{E}[D|X] \geq 0 \text{ p.s.}$$

3. Si $\mathbb{E}[X] \leq \mathbb{E}[Y]$ et s'il existe c telle que :

$$F_X(d) \leq F_Y(d) \quad \text{pour } d \in]-\infty, c]$$

$$F_X(d) \geq F_Y(d) \quad \text{pour } d \in [c, \infty[$$

alors $X \preceq_{DS2} Y$.

4. Si $X \preceq_{DS1} Y$, alors $X \preceq_{DS2} Y$.

Preuve de la proposition 2.

1. Posons tout d'abord $v(x) = -u(-x)$. Ainsi les deux inégalités suivantes sont équivalentes :

$$\mathbb{E}[u(-X)] \geq \mathbb{E}[u(-Y)]$$

est équivalent à :

$$\mathbb{E}[v(X)] \leq \mathbb{E}[v(Y)]$$

pour toute fonction non-décroissante et convexe v.
Sens $\Leftarrow$:
Trivial car la fonction $x \to (x-t)_+$ est convexe pour tout $t \in \mathbb{R}$.
Sens $\Rightarrow$:
Notons que toute fonction convexe v est une limite d'une suite croissante de fonctions de la forme :

$$v_n(x) = \alpha_1 + \alpha_2 x + \sum_{j=0}^{n} \beta_j^{(n)} \left(x - t_j^{(n)}\right)_+$$

avec $\beta_j^{(n)} \geq 0$. Par conséquent, nous pouvons écrire :

$$\begin{aligned}
\mathbb{E}[v_n(X)] &= \mathbb{E}\left[\alpha_1 + \alpha_2 X + \sum_{j=0}^{n} \beta_j^{(n)} \left(X - t_j^{(n)}\right)_+\right] \\
&= \alpha_1 + \alpha_2 \mathbb{E}[X] + \sum_{j=0}^{n} \beta_j^{(n)} \mathbb{E}\left[\left(X - t_j^{(n)}\right)_+\right] \\
&\leq \alpha_1 + \alpha_2 \mathbb{E}[Y] + \sum_{j=0}^{n} \beta_j^{(n)} \mathbb{E}\left[\left(Y - t_j^{(n)}\right)_+\right] = \mathbb{E}[v_n(Y)]
\end{aligned}$$

pour tout n. Par passage à la limite puis en utilisant le théorème de convergence monotone nous obtenons :

$$\mathbb{E}[v(X)] \leq \mathbb{E}[v(Y)].$$

2. Nous démontrerons uniquement le sens $\Leftarrow$. En utilisant l'inégalité de Jensen (conditionnelle) :

$$\begin{aligned}
\mathbb{E}[(Y-d)_+] &= \mathbb{E}[(X+D-d)_+] \\
&= \mathbb{E}_X[\mathbb{E}_{D|X}[(X+D-d)_+|X]] \\
&\geq \mathbb{E}_X[(X+\mathbb{E}_{D|X}[D|X]-d)_+|X] \\
&\geq \mathbb{E}[(X-d)_+].
\end{aligned}$$

3. Remarquons que

$$\begin{aligned}
\mathbb{E}[(X-d)_+] &= \int_d^\infty (x-d) dF_X(x) \\
&= -[(x-d)(1-F_X(x))]_d^\infty + \int_d^\infty \bar{F}_X(x) dx \\
&= \int_d^\infty \bar{F}_X(x) dx
\end{aligned}$$

Ainsi, en dérivant par rapport à d :

$$\frac{\partial \mathbb{E}[(X-d)_+]}{\partial d} = -(1-F_X(d)) = -\bar{F}_X(x).$$

De plus, $\lim_{d\to-\infty} \mathbb{E}[(X-d)_+] = 0$ et, puisque $\mathbb{E}[(X-d)_+] + d = \mathbb{E}[\max(X,d)]$, nous avons :

$$\lim_{d\to-\infty} (\mathbb{E}[(X-d)_+] + d) = \mathbb{E}[X].$$

Considérons maintenant la fonction $\phi(d) = \pi_Y(d) - \pi_X(d)$ où $\pi_X(d) = \mathbb{E}[(X-d)_+] + d$. Nous avons $\lim_{d\to-\infty} \phi(d) = \mathbb{E}[Y] - \mathbb{E}[X] \geq 0$, $\lim_{d\to\infty} \phi(d) = 0$ et $\phi'(d) = F_Y(d) - F_X(d)$, ce qui permet de conclure.

4. $X \preceq_{DS1} Y$ si et seulement si

$$\mathbb{E}[u(-X)] \geq \mathbb{E}[u(-Y)]$$

pour toute fonction non-décroissante u. Ainsi, l'inégalité est vraie pour toute fonction non-décroissante et concave u et $X \preceq_{DS2} Y$.

□