IRM | institut du risk management BY Institut.

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

Par : Matthieu de CHEVIGNY

Titre : Application de la théorie de la crédibilité hiérarchique et des valeurs extrêmes à la tarification du risque incendie des risques industriels.
Confidentialité : ☐ NON ☒ OUI (Durée : ☐ 1an ☒ 2 ans)
Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de l'Institut des actuaires :

_____

_____

_____

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

_____

_____

_____

_____

_____

_____

_____

_____

Secrétariat :

Bibliothèque :

Entreprise :

Nom : AXA France IARD

Signature et Cachet :

AXA France I.A.R.D.
Société Anonyme au Capital de 214 799 030 €
Entreprise régie par le Code des Assurances
Siège social : 313, Terrasses de l'Arche
92727 NANTERRE CEDEX
722 057 460 RCS Nanterre

Directeur de mémoire en entreprise :

Nom : Mayeul COTHENET

Signature :

Invité :

Nom : _____

Signature :

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

Signature du responsable entreprise

Signature(s) du candidat(s)

Remerciements

Je remercie Véronique MARPILLAT pour son accueil au sein de l'équipe APDS AXA Entreprise, pour ses précieux conseils et son soutien. Être accueilli à un poste avec comme objectif de réaliser un mémoire est une vraie marque de confiance professionnelle.

Je remercie Mayeul COTHENET pour son soutien, sa patience et son aide au quotidien pour cette étude. Relire une telle étude pendant son temps personnel est un grand soutien.

Merci également à l'équipe DAB, sans qui ces travaux n'auraient pu aboutir à temps.

Je remercie également tous les membres de l'APDS Entreprise pour leur soutien, leurs conseils et leurs relectures. Ce service est un vrai puit de science actuarielle.

Je remercie aussi Arthur JEANNIN, Fabien KUBLER et Julien TREZZA avec qui le CEA fut motivé par une ambition de réussite commune.

Merci à ma maman d'avoir relu. Merci à Edgar, Sybille et Astrid pour leur confiance.

Je remercie ma tutrice académique, Anne-Charlotte BONGARD, qui a su rapidement prendre connaissance de l'étude et orienté mes travaux et mes explications.

Résumé

L'objectif actuariel de ce mémoire est l'amélioration de la méthode de tarification du risque incendie. Ce mémoire part de la méthode de la crédibilité hiérarchique et la complète de deux étapes actuarielles :

- La détermination des seuils de sinistralité pour l'écrêtement et la mutualisation des sinistres extrêmes via la théorie des valeurs extrêmes.
- L'utilisation d'ANOVA et de Machine Learning pour la détermination des niveaux hiérarchique

La refonte tarifaire s'appuie sur les différents travaux réalisés dans l'actuariat sur le risque incendie avec l'apport de la crédibilité par « TARIFICATION DE L'INCENDIE DES RISQUES INDUSTRIELS FRANCAIS PAR LA METHODE DE LA CREDIBILITE » (Cohen Dupin et Levy, 1985) et par l'apport du mémoire « TARIFICATION DU RISQUE INDUSTRIEL » (DOUVILLE, 2004) avec l'utilisation de la crédibilité hiérarchique.

Dans une première partie, l'étude se concentre sur l'introduction du vocabulaire et de la branche Risques industriels, les travaux importants réalisés par France Assureurs ayant conduit à un mode de tarification par rubrique d'entreprise (TRE) base de la tarification d'AXA France.

La seconde partie aborde l'apport de la théorie des valeurs extrêmes et d'ANOVA et du Machine Learning sur les modèles de la théorie de la crédibilité hiérarchique permettant la détermination du niveau de prime pure pour le risque incendie.

Les résultats de l'étude montrent que l'apport de la théorie des valeurs extrêmes et de l'ANOVA permet de se détacher des avis d'experts en apportant à l'actuaire le recul nécessaire à une tarification technique. Le choix des paramètres de tarification gagne en pertinence actuarielle.

L'objectif économique de cette étude est la refonte de la tarification de la garantie incendie des risques industriels d'AXA France.

Mot clés : crédibilité, Bühlmann, Jewell, Hiérarchique, Vylder, RandomForest, GBM, Boosting, Incendie, ANOVA.

Abstract

The actuarial objective of this thesis is the improvement of the fire risk pricing methodology. This thesis starts with the hierarchical credibility method and completes it with two actuarial steps:

- The choice of loss thresholds to capp and pool extreme claims thanks to the theory of extreme values.
- The use of ANOVA and Machine Learning to choose headings of companies' groupings.

The pricing overhaul is based on various works dealing with the fire risk within the actuarial field with the contribution of two thesis :

- "TARIFICATION DE L'INCENDIE DES RISQUES INDUSTRIELS FRANCAIS PAR LA METHODE DE LA CREDIBILITE" (Levy, 1985)
- "TARIFICATION OF INDUSTRIAL RISK" (DOUVILLE, 2004).

To bring its contribution,

In the first part, the study deals with the vocabulary introduction's and the Industrial Risks' branch Major works carried out by France Assureurs lead to a pricing model by heading of company (TRE) baseof AXA France's pricing

The second art tells the contribution of extreme values' theory and ANOVA & Machine Learning on the theory of hierarchical credibility models' allowing the evaluation of pure premium forfire risk's level.

The economic objective of this thesis is an Industrial Risks branch's overhaul pricing of the insurance contracts' fire risk guarantee to update the pricing of AXA France ENTREPRISE.

Keywords: Credit, Bühlmann, Jewell, Hierarchical, Vylder, RandomForest, GBM, Boosting, Fire, ANOVA.

INTRODUCTION

Objectif

L'objectif actuariel de ce mémoire est l'amélioration de la méthode de tarification du risque incendie. Ce mémoire part de la méthode de la crédibilité hiérarchique et la complète de deux étapes actuarielles :

- La détermination des seuils de sinistralité pour l'écrêtement et la mutualisation des sinistres extrêmes via la théorie des valeurs extrêmes.
- L'utilisation d'ANOVA et de Machine Learning pour la détermination des niveaux hiérarchique

L'ambition économique de cette étude est une refonte de la tarification du risque incendie de la branche Risques Industriels d'AXA France.

Pour des raisons de confidentialité, les résultats chiffrés présentés ont été modifiés tout en conservant les ordres de grandeur pour maintenir le sens des conclusions.

L'assurance risque industriel

L'assurance des risques industriels est spécifique aux entreprises du secteur de l'industrie, ce qui couvre de nombreux secteurs d'activité. Elle couvre notamment les risques d'incendies, de dégâts des eaux, de vol ou tentative de vol, de bris de glace, de vandalisme, de catastrophes naturelles.

Le risque industriel est un risque sensible. Il existe deux typologies de risques dans les risques sensibles. Les risques émergents, dont la modélisation est compliquée, par leur absence d'historique statistique et donc une donnée insuffisante, l'assurance cyber est un risque sensible rentrant dans cette définition. La seconde typologie de risque sensible regroupe les risques connus dont la sinistralité est importante en charge et faible en fréquence.

Le risque industriel entre dans la seconde topologie. De ce fait la rentabilité de cette branche peut être rapidement dégradée. L'assureur dans le cadre du risque industriel doit avoir la possibilité de mutualiser au maximum son risque pour pallier les évènements extrêmes.

Tarification du risque incendie

La tarification du risque incendie nécessite une grande rigueur par le fait que la tarification du risque industriel d'une entreprise n'est pas réalisée de manière globale pour toutes les garanties comme dans un contrat standardisé. Dans un premier temps, le travail de l'assureur consiste à déterminer, pour une entreprise, le taux correspondant à la garantie de base incendie, ce taux appliqué à l'ensemble des capitaux assurés contre le risque d'incendie donne la prime pure de l'entreprise. Cette garantie couvre uniquement les évènement incendie, explosion, chute directe de la foudre et fuite accidentelle de sprinklers. Dans un second temps, le taux calculé sert de base aux autres garanties, par exemple la perte d'exploitation est un coefficient appliqué au taux incendie. Les autres garanties, dites garanties annexes, seront majoritairement extrapolés à partir du taux de base incendie.

La tarification du risque incendie est un élément important dans la maitrise du risque incendie portée par l'assureur, cependant le suivi du souscripteur par l'entreprise d'assurance est un autre pilier de cette maitrise. En effet, les efforts de prévention et de protection contre l'incendie réalisés par le souscripteur sont des éléments clés de l'atténuation du risque. Ces actions sont des éléments qui permettent de personnaliser le taux de base incendie au profil de l'assuré à travers des rabais ou de majoration.

En effet, la tarification du risque incendie est basée sur le regroupement des entreprises en rubrique. Une rubrique est une classe d'activité industrielle, par exemple « Fabrication et travail du verre ». Toute

rubrique a son propre taux de base incendie à appliquer aux capitaux incendie de l'entreprise pour définir la prime de l'entreprise, avant les actions d'ajustement au profil de prévention et protection.

Une petite erreur de calcul du taux de base incendie impacte l'ensemble de la tarification du risque industriel de l'entreprise.

Afin de développer au maximum le marché des Risques Industriels, France Assureurs a mis disposition à l'ensemble du marché de l'assurance une proposition tarifaire du risque incendie pour le risque industriel.

Cette tarification est basée principalement sur les travaux suivants :

- « Tarification de l'incendie des risques industriels français par la méthode de la crédibilité » (Cohen Dupin et Levy, 1985) : Cette étude propose de perfectionner la méthode d'écrêtement et mutualisation des sinistres, qui étaient la méthode historique de tarification du risque incendie dans le risque industriel en appliquant la théorie de la crédibilité développée récemment.
- « TARIFICATION DU RISQUE INDUSTRIEL » (DOUVILLE, 2004) : Cette étude propose de perfectionner l'étude précédente en s'appuyant sur les fascicules regroupant les rubriques d'activités et ainsi utiliser la crédibilité hiérarchique.

L'objectif de France Assureurs étant de favoriser l'entrée sur ce secteur des sociétés d'assurance sans expérience dans le secteur.

En 2012, AXA a travaillé sur une mise à jour de l'ensemble des taux rubriques d'activités d'entreprises :

- Une segmentation plus précise des rubriques d'entreprise et donc des taux (Une rubrique d'entreprise (TRE) selon la segmentation de France Assureurs (TE208) pouvant être divisé en plusieurs TRE selon la segmentation AXA (T208A, T208B))
- Une actualisation des taux de rubrique d'entreprise (TRE) à partir des études France Assureurs et du portefeuille AXA.

L'évaluation de ces taux de rubriques par AXA France n'ayant pas été revue depuis, l'objectif économique de ce mémoire est donc la tarification de la garantie incendie des risques industriels à partir de la méthode actuarielle proposée dans l'étude.

Problème

L'objectif actuariel de cette étude est l'amélioration de la méthode de tarification du risque incendie. Ce mémoire part de la méthode de la crédibilité hiérarchique mais propose l'ajout de deux étapes actuarielles :

- L'utilisation d'ANOVA et de Machine Learning pour la détermination des regroupements des rubriques d'entreprises.
  Le regroupement actuel par fascicule, c'est-à-dire par groupe de catégories d'activité d'entreprise similaires (ex : Extraction de minerais et minéraux divers, de combustibles solides. Métallurgie) montrant peu à peu ses limites au vu des types d'activités émergent (DATA-Center…) sortant du cadre proposé, ainsi l'utilisation d'une méthode actuarielle accompagnera l'actuaire dans le choix de ces niveaux hiérarchiques, éléments critique de la crédibilité hiérarchique.
- La détermination des seuils de sinistralité pour l'écrêtement et la mutualisation des sinistres extrêmes via la théorie des valeurs extrêmes. Ainsi permettre à l'actuaire de se détacher des dires d'experts dans un milieu qui évolue rapidement et marqué actuellement par une forte inflation.

En 2021, à la suite d'une étude sur la sinistralité incendie de son portefeuille, AXA France est arrivé au constat que les taux incendie précédemment calculés étaient décorrélés de la sinistralité incendie. En effet, au cours du temps le taux incendie ayant servi de variable d'ajustement de la rentabilité des rubriques, quel que soit l'origine de sa dégradation (PE, Grêle…).

L’objectif second de l’étude est donc une remise à plat des taux incendies. Cette étude est les prémices d’une refonte complète totale de la tarification du risque industriel d’AXA France.

1 CONTEXTE, OBJECTIFS ET PERIMETRE DE L'ETUDE

1.1 LE GROUPE AXA ET AXA FRANCE

Selon le livre « L'histoire d'AXA » (Desaegher), la société AXA a été fondée en Normandie, à partir de la mutuelle de Rouen, qui était une petite société d'assurance française, créée en 1817, issue de la Compagnie mutuelle contre l'incendie, Axa est devenue depuis l'un des leaders internationaux de l'assurance et de la gestion d'actifs. La création officielle d'AXA voit le jour en 1986 sous la direction de Claude Bébéar. A partir de cette date, le Groupe enchaine les fusions et surtout les acquisitions, dont la principale est l'Union des Assurances de Paris (l'UAP) en 1996, qui lui permet de devenir la plus importante entreprise française par le chiffre d'affaires et le numéro un mondial de l'assurance. A l'aube des années 2000, AXA est déjà un leader mondial, mais le groupe continue de s'agrandir, notamment avec sa dernière acquisition datant de 2018, celle du groupe XL, l'un des principaux acteurs de l'assurance des dommages des entreprises et de la réassurance.

En 2022, AXA compte un peu moins de 100 millions de clients et 150 000 collaborateurs. L'une des forces du groupe est son offre : assurance dommages, assurance vie, épargne, retraite et santé, ainsi que la gestion d'actifs.

Le chiffre d'affaires d'Axa est estimé à 100 Mds d'euros en 2021. Le groupe poursuit sa croissance sur des segments prioritaires et renforce ses engagements RSE avec notamment la lutte contre le réchauffement climatique.

Au niveau du marché français, Axa France est le 2ème assureur en France devancé par Crédit Agricole Assurance, avec un chiffre d'affaires de 28,3 Mds €, soit 28% du CA du groupe. Ce classement peut se décomposer avec une 2ème place en assurance dommages soit 12,9% des parts de marché, et une 3ème place en assurance vie, épargne et retraite soit 8,4% des parts de marché.

L'entité d'Axa France IARD Entreprises génère près de 3 Mds € de chiffre d'affaires pour l'année 2021 et propose une variété de produits couvrant les risques encourus par différentes entreprises (IARD est l'acronyme de : « incendies, accidents et risques divers »).

Ces produits sont élaborés par la Direction Actuariat et Pilotage Entreprises (DAPE). Au sein de cette direction se trouve la partie DAB (Dommage aux biens) contenant le RI (Risque industriel) contexte de ce travail de recherche.

1.1 PRESENTATION DE L'ASSURANCE DE DOMMAGES

Le code des assurances fait la distinction entre les différents risques (Vie/Non Vie), l'article L. 310-1 du Code des assurances distingue trois types d'assurance :

- assurance vie : assurance dont l'aléa dépend de la durée de la vie humaine (Ex: temporaire décès.)
- assurance de dommages corporels (Ex: Incapacité-Invalidité)
- assurance d'autres risques (Ex: automobile)

Le principe de spécialisation (L. 321-1) précise que l'assurance vie et l'assurance des « autres risques » sont incompatibles. Il est en revanche possible de pratiquer le dommage corporel et la vie en société mixte.

L'assurance de dommages correspond au 3ème point de l'art L310-1 (terme équivalent : assurance IARD pour Incendie, Accidents et Risques Divers)

L'assurance IARD comprend :

- les assurances de biens, qui couvrent un risque relatif à un élément d'actif patrimonial,
- les assurances de responsabilité, qui couvrent les dettes liées à l'obligation de réparer les dommages causés à autrui, y compris éventuellement les dommages corporels.

Les principales assurances de IARD sont :

- l'assurance des biens particuliers (contrats MultiRisques Habitation MRH),
- l'assurance des biens professionnels (Risques industriels Entreprise, agriculteurs, commerçants, artisans et prestataires de services, collectivités locales…),
- l'assurance de construction,
- l'assurance automobile,
- l'assurance de transports (assurances ferroviaire, maritime, fluviale, aérienne, spatiale, marchandises transportées),
- l'assurance de responsabilité civile,
- l'assurance crédit,
- l'assurance de protection juridique.

1.2 POSITIONNEMENT D'AXA FRANCE SUR LE DOMMAGE AUX BIENS

La Fédération Française de l'Assurance (France Assureurs) publie chaque année une synthèse d'un questionnaire, rempli par une majorité de sociétés d'assurance, sur la production de contrats et sinistres de la branche Dommages Aux Biens (DAB).

Cette synthèse permet de comprendre le positionnement d'AXA sur le marché du DAB.

Pour l'année 2021, voici le positionnement d'AXA selon 4 critères :

- Le nombre de contrats
- Le montant de primes
- Le montant de sinistre
- Le ratio S/P. Ce ratio est le montant de sinistre sur le montant de prime, il est un des indicateurs économiques les plus important sur le marché de l'assurance.

Figure 1 - positionnement d'AXA sur le DAB

Le présent mémoire se concentrant sur le risque industriel, voici le détail d'AXA par rapport au marché du dommage aux biens en France découpé par sous marché :

- Risque industriel
- Garage
- Collectivités locales
- Risque technique
- Immeubles
- ACPS (Artisans, Commerçants et Professions de Services)

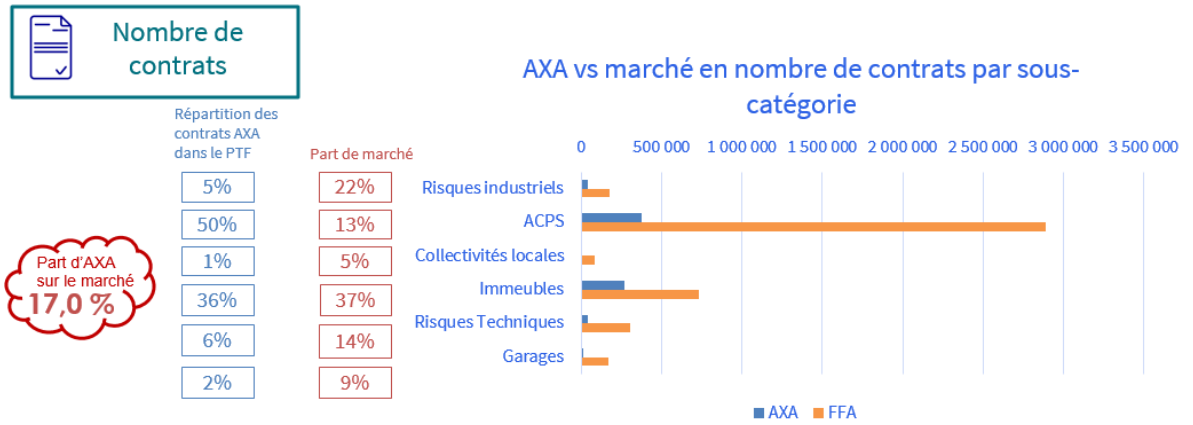


Figure 2 - Positionnement d'AXA par sous-catégorie (Nombre de contrat)

Axa France possède 22% des contrats des risques industriels.

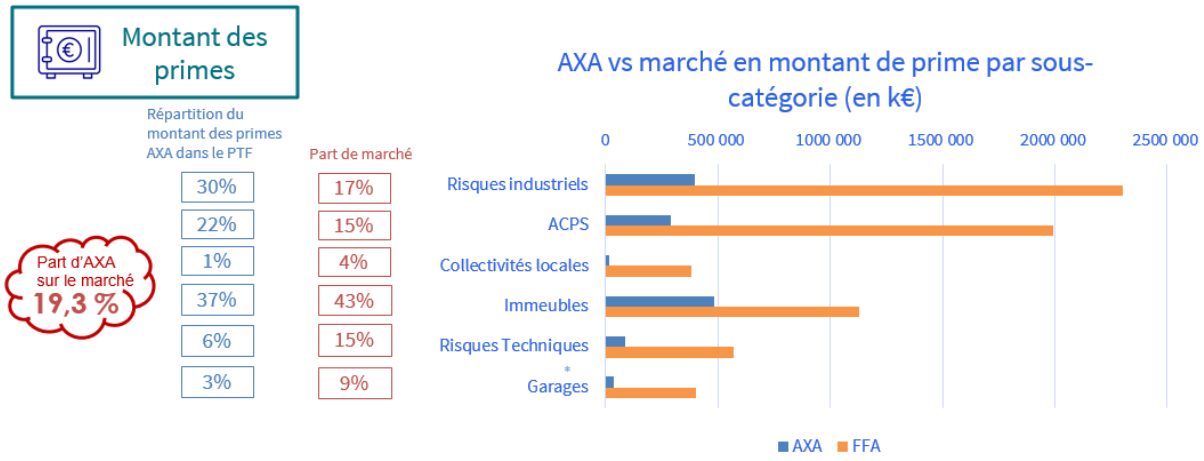


Figure 3 - Positionnement d'AXA par sous-catégorie (Montant de primes)

Axa France représente 17% des primes des risques industriels du marché.

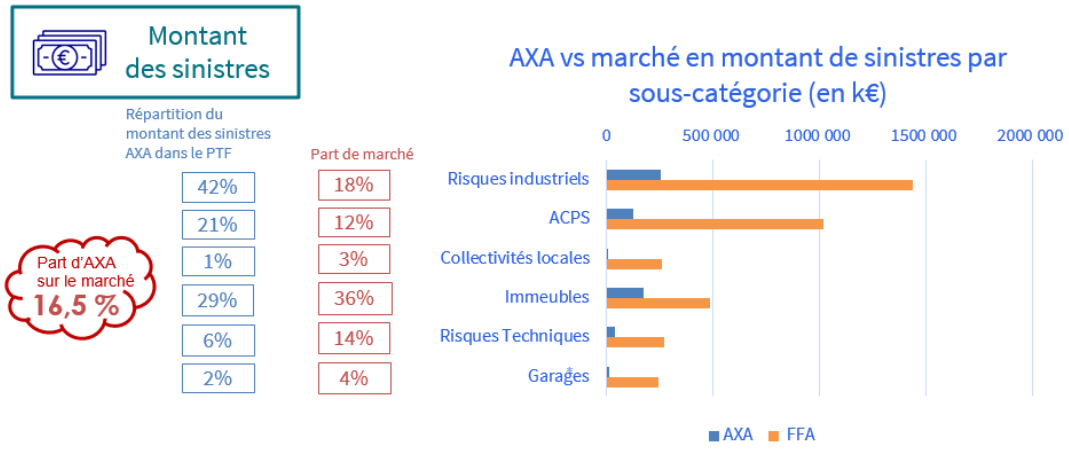


Figure 4 - Positionnement d'AXA par sous-catégorie (Montant des sinistres)

18 % des sinistres des risques industriels sont dans le portefeuille d'AXA France.

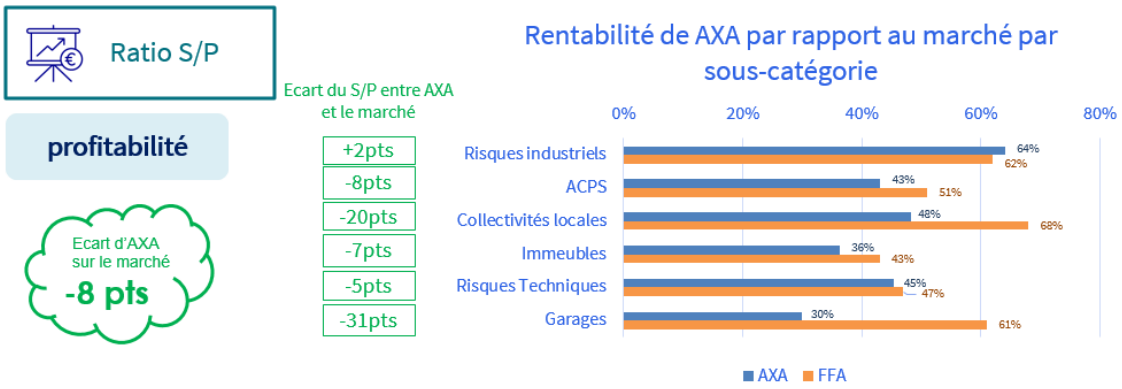


Figure 5 - Positionnement d'AXA par sous-catégorie (Sinistres sur primes)

Le ratio S/P représente la part de la prime allant directement dans le règlement de sinistre. AXA a un ratio supérieur à l'ensemble du marché. Hors chargement, AXA a donc une rentabilité inferieure que le marché. L'étude tente de comprendre les incohérences tarifaires pouvant dégrader cette rentabilité.

1.3 HISTOIRE ET PRINCIPES DE L'ASSURANCE DES RISQUES INDUSTRIELS (RI)

L'assurance des Risques Industriels fait partie du dommage aux biens de la partie IARD Entreprise d'AXA.

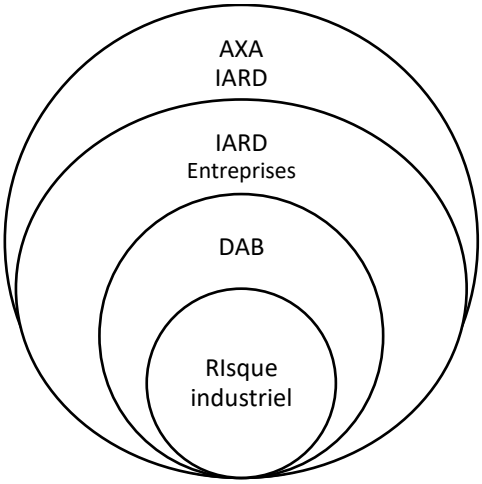


Figure 6 - Le risque industriel chez AXA France

L'assurance des Risques Industriels permet aux entreprises d'assurer les bâtiments, le mobilier, le matériel et même les biens pour se prémunir contre les dommages aux personnes comme aux biens (incendie, explosion, dégâts électriques, dégâts des eaux, vol, vandalisme, attentat, évènements météorologique (tempêtes, grêle), neige, inondations, glissements de terrain, etc.).

Le principal risque des entreprises dans le risque industriel est le risque portant sur l'incendie.

Historiquement, le principe de l'assurance des Risques Industriels est apparu à la naissance de l'assurance des dommages aux biens à la fin du XVII^{ème} siècle, après le grand incendie de Londres de 1666 qui a causé la destruction de plus de 13 000 bâtiments.

C'est l'organisation du « Fire Office » en 1667, à l'initiative de Monsieur Nicholas Barbon, qui engendrera la naissance de plusieurs compagnies d'assurances de dommages dont la « Hand in Hand » qui est la compagnie la plus ancienne du monde.

En France, la première société mutuelle concernant l'incendie est créée en 1717, appelée aussi bureau des incendies. Cette société est gérée par l'Église catholique pour procurer une aide financière en cas de dommages après un incendie.

Par la suite, de nombreuses compagnies d'assurance de dommages ont vu le jour et c'est en 1936 qu'est créée la Fédération Française des Sociétés d'Assurance (FFSA). En juillet 2016, la création de la Fédération française de l'assurance (France Assureurs) marque la réunion au sein d'une seule organisation de la Fédération française des sociétés d'assurances (FFSA) et du Groupement des entreprises mutuelles d'assurance (GEMA).

Le risque industriel se base principalement sur le risque incendie dont l'indemnisation est basée sur la notion de « capitaux incendie » c'est-à-dire la valeur potentielle de la perte de l'entreprise en cas d'incendie. Cette valeur ne prend pas en compte la perte d'exploitation mais bien celle des capitaux matériels de l'entreprise. Le but du souscripteur dans le cas du RI est d'évaluer à partir des différents capitaux incendie d'un contrat le sinistre maximum possible (SMP) qui dépend, en plus du capital Incendie, de l'éloignement des différents sites ou bâtiments ainsi que des précautions prises pour empêcher la propagation de l'incendie (portes coupe-feu, etc…). Ce travail d'évaluation est l'un des principaux piliers de la tarification du risque.

Pour chaque type de bien, il existe plusieurs façons de l'assurer ou de définir la valeur déclarée :

- Les bâtiments peuvent en effet être assurés en valeur de reconstruction selon deux formules : vétusté déduite ou en valeur à neuf. La garantie peut être complétée par la prise en compte de certains frais de remise en état comme les frais de démolition et de déblai.
- Le mobilier et le matériel sont le plus souvent assurés en valeur de remplacement à neuf, mais il est possible de les assurer en valeur d'usage, c'est-à-dire en valeur de remplacement à neuf, vétusté déduite.
- En ce qui concerne les marchandises, il faut distinguer trois catégories. La première regroupe les matières premières et les emballages qui sont assurés au prix d'achat incluant les éventuels frais de transport et de manutention. La deuxième concerne les objets fabriqués ou en cours de fabrication qui sont assurés au coût de production (prix d'achat des matières premières et produits utilisés et frais de fabrication). Enfin, la troisième est constituée des marchandises vendues mais non encore livrées qui sont assurées au prix de vente, déduction faite des frais de livraison non engagés.

1.4 LES OFFRES DE LA BRANCHE RI D'AXA FRANCE

Au sein d'AXA, la branche RI est composée de 2 produits avec deux segmentations différentes, le bas de segment et le haut de segment.
La frontière entre les différents produits est définie tout d'abord par le type d'activité de l'entreprise :

- Activités commerciales et artisanales.
- Commerces de gros et petites activités industrielles.
- Activités industrielles Grandes activités commerciales).

Puis par des critères de surface, chiffre d'affaires et de contenu incendie.
Ci-dessous un schéma et un tableau qui illustrent la distinction des 2 produits avec les différents critères :

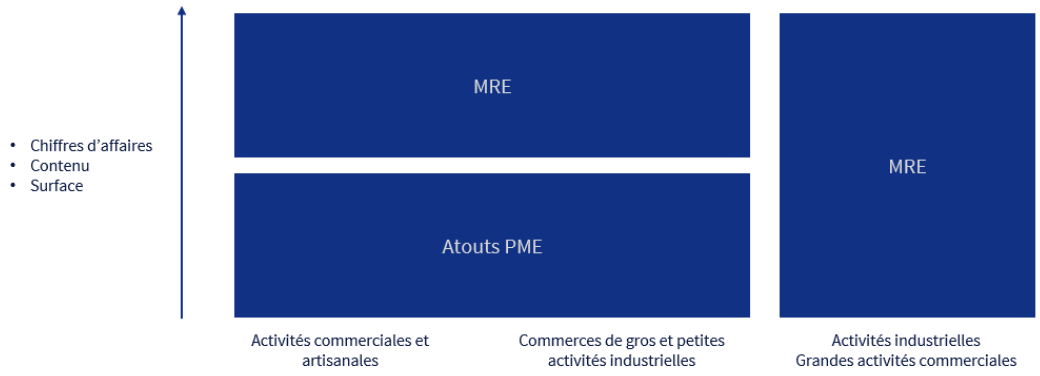


Figure 7 - Les produits AXA France Entreprise

1.4.1 Bas de segment

Le bas de segment de la branche RI, comprend le produit Atouts PME, qui regroupe les anciennes gammes de produit « Multirisques Professionnels Particuliers (MRP PP) », « Multirisques Entreprises (MRP EN) » et « Multirisques Petites et Moyennes Entreprises (MPME) »

Quelques exemples d'activité pour ces deux produits : salon de coiffure, auto-école, vente de produits de beauté, station de lavage automobile en libre-service, tailleur, etc. Dans le bas de segment, l'identification de l'entreprise se fait par code d'activité.

Pour distinguer les différentes activités pour ces trois produits, il existe un code activité sur sept caractères dont les quatre premiers caractères permettent de déterminer le groupe d'activité. Les trois autres caractères permettent à AXA de faire une différenciation des sous-activités à une maille plus fine.

1.4.2 Haut de segment

Le haut de segment de la branche RI se caractérise par le produit Multirisques Entreprises.

Multirisques Entreprises (MRE), commercialisées depuis 2004, avec deux approches :

- **Multirisques Entreprises simplifiées (MRES)**, où le SMP est inférieur à 20 M€ de capitaux assurés et correspond à un monosite
- **Multirisques Entreprises complexes (MREC)**, où le SMP est supérieur ou égal à 20 M€ de capitaux assurés

Dans le haut de segment, l'identification de l'entreprise se fait par rubrique d'activité.

Quelques exemples de rubrique d'activité (TRE) pour ce produit : hôtel 5 étoiles, industrie pharmaceutique, travail des métaux, piscine, gymnase, patinoire, laboratoire de recherches, …

Ci-dessous un schéma descriptif détaillé du périmètre de la branche RI avec des exemples d'activités par segment (portefeuille 2010-2018).

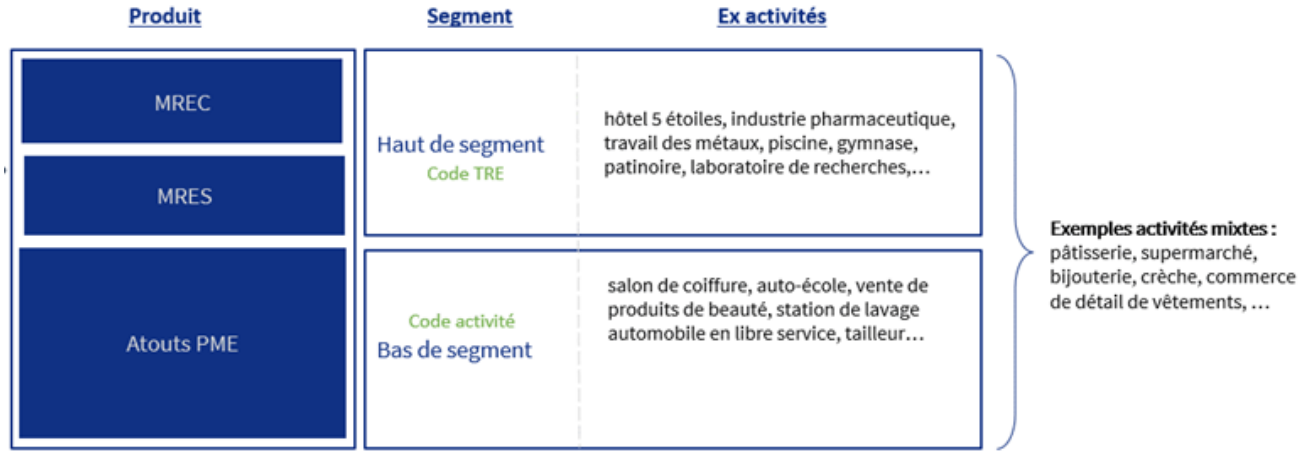


Figure 8 - Exemples d'activités par segment

Le périmètre de l'étude se concentre sur la tarification de la MREC. Les produits MRES et Atouts PME étant plus standardisés.

1.5 TARIFICATION DE L'ASSURANCE DES RISQUES INDUSTRIELS ET EVOLUTIONS

Afin de faciliter aux lecteurs la compréhension de l'environnement du risque entreprise, il est important de comprendre la notion de rubrique d'activité entreprise (TRE), une rubrique est un ensemble homogène d'entreprises, par exemple le TE928 définit les « Hangars pour aéronefs ; aéronefs au sol », le TE778 définit les entreprises d'« Abattage de volailles, lapins, gibiers ».

Historiquement appelé le « tarif rouge », les premiers travaux en « Incendie Industriel » furent un calcul par l'écrêtement-mutualisation destiné à réduire la volatilité induite par les sinistres les plus importants. Cependant cette méthode présentait un biais car elle limitait la mutualisation sur les différents exercices d'une rubrique d'activité entreprise (TRE) ; ainsi une rubrique d'activité entreprise (TRE) avec des capitaux faibles et donc de faibles encaissements était toujours plus exposée qu'une rubrique à capitaux important.

La problématique de la tarification des risques industriels est premièrement abordée d'un point de vue actuariel par les travaux « Tarification de l'incendie des risques industriels français par la méthode de la crédibilité » (Cohen Dupin et Levy, 1985). L'approche que les auteurs proposent est une mutualisation des rubriques par l'utilisation de la crédibilité en complément de la méthode d'écrêtement de la sinistralité.

Cette double approche mutualisation et écrêtement des sinistres et utilisation de la crédibilité est une base importante de la tarification du risque « incendie ».

Pour rappel la théorie de la crédibilité est une théorie mathématique développée à la fin des années 1960 par un actuaire suisse, Hans Bühlmann (Hans Bühlmann, 2005).

La théorie de la crédibilité se base sur l'hypothèse selon laquelle les données observées sont censées refléter une réalité sous-jacente, mais qu'il y a un certain degré d'incertitude, elle cherche donc à charger assez de primes pour couvrir les sinistres tout en répartissant équitablement la prime entre les assurés.

De fait, la théorie de la crédibilité s'applique particulièrement bien aux risques d'entreprise, qui se caractérisent par un relativement faible nombre de contrats et de sinistres avec une forte intensité, ce qui empêche l'utilisation fiable des modèles linéaires généralisés comme en risques de masse.

L'importance de l'étude présentée ci-dessus permet à la FFSA une refonte du tarif RI s'appuyant sur la théorie de la crédibilité. En découle en 1990, le premier traité d'assurance incendie des risques d'entreprises (TRE), un tarif analytique différencié par activité, l'appelation « Tarif Rouge » est remplacée par « Traité du risque entreprise » (Le « Tarif rouge » est dépoussiéré pour encourager la prévention, 2004).

En 2004, le mémoire « TARIFICATION DES RISQUES INDUSTRIELS PAR LE MODELE DE CREDIBILITE » (DOUVILLE, 2004), propose une nouvelle approche du tarif, s'appuyant sur le modèle de crédibilité hiérarchique de Jewell, afin de tenir compte des évolutions significatives du marché de l'assurance des Risques Industriels en France en termes de :

- Structure du marché : le nombre d'intervenants est de plus en plus restreint du fait « *d'une technicité pointue et des capacités importantes exigées pour intervenir sur ce marché* ».

- Population des risques assurés : une concentration des valeurs assurées est observée et se retrouve « *dans la diminution du nombre de risques assurés et la hausse du capital moyen garanti en risque direct* ». Ce phénomène est expliqué par « *des mouvements de concentration et de redistributions sectorielles importants* » et de « *nouvelles technologies utilisées de plus en plus complexes et coûteuses* ».

- Le contenu des contrats d'assurance : le poids des garanties annexes, autres que la garantie Incendie Risques Directs, est en constante augmentation du fait que les entreprises sont de plus en plus en recherche « de nouvelles couvertures d'assurance prenant mieux en compte l'ensemble des risques auxquels elles peuvent être exposées ». Avec une importante diffusion depuis le début des années 1990, la garantie Perte d'Exploitation en est d'ailleurs un parfait exemple. La sinistralité climatique est aussi en forte augmentation, ce qui pousse les acteurs du marché à développer une tarification plus pertinente de cette garantie.

L'avancé du machine Learning est une ouverture possible pour le futur du risque industriel, cependant cette branche est marquée par deux limites à l'utilisation de ces avancées : le faible nombre de contrat et donc de sinistres incendie, la forte variance de l'intensité de la sinistralité incendie. Dans le cas d'autres garanties, cette approche est possible, par exemple le risque climatique.

1.6 TARIFICATION DE L'INCENDIE DE L'ASSURANCE DES RISQUES INDUSTRIELS CHEZ AXA FRANCE

La première méthode de tarification au niveau d'AXA France se basait principalement sur le tarif des risques d'entreprise de France Assureurs, le traité des risques entreprises. Les souscripteurs avaient cependant le choix dans la fixation du prix grâce à des coefficients tarifaires applicables au taux France Assureurs.

Rapidement le travail de France Assureurs devient obsolète et montre des limites concurrentielles, il est donc nécessaire à AXA de maitriser sa tarification.

Dès 2006, AXA a donc personnalisé son tarif à partir d'une NQI (Note qualité incendie), cette note permet de personnaliser le taux France Assureurs à l'entreprise, en réalisant une étude de prévention et de protection sur l'incendie. Cette étude permet de dégager un coefficient qualité propre à l'entreprise tarifé, qui permet une personnalisation du taux incendie de la rubrique de l'entreprise.

Comme énoncé dans l'introduciton, en 2012, AXA a travaillé sur une mise à jour de l'ensemble des rubriques d'activités d'entreprises (TRE) et des taux applicables, dans la continuité de ces travaux, AXA souhaite revoir actuellement ses critères de tarification :

- Pertinence des taux des rubrique d'entreprise (TRE) AXA
- Pertinence de la NQI.

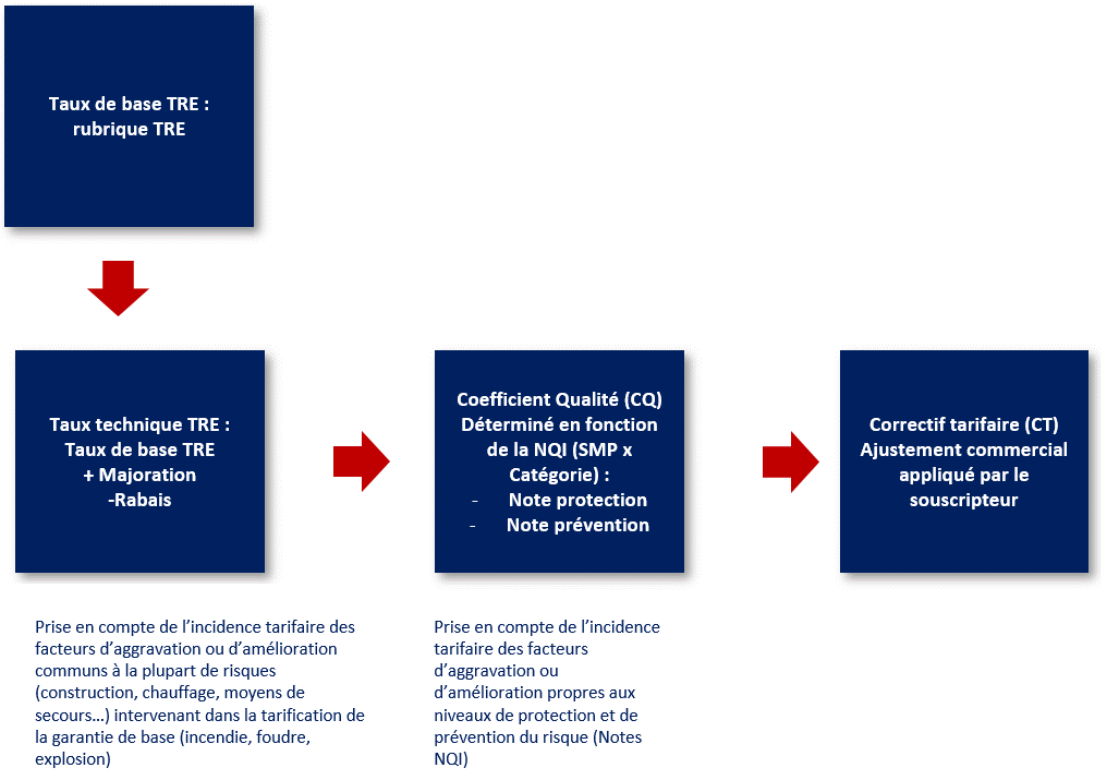


Figure 9 - Méthodologie de tarification en risque industriel

1.7 LE TRAITE DE RISQUE ENTREPRISE DE FRANCE ASSUREURS ET AXA

Le Traité des Risques d'Entreprises comprend trois tomes, il a été produit par France Assureurs. AXA a réalisé de nombreuses études afin de l'adapter à son processus de souscription et à son portefeuille.

1.7.1 TOME I. - RISQUES DIRECTS : DISPOSITIONS GENERALES ET CLAUSES

Les Dispositions Générales portent sur l'appréciation de l'incidence tarifaire des facteurs d'aggravation ou d'amélioration communs à la plupart des risques (construction, chauffage, moyens de secours, etc) intervenant dans la tarification de la garantie de base (incendie, foudre, explosion).

Le tome I traite également des extensions de la garantie de base à d'autres événements, à d'autres biens, à des frais, pertes et responsabilités ainsi que de modalités d'assurance particulières.

1.7.2 TOME II. - PERTES D'EXPLOITATION

Ce tome traite dans une première partie des principes généraux de l'assurance des Pertes d'Exploitation (objet de l'assurance, mécanismes comptables, éléments du contrat, etc) et du règlement des sinistres.

Une deuxième partie est consacrée à la tarification des garanties.

1.7.3 TOME III. - TARIFICATION ANALYTIQUE

La Tarification Analytique (TA) se compose de 124 rubriques (ou classes d'activité) réparties en 10 fascicules (ou familles d'activités), numérotés de 0 à 9. A chaque rubrique est indiquée une appréciation de l'incidence tarifaire des facteurs d'aggravation ou d'amélioration propres à chaque activité.

Dans un onzième fascicule, numéroté 10, sont proposées les dispositions tarifaires applicables à des activités accessoires non prévues ailleurs à la TA.

Sauf dérogations prévues à certaines rubriques de la TA, les Dispositions Générales sont applicables à toutes les entreprises entrant dans le champ d'application du Traité.

Le TOME III France Assureur présente un taux par rubrique, avec un tableau de pénalisation pour mieux qualifier le l'activité.

DEGRE DE DANGER
d'une activité par rapport à l'ensemble des activités d'une même rubrique

Le degré de danger de l'ensemble des activités d'une rubrique considérée est approché par le taux de prime pure statistique donné dans le tableau figurant sous l'intitulé de chaque rubrique.

Classe de danger	Degré de danger de l'activité par rapport à l'ensemble des activités de la rubrique
+F	extrêmement supérieur
+E	très fortement supérieur
+D	fortement supérieur
+C	supérieur
+B	faiblement supérieur
+A	très faiblement supérieur
RE	**Activité à risques équivalents**
-A	très faiblement inférieur
-B	faiblement inférieur
-C	inférieur
-D	fortement inférieur
-E	très fortement inférieur
-F	extrêmement inférieur

Figure 10 - Classe de danger France Assureurs

Une fois le taux fixé, les majorations et rabais pouvant être apportés sont qualifiés par un système de pictogramme semblables.

INCIDENCE TARIFAIRE DES FACTEURS D'AMELIORATION ET D'AGGRAVATION DES RISQUES

Le tableau ci-dessous s'applique aux facteurs d'amélioration et d'aggravation de risques mentionnés aux Dispositions Générales des Tomes I et II du TRE ainsi qu'aux dispositions spécifiques des Tomes II et III du TRE. Il appartient aux sociétés d'assurance de fixer les rabais et majorations correspondant aux niveaux d'incidence tarifaire.

Niveau d'incidence		Incidence
Facteurs d'aggravation	Facteurs d'amélioration	
→	→	Aucune
1↗	1↘	Extrêmement faible
2↗	2↘	Très faible
3↗	3↘	Faible
4↗	4↘	Moyenne
5↗	5↘	Forte
6↗	6↘	Très forte
7↗		Extrêmement forte

Figure 11 - Pictogramme du niveau d'incidence France Assureurs

Les travaux de France Assureurs ont permis la production d'une base de travail et de tarification pour AXA, motivés par les limites de la tarification analytique de France Assureurs :

- **Une création du tarif complexe : Taux x Pictogramme « DANGER » x Pictogramme « Facteur ».**
- **AXA ne maitrise pas son tarif et celui n'est pas personnalisé à son portefeuille.**
- **Une uniformisation du tarif sur le marché**
- **Une obsolescence rapide, FRANCE ASSUREURS n'apportant pas une mise à jour annuelle de ses travaux.**

1.7.4 Personnalisation du Tarif Analytique (TA) par AXA.

En 2012, AXA édite son propre TOME III qui présente les taux de prime « chargés » avec application en sus du chargement spécifique.

Les travaux d'AXA présentent les intérêts suivants :

- Une segmentation plus claire du niveau de risque :
  - Cible
  - Standard
  - Lourd
  - Exclus
- Une granularité des rubriques plus importante
- Un chiffrage des rabais et majoration

L'étude actuelle est une remise à plat de ces travaux.

1.7.5 Illustration entre France Assureurs – AXA version 2012

Les 2 figures ci-dessous montrent la différence entre une fiche de tarification France Assureurs et AXA France.

Taux de prime pure : 1,09 °/°°

Ateliers d'emplissage :

	I	II
Taux de base :	**RE**	**+F**
Majorations :		
• chauffage :	**4 à 5↗**	**-**
• non-insertion des clauses n° 26-P (Absence de foyer), n° 27-A (Installations électriques) et n° 27-C (Thermographie IR)(2) :	**-**	**7↗**
• pour ateliers en sous-sols :	**-**	**4 à 5↗**
• non-insertion de la clause n° 80-F (Evacuation des gaz et vapeurs inflammables) :	**-**	**2↗**

Figure 12 - Fiche de tarification des ateliers d'emplissage par France Assureurs

La version de France Assureurs définit un taux de prime puis oriente la souscription à partir d'indicateurs (RE = Risque Equivalent, +F extrêmement supérieur..).

Ateliers d'emplissage :

	I	II
Taux de base :	**2,03°/°°**	**7,13°/°°**
Majorations :		
• chauffage :	**50%**	**-**
• non-insertion des clauses n° 26-P (Absence de foyer), n° 27-A (Installations électriques) et n° 27-C (Thermographie IR)(2) :	**-**	**200%**
• pour ateliers en sous-sols :	**-**	**50%**
• non-insertion de la clause n° 80-F(Evacuation des gaz et vapeurs inflammables)	**-**	**20%**

Figure 13- Fiche de tarification des ateliers d'emplissage par AXA France

La version d'AXA France est plus stricte, avec un ensemble de taux prédéfinis et sans modulation possible.

1.8 SYNTHESE

La tarification du risque incendie est un domaine possédant 30 ans de travaux actuariels qui se nourrit des évolutions sur la théorie de la crédibilité (Bühlman-Straub puis hiérarchique).

Les évolutions précédentes se sont concentrées sur la théorie de la crédibilité, ce mémoire complète la théorie de la crédibilité par l'ajout de deux étapes préparatives au modèle et aux données, la recherche du meilleur critère hiérarchique et l'étude de la sinistralité.

2 SELECTION ET PREPARATION DES DONNEES

2.1 PRESENTATION DES DONNEES POUR L'ETUDE

La segmentation des rubriques d'activités industrielles selon AXA France date de 2012.

L'étude ayant une volonté d'avoir la profondeur la plus importante, mais de garder de la cohérence, il a donc été décidé de limiter l'historique à 2011, l'information de la rubrique entre 2011 et 2012 étant facile à réaliser. Il a donc été gardé comme historique 11 ans de portefeuille.

Les activités industrielles dont l'effectif n'est pas significatif sur les 5 dernières années ont été écartées. Ainsi que les activités industrielles dont l'effectif des 5 dernières années était inférieur à 0,01% du total observés sur 5 ans.

Cela implique une perte de moins de 1% des contrats et des primes.

Les contrats donc le capital incendie était manquant ont été exclus. Cette absence d'information peut être expliquée par 2 facteurs :

- Perte de l'information sur des contrats avec une ancienneté importante et dont le passage d'un système de gestion à un autre à engendré des pertes.
- Absence de l'information sur des contrats sur mesure.

Cela implique une perte de moins de 0,5% des contrats et des primes.

2.1.1 Variables sélectionnées

Les bases de données AXA sur le risque industriel comportent l'ensemble des éléments propres à l'entreprise (SIRET, Adresse du siège, ...), l'activités industrielles étant associée au secteur de l'entreprise et non à l'entreprise elle-même seuls les éléments suivants ont été conservés : le numéro de contrat, l'activité industrielle, les capitaux incendie, l'année du contrat, le nombre de sinistre, le cout des sinistres ainsi que les 3 mailles envisagées pour l'étude (Fascicule, niveau d'acceptation du risque et Segment).

2.1.1.1 Présentation des fascicules

Dans le traité d'assurance incendie des risques d'entreprises produit par France Assureurs, les rubriques d'activité d'entreprise (TRE) sont regroupées dans dix fascicules, numérotés de 0 à 10. Par exemple, le code TRE « TE601 », qui représente les activités de scieries, fait partie du fascicule n°6 qui regroupe les industries du bois.

La nomenclature des onze fascicules est la suivante :

- Fascicule 0 : Extraction de minerais et minéraux divers, de combustibles solides. Métallurgie.
- Fascicule 1 : Production de matériaux de construction. Industries des céramiques. Industries du verre.
- Fascicule 2 : Travail des métaux. Industries électriques et électroniques. Construction automobile, aéronautique et navale. Carrosserie et réparation de véhicule en tous genre Garages et stations-service.
- Fascicule 3 : Industries chimiques et para-chimiques. Transformation de matières plastique et de caoutchouc.
- Fascicule 4 : Industries textiles. Bonneterie. Confection de vêtements et autres articles textiles.
- Fascicule 5 : Industries du papier et du carton. Imprimeries. Industries du cuir et du délainage.
- Fascicule 6 : Industries du bois.
- Fascicule 7 : Industries agro-alimentaires.
- Fascicule 8 : Traitement des déchets urbains et industriels. Production et distribution d'énergie.
- Fascicule 9 : Autres risques d'entreprises.

- Fascicule 10 : Services généraux et risques annexes concourant à l'exploitation de l'établissement assuré

En 2012, AXA ajoute le fascicule 11, « Autres risques simples ».

2.1.1.2 La catégorie du risque, ou segment

Il existe une segmentation par typologie de risque propre à AXA France à teneur commerciale. Cette segmentation est fondée sur le classement des activités industrielles par fréquence de graves Incendie. Au sein du département Risque Industriel d'AXA, un sinistre grave est un sinistre dont la charge dépasse 150 k€. Cette segmentation AXA comporte 5 segments de risque :

- Le segment « Cible » qui regroupe les activités du portefeuille pour lesquelles la survenance de sinistres graves est la plus faible. Ce sont ces activités qui doivent d'ailleurs être souscrites en priorité.
- Le segment « Standard » constitué des activités dont les résultats en matière de fréquence de graves gravitent autour de la moyenne. Pour ces activités, la souscription ne fait l'objet d'aucune restriction et est d'ailleurs assez encouragée.
- Le segment « Lourd » dans lequel sont regroupées les activités dont la fréquence de grave commence à être élevée par rapport à la moyenne. La souscription de ces activités est un peu plus surveillée et peut faire, de rares fois, l'objet d'études plus approfondies en comité de direction.
- Le segment « Lourd Réservé » qui regroupe les activités ayant de très mauvais résultats en termes de fréquence grave. La souscription de ces activités est très encadrée, car AXA France ne souhaite pas se développer sur ce segment. En général, la souscription de tels risques est motivée par un contexte client mais est le plus souvent refusée.
- Le segment «Exclu» qui regroupe les activités auxquelles AXA France ne souhaite pas souscrire. Ce sont les risques exclus du « Groupe Axa » pouvant générer des sinistres non mutualisables avec le reste du portefeuille RI ou relevant d'exclusion des traités de réassurance. Il n'y a aucune possibilité de souscrire de risques exclus dans le cadre d'AXA Entreprises.

Cette segmentation est revue de manière régulière, mais les retouches effectuées sont minimes et ne concernent qu'une dizaine d'activités tout au plus qui sont transférées d'un segment à un autre suivant l'évolution de leurs résultats.

2.1.1.3 Niveau Acceptation Du Risque : NARI

Afin d'évaluer le niveau de risque d'une entreprise, il existe au sein d'AXA France une catégorisation du risque qui définit si l'entreprise doit être visitée par un collaborateur d'AXA. En fonction du niveau de risque, du plus risqué au moins risqué, le collaborateur peut être :

- Un ingénieur, qui fournira un rapport détaillé à la souscription du niveau de risque
- Un inspecteur expérimenté (IREX) qui possède un pouvoir de souscription, il ne s'agit pas d'un ingénieur mais l'IREX possède une formation sur l'évaluation de la protection et de la prévention de l'incendie dans les entreprises.
- Un inspecteur (IRE).

Les tableaux ci-après définissent les seuils d'intervention des souscripteurs, des IRE, des IREX et de l'ingénierie, en fonction du classement du Niveau Acceptation Du Risque (NARI). Ce classement est lié à la sensibilité des sources potentielles d'éclosion et de propagation à partir des informations disponibles sur des bases techniques reconnues définies par différentes associations internationales (ex : CNPP Centre national de prévention et de protection, National Fire Protection Association).

Ces seuils varient en fonction de la sensibilité des activités et leur taille, selon le classement NARI ci-dessous :

- Groupe 1 : Exclus et activités très sensibles
- Groupes 2 à 4 : des activités industrielles les plus sensibles aux activités les moins dangereuses
- Groupe 5 : activités non industrielles hors groupe 6
- Groupe 6 : entrepôts, commerces et distribution (ERP de type M)

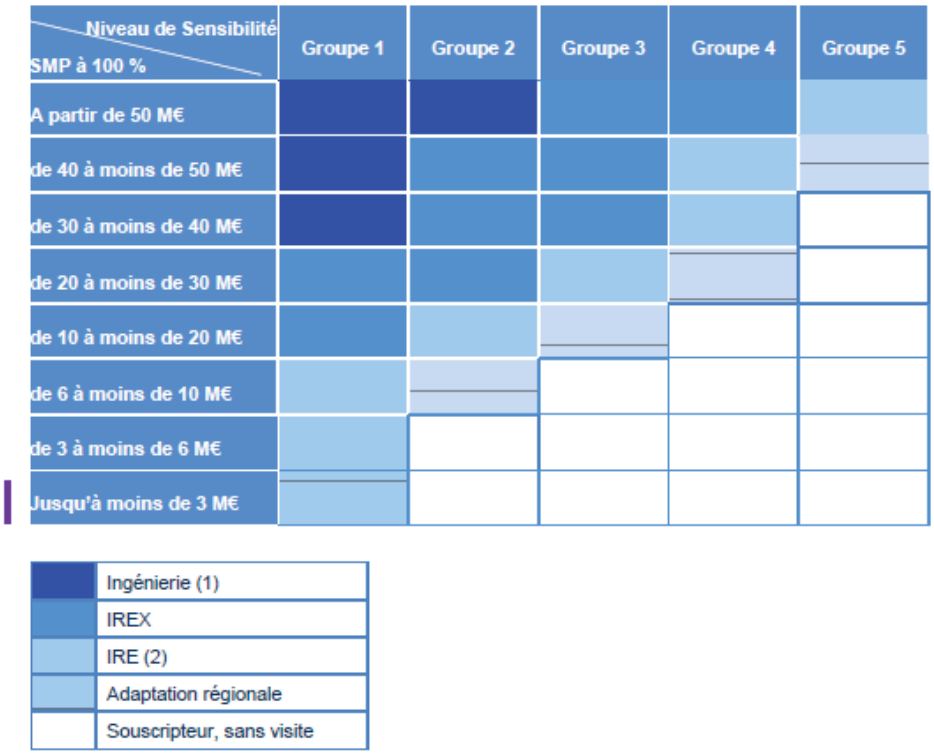


Figure 14 - Guide des visites par groupe de risque (NARI) et sinistre maximum possible

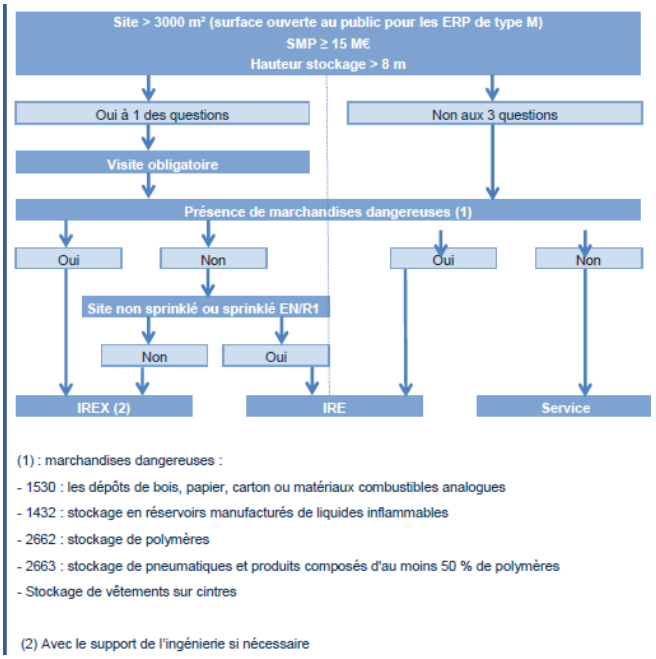


Figure 15 - Guide des visites pour le groupe de risque 6 (NARI 6)

2.1.2 Statistiques descriptives

Pour donner suite à notre sélection de données, l'étude présente des analyses univariées sur chaque variable afin de permettre au lecteur de mieux comprendre la base de données

2.1.2.1 L'année police

L'année police étant avec une vision annuelle, car le détail de la donnée est à l'année, elle est comprise entre 0 et 1.

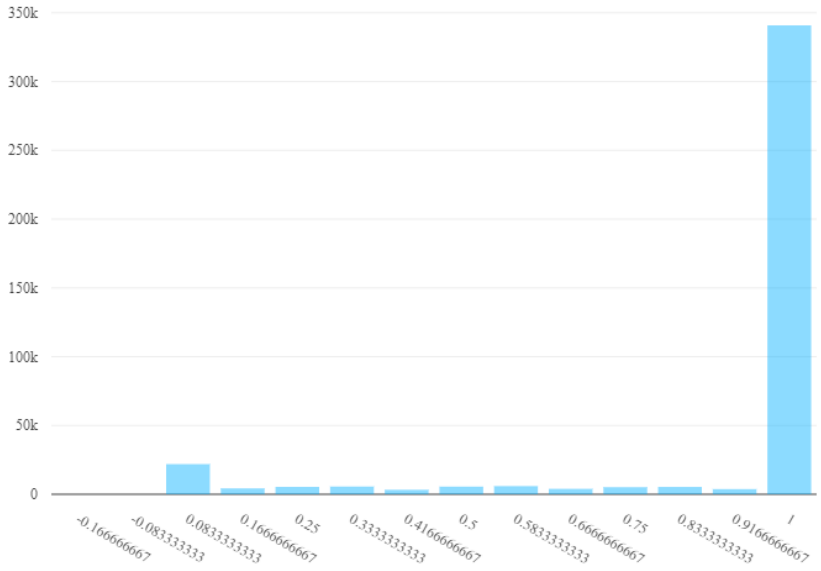


Figure 16 - Année police par contrat

La majorité des contrats du portefeuilles ont une durée annuelle, ce qui s'explique par une grande stabilité du portefeuille d'AXA France. Les années polices inférieures à 0 sont exclues de la base. Uniquement 2 lignes présentent cette anomalie, il s'agit d'erreur de saisie.

2.1.2.2 Contrat par année

Le portefeuille d'AXA étant mature, le nombre de contrats par année est stable.

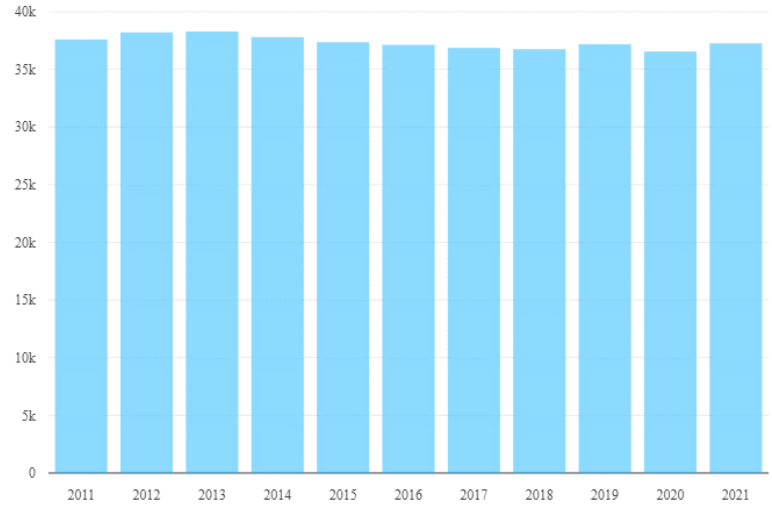


Figure 17 - Nombre de contrats par année

2.1.2.3 NARI (Niveau d'acceptation du risque)

La répartition des contrats est la suivante :

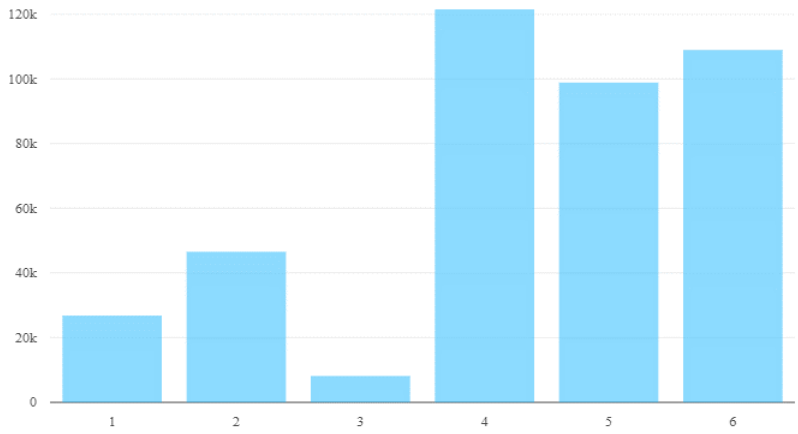


Figure 18 - Nombre de contrats par groupe de risque (NARI)

Pour mieux comprendre le décalage entre les niveaux d'acception du risque (NARI) 2,3 et 4, une étude de la répartition des rubriques d'entreprise (TRE) est nécessaire. Une rubrique appartement à un seul niveau d'acceptation du risque (NARI).

NARI	Nombre de rubrique (TRE)
1	44
2	24
3	15
4	28
5	34
6	18

Tableau 1 - Nombre de rubrique par niveau d'acceptation du risque

Le niveau de risque 3 est sous-représenté dans les rubriques (TRE). Cela s'explique principalement par la facilité de classer les TRE par sensibilité binaire (sensible pour NARI 4 et non sensible pour le NARI 2) plutôt qu'apporter une nuance d'entre deux (NARI 3). Pour rappel les groupes ou niveaux 2 à 4 sont les activités industrielles les plus sensibles aux activités les moins dangereuses.

2.1.2.4 Rubrique entreprise (TRE)

Le nombre de rubriques (TRE) étant très important, graphiquement la présentation serait illisible, voici les 10 plus gros TRE du portefeuille avec la part de la charge et le nombre en % du portefeuille

TRE	En charge	En nombre
TRE 1	6,44%	9,35%
TRE 2	3,44%	7,15%
TRE 3	6,38%	4,59%
TRE 4	4,16%	4,34%
TRE 5	1,92%	4,16%
TRE 6	4,95%	3,23%
TRE 7	3,47%	3,15%
TRE 8	1,28%	2,53%
TRE 9	1,59%	2,51%
TRE 10	2,13%	2,38%

Tableau 2 - Top 10 des rubriques AXA France (TRE anonymisés)

Les 10 premiers TRE du portefeuille représentent 35% de la charge total pour 43% des contrats.

2.1.2.5 FASCICULE

La répartition des contrats par fascicule est la suivante :

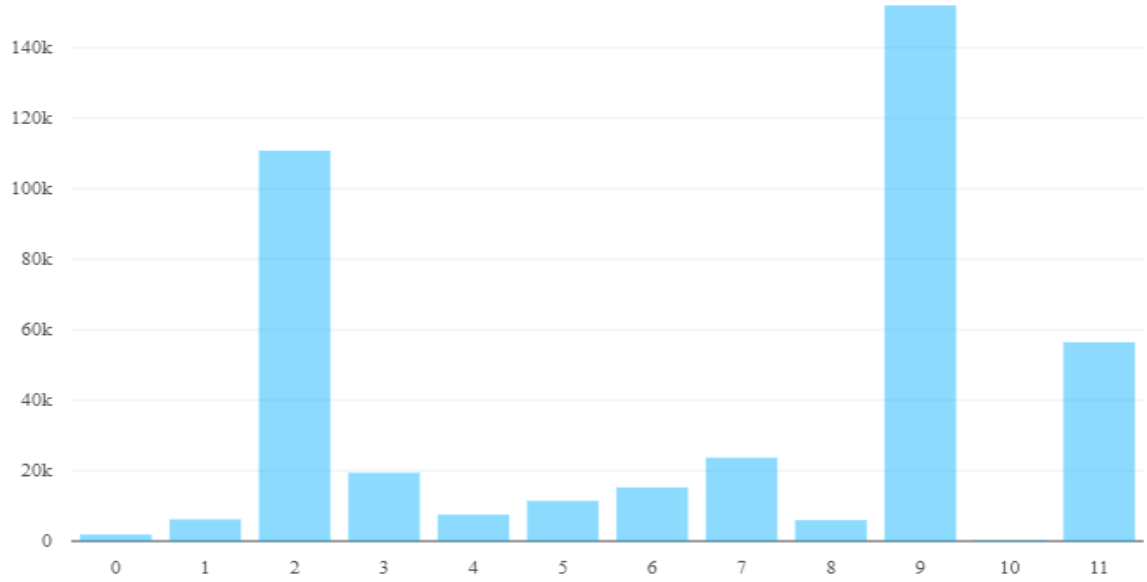


Figure 19 - Contrats par fascicule

La majorité des contrats du portefeuille d'AXA France sont regroupés dans les fascicules 2 et 9 :

- Fascicule 2 : Travail des métaux. Industries électriques et électroniques. Construction automobile, aéronautique et navale. Carrosserie et réparation de véhicule en tous genre Garages et stations-service.
- Fascicule 9 : Autres risques d'entreprises

2.1.2.6 SEGMENT

La répartition des contrats par segment (Cible, standard, Lourd et Exclus) est la suivante

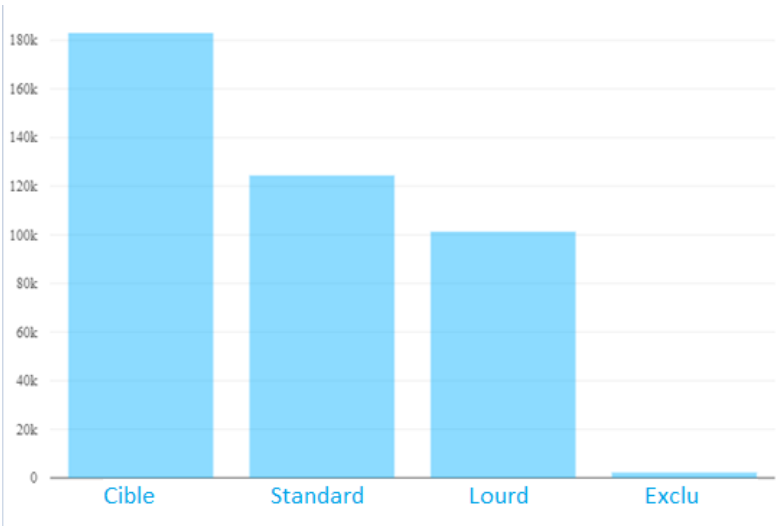


Figure 20 – Contrats par segment (Cible, Standard, Lourd (yc Lourd Réservé) Exclus)

La figure ci-dessus montre que la politique d'AXA France sur les exclus permet une souscription limitée des rubriques concernées.

La sinistralité incendie est présentée dans la présentation de la théorie des valeurs extrêmes.

2.2 TRAVAUX SUR LES DONNEES

2.2.1 Inflation

Afin d'adapter la sinistralité passée pour la tarification 2022, il convient de de prendre en compte ce paramètre, 150 000 € en 2021 n'ont pas la même valeur qu'en 2011.

Afin de ne pas sous-estimer le poids des exercices les plus anciens, les données extraites ont été revalorisées suivant l'indice RI au 1er janvier 2021. Police par police et exercice par exercice, il a été calculé la variation de l'indice risque industriel (RI) entre le 1er janvier de l'exercice considéré et le 1er janvier 2021. Cette variation est alors appliquée aux capitaux Incendie assurés et aux règlements effectués sur les sinistres.

L'indice risque industriel est l'indice sur lequel sont indexés tous les contrats d'assurances dommages des entreprises dont le contenu à assurer (matériel et/ou marchandises) a une valeur supérieure à 150 fois la valeur en euros de l'indice RI. L'indice RI tient compte de l'évolution des prix relatifs aux bâtiments, au matériel, aux marchandises et à la main d'œuvre. L'indice RI est calculé trimestriellement selon la formule suivante :

Formule de l'indice : $I = 45 + 2{,}26\,A + 19{,}43\,B + 5{,}64\,C + 8{,}60\,D$ (indice composite sur terme à échoir).

Où :

- A : l'indice FFB (ex FNB) du coût de la construction (base 1 au 1er janvier 1941).
- B : l'indice du coût de la main d'oeuvre pour les industries mécaniques et électriques (INSEE, base 100 moyenne de l'année 1985).
- C : l'indice du prix de vente industriel des métaux (INSEE, base 100 moyenne de l'année 1985).
- D : l'indice du prix de vente des biens intermédiaires (INSEE, base 100 moyenne de l'année 1985).

L'historique de l'indice risque industriel est le suivant :

Années	Au 1er janvier	Au 1er avril	Au 1er juillet	Au 1er octobre	Indice i / Indice 2021
2 021	6 186	6 206	6 280	6 385	100,00%
2 020	6 206	6 205	6 203	6 195	99,03%
2 019	6 134	6 157	6 179	6 194	98,44%
2 018	5 987	6 019	6 052	6 100	96,42%
2 017	5 807	5 846	5 899	5 948	93,79%
2 016	5 840	5 819	5 796	5 783	92,76%
2 015	5 783	5 805	5 814	5 819	92,68%
2 014	5 746	5 751	5 758	5 772	91,91%
2 013	5 711	5 716	5 769	5 753	91,60%
2 012	5 627	5 637	5 683	5 690	90,35%
2 011	5 321	5 388	5 496	5 573	86,91%

Tableau 3 - Indice RI 2010 - 2021

L'inflation suivante a été appliquée :

$$Sinistre_{\ i\ vision\ 2021} = Sinistre_i \times \frac{Indice_{2021}}{Indice_i}$$

et

$$Capitaux_{i\ vision\ 2021} = Capitaux_i \times \frac{Indice_{2021}}{Indice_i}$$

2.2.1.1 Projection de la sinistralité

La Direction Actuariat & Pilotage Entreprise établit les coefficients de vieillissement de la sinistralité du RI.

Pour rappel, le provisionnement en assurance consiste pour l'assureur à mettre de côté de l'argent dans le but de pouvoir faire face à ses engagements vis-à-vis des assurés. Il joue un rôle très important car l'assureur collecte les primes sans savoir exactement quand et combien il devra payer lors de la survenance des sinistres.

La Chain Ladder (CL) est la méthode de provisionnement la plus connue et la répandue sur le marché de l'assurance non-vie. Elle permet d'estimer les sinistres survenus mais non encore déclarés et de projeter des valeurs observées jusqu'à la fin de tous mouvements des sinistres, c'est-à-dire jusqu'à l'ultime.

Actuariellement cette méthode consiste dans le calcul du facteur de développement individuel $f_{i,j} = \frac{C_{i,j+1}}{C_i}$ pour i=1…n, j=1…n et en partant du principe que pour j allant de 1 à n, les facteurs de développement $f_{i,j}$ sont indépendants de l'année de survenance i. Ainsi le coefficient de passage d'une année à l'autre, commun pour les années de survenance, et dont l'estimation est donnée par :

$$\hat{f}_i = \frac{\sum_{i=0}^{n-j+1} C_{i,j+1}}{\sum_{i=0}^{n-j+1} C_{i,j}}, j\epsilon[0,n]$$

Soit une projection à l'ultime de la charge par la formule suivante :

$$\hat{C}_{j,n} = C_{j,n-j} \times \prod_{i=n-j}^{n-1} \hat{f}_i$$

Le graphique suivant montre le vieillissement de la charge de sinistre du RI :

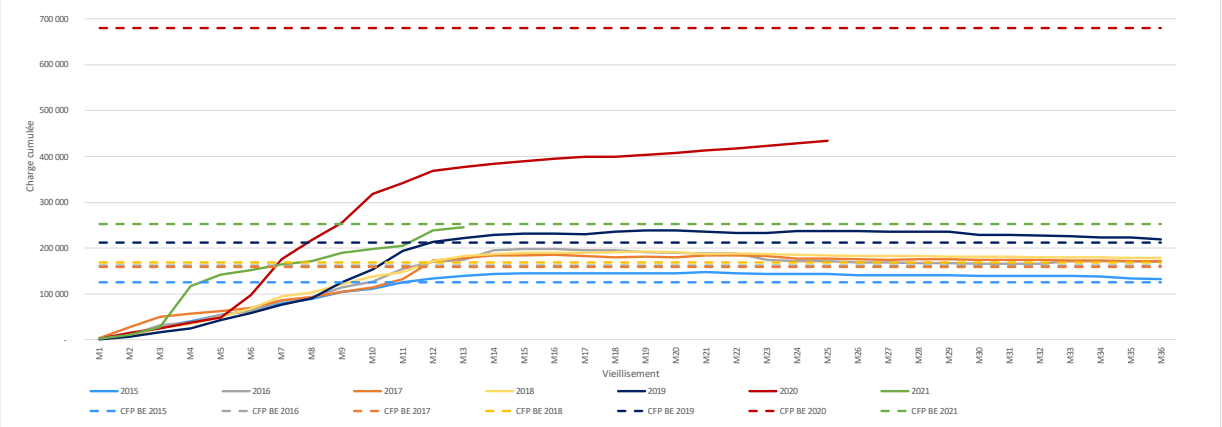


Figure 21 - Vieillissement de la sinistralité par année

Le vieillissement dans le cas du risque industriel est assez rapide, l'année 2020 étant particulière car touchée par les pertes d'exploitations sans dommages liées au COVID.

Sur la sinistralité incendie, les années 2019 à 2021 ont donc été actuariellement vieillies à partir de ces triangles de cadencement.

2.2.2 Mutualisation des sinistres graves, atypiques et super-atypiques

Le marché des Risques d'entreprises est un marché à très faible fréquence de sinistres graves qui font la majorité de la charge (plus de 70% en Risques Industriels). Par conséquent, pour construire un modèle fiable, la mutualisation de la charge est un élément important de l'étude.

Cette méthode est historiquement utilisée dans le RI, cependant comme le précisent les travaux de COHEN, DUPIN et LEVY, en 1985, les rubriques d'entreprise (TRE) avec des capitaux importants ont peu de chance de connaitre des sinistres susceptibles d'être écrêtés si cet écrêtement est basé sur la prime encaissée.

C'est pour cela qu'il ne faut pas se limiter à cette étape mais l'utiliser l'écrêtement comme base de la Théorie de la crédibilité et redéfinir les seuils d'écrêtement pour qu'ils soient communs à l'ensemble des rubriques.

Une police fortement sinistrée pénalisera de manière importante un TRE. Les charges importantes en contrepartie d'une fréquence faible représentent une partie importante de la sinistralité.

Il ne permet pas d'avoir une application de la crédibilité de manière fiable sur le portefeuille AXA. Certaines années sont trop sinistrées et la variance entre les TRE est trop importante.

Le graphique ci-dessous présente la diversité des S/P observés sur 10 ans.

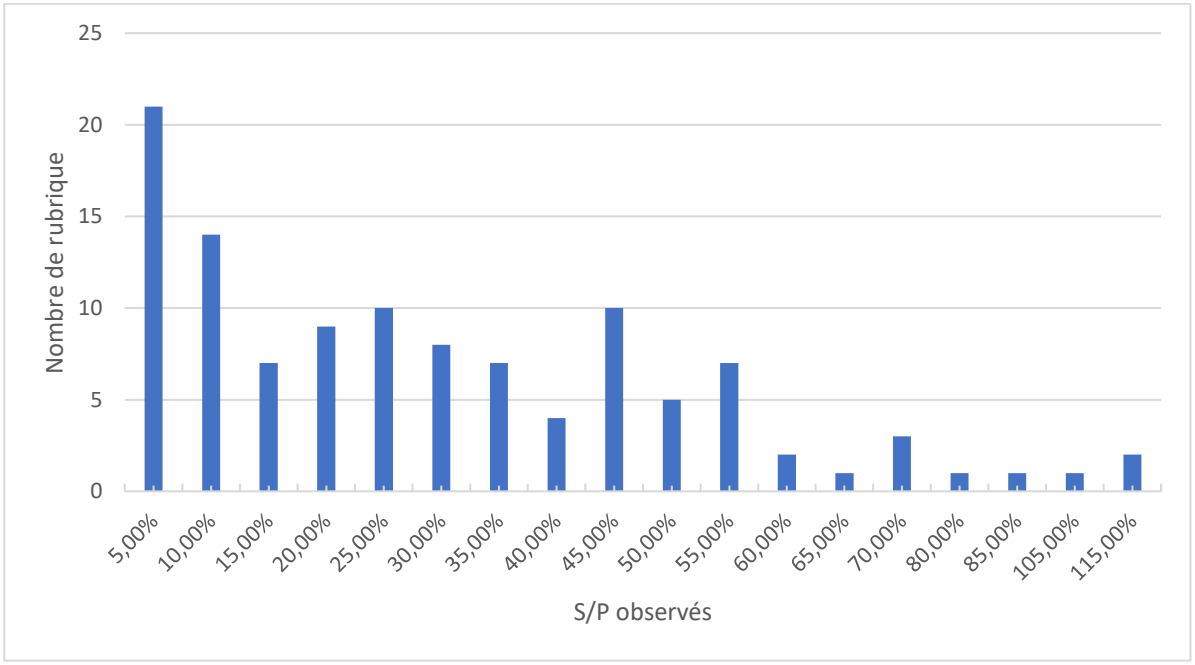


Figure 22 - Répartition des contrats par S/P

En mutualisant la charge par 3 seuils, l'information individuelle se renforcera sur l'information collective (niveau hiérarchique).

Le premier levier est la mutualisation de la charge grave sur l'historique du TRE, afin de conserver le maximum de poids sur la rubrique d'activité (TRE) mais en écrêtant les charges dépassant le premier seuil.

Le second levier est la mutualisation de la charge atypique dépassant le second seuil sur le niveau hiérarchique considéré pour l'étude. Cela permet de renforcer l'effet du niveau hiérarchique.

Le dernier levier est la mutualisation de la charge super-atypique dépassant le troisième seuil sur le portefeuille d'AXA. Ainsi la mutualisation de conserver l'une des bases de l'assurance, la mutualisation du risque. Une entreprise ne rentre pas dans un portefeuille lambda mais dans le portefeuille AXA.

Les 3 leviers déterminés, le choix de la méthode de mutualisation reste à déterminer. Deux axes sont possibles : l'écrêtement dès le premier euro contre l'écrêtement à partir du nièmes euros.

Par soucis de simplicité, la seconde solution a été sélectionnée.

L'application d'une méthode d'écrêtement-mutualisation a pour objectif d'atténuer la volatilité de la sinistralité caractéristique des risques d'intensité. Le but premier est de lisser la charge des graves d'une année (ou de plusieurs années) d'un TRE sur l'historique de la rubrique.

Les seuils proposés :

- Les sur-crêtes (S_i) comprises entre le premier et le second seuil sont mutualisées sur chaque exercice de la rubrique d'activité au prorata du poids des capitaux assurés de chaque exercice ($w_{i,j}$) pour la rubrique d'activité dans l'assiette totale des capitaux de la rubrique d'activité (TRE) sur l'ensemble de la période (w_i) : mutualisation sur la période d'une activité.

$$Charge\ mutualisée\ en\ année\ j\ =\ S_i \times \frac{w_{i,j}}{w_i}$$

- Les sur-crêtes sont mutualisées entre le seuil 2 et le seuil 3 (des atypiques) au niveau de la classe de regroupement de la rubrique d'activité, la mutualisation est donc sur la période au niveau de la classe selon le même prorata : mutualisation sur les rubriques d'activités à l'intérieur d'une classe.
- La dernière mutualisation pour les sur-crêtes supérieures à S3 (des super-atypiques) au niveau du portefeuille.

Charge	Charge conservée sur l'année i du TRE	Mutualisation dans le TRE sur la période	Mutualisation dans le groupe sur la période	Mutualisation dans le portefeuille sur la période
Non Grave	Charge			
Grave	0 € à Seuil 1	Sur-crête du Seuil 1		
Atypique	0 € à Seuil 1	Seuil 1 à Seuil 2	Sur-crête du Seuil 2	
Super- Atypique	0 € à Seuil 1	Seuil 1 à Seuil 2	Seuil 2 à Seuil 3	Sur-crête du Seuil 3

Tableau 4 - Méthodologie Ecrêtement-Mutualisation de la charge

Exemple, pour les seuils 1, 2 et 3 à respectivement 1, 3 et 5 M €:

	Année	Charge année
Contrat 1 (TRE 1 & Groupe 1)	1	**2 500 000 €**
Contrat 1 (TRE 1 & Groupe 1)	2	**500 000 €**
Contrat 2 (TRE 1 & Groupe 1)	1	**4 500 000 €**
Contrat 3 (TRE 2 & Groupe 1)	1	**250 000 €**
Contrat 4 (TRE 3 & Groupe 2)	2	**7 000 000 €**

Tableau 5 - Contexte de l'exemple de la mutualisation de la charge

A la maille contrat la mutualisation sera la suivante :

	Année	Charge conservée sur l'année i du TRE	Mutualisé sur la période du TRE	Mutualisé sur le groupe du TRE	Mutualisé sur le portefeuille
Contrat 1 (TRE 1 & Groupe 1)	1	1 000 000 €	1 500 000 €	- €	- €
Contrat 1 (TRE 1 & Groupe 1)	2	500 000 €		- €	- €
Contrat 2 (TRE 1 & Groupe 1)	1	1 000 000 €	2 000 000 €	1 500 000 €	- €
Contrat 3 (TRE 2 & Groupe 1)	1	250 000 €	- €	- €	- €
Contrat 4 (TRE 3 & Groupe 2)	2	1 000 000 €	2 000 000 €	2 000 000 €	2 000 000 €

Tableau 6 - Illustration de la mutualisation - maille contrat

Pour l'étude, la mutualisation par rubrique d'activité (TRE) correspondra au tableau ci-dessous :

	Année	Charge année	Charge conservée sur l'année i du TRE	Mutualisé sur la période du TRE	Mutualisé sur le groupe du TRE	Mutualisé sur le portefeuille
TRE 1	1	*7 000 000 €*	2 000 000 €	3 500 000 €		
	2	*500 000 €*	500 000 €		1 500 000 €	
TRE 2	1	*250 000 €*	250 000 €	0 €		2 000 000 €
	2	*0 €*	0 €			
TRE 3	1	*0 €*	0 €	2 000 000 €	2 000 000 €	
	2	*7 000 000 €*	1 000 000 €			

Tableau 7 - Illustration mutualisation de la charge - Maille rubrique (TRE)

2.3 SYNTHESE

La base de données comprend l'ensemble des informations nécessaires pour la tarification de la garantie incendie.
Des travaux de projection et d'actualisation (inflation) sont nécessaires afin de rendre la base de données cohérente sur l'historique étudié.

Pour éviter les effets d'intensité des sinistres, la mutualisation de la sinistralité à partir de certains seuils est primordial, la théorie des valeurs extrêmes va permettre de déterminer ces seuils. Une fois cette étape réalisée, l'étude appliquera les travaux de crédibilité à la base de données présentée dans cette partie.

3 TARIFICATION INCENDIE VIA LA THÉORIE DE LA CRÉDIBILITÉ

Cette partie parcours l'ensemble de la crédibilité, de la fluctuation limitée à la crédibilité hiérarchique dans le but d'accompagner le lecteur dans l'évolution de cette théorie, et pour l'aider à comprendre l'origine des estimateurs calculés.

La théorie de la crédibilité regroupe des méthodes mathématiques de tarification en assurance fréquemment utilisées par les actuaires. Deux concepts fondamentaux lui sont sous-jacents : le risque collectif, c'est-à-dire l'évaluation d'un risque à l'échelle d'un groupe d'assurés pour en capturer les tendances, et le risque individuel, c'est-à-dire l'expérience de sinistres propre à un assuré.

Plus généralement, le terme « crédibilité » fait référence au poids, ou niveau de confiance, qu'un organisme d'assurance décide d'attribuer à l'expérience individuelle d'une entreprise (ou dans le cadre de l'étude, d'une rubrique) lors de la construction de sa prime de souscription. Ce principe est illustré par la formule suivante :

$$\text{Prime finale} = (1 - \text{Z}) * \text{Prime collective} + \text{Z} * \text{Prime individuelle}$$

avec $Z \in [0,1]$ le facteur de crédibilité.

Ainsi, plus une rubrique d'entreprise sera de grande taille et disposera d'une expérience de sinistres importante, plus l'organisme d'assurance lui accordera de confiance, et plus le facteur Z sera proche de 1. À l'inverse, plus une entreprise sera de petite taille, moins elle disposera de données de sinistres, plus sa tarification sera basée sur l'expérience collective du portefeuille global et Z sera proche de 0.

3.1 MODELE DE LA FLUCTUATION LIMITEE

3.1.1 Les origines

D'après la publication « Mathématiques de l'assurance Non-Vie » (Charpentier & Denuit, 2004) en 1910, l'entreprise $General\ Motors$ assurée contre les accidents du travail chez $Allstate$ remarque que sa prime d'expérience, construite sur l'expérience de sinistres propre à son entreprise, est sensiblement inférieure à la prime collective réclamée à l'ensemble des autres entreprises assurées. En argumentant que son exposition au risque est suffisamment importante, $General\ Motors$ réclame alors un tarif fondé exclusivement sur sa propre sinistralité et non celle de l'ensemble du portefeuille assuré. Au même moment, le constructeur indépendant $Tucker$ fait une demande similaire.
La question de l'historique nécessaire à une entreprise pour avoir une tarification personnalisée est alors émise.
$Mowbray$ puis $Whitney$ respectivement en 1918 apportent les premières réponses à cette question en fondant la théorie de la fluctuation limitée ou crédibilité américaine. La théorie de la fluctuation limitée constitue la première approche de crédibilité qui a été utilisée dans l'histoire pour inclure l'expérience de sinistres propre des entreprises dans leur tarification.

3.1.2 Fluctuation limitée

Deux situations sont possibles :

- Si $\mathbf{Z = 1}$, il est question de crédibilité totale. L'organisme d'assurance a suffisamment confiance en la taille et l'historique de sinistres de l'entreprise pour lui réclamer sa prime d'expérience :

$$Prime\ finale_{entreprise} = Prime_{exp.\ ind.\ entreprise}$$

- Si $\mathbf{Z} \in]\mathbf{0}, \mathbf{1}[$, il est question de crédibilité partielle. L'organisme d'assurance réclame alors une prime finale fonction de l'expérience collective et de l'expérience individuelle pondérée par le facteur de crédibilité Z :

$$Prime\ finale_{entreprise} = Z * Prime_{exp.\ ind.\ entreprise} + (1 - Z) * Prime_{exp.\ coll.\ ptf}$$

Le problème de la crédibilité réside donc dans le choix entre la prime individuelle et de la crédibilité totale, principalement sur l'exposition nécessaire à l'assuré pour appliquer une prime d'expérience individuelle.

3.1.3 Crédibilité totale

Soit N la variable aléatoire associée à la fréquence de sinistres. La loi attribuée au risque de fréquence est la loi de Poisson.

En partant du postulat que $\mathrm{N} \sim P(n)$ donc $\mathbb{E}[\mathrm{N}] = \mathbb{V}[\mathrm{N}] = n$, représentant le nombre de sinistres moyen d'un assuré par an. N prendra ses valeurs au voisinage de n d'une année sur l'autre.

Soit S la charge de sinistre avec $S = \sum_{k=I}^{N} X_k$ avec (X_k) une suite v.a.i.i.d. de moyenne μ et de variance σ^2 indépendante de N.

$$E[S] = n \times \mu$$

$$\mathrm{Var}[S] = \mathrm{E}[\mathrm{N}] \times (Var[X_i] + E^2[X_i]) = \mathrm{n}(\sigma^2 + \mu^2)$$

L'entreprise d'assurance accordera à l'assuré une crédibilité totale si la probabilité que la charge de sinistres s'éloigne de la moyenne est suffisamment faible, i.e. si

$$\mathbb{P}\left[\frac{|\mathrm{S} - \mathrm{E}[\mathrm{S}]|}{\mathrm{E}[\mathrm{S}]} \geq \mathrm{c}\right] \leq \epsilon$$

où c et ϵ sont des constantes fixées a priori suffisamment petit.

A partir du théorème central-limite (applicable à une somme aléatoire), pour un nombre important d'observations, il est possible d'approcher :

$$\frac{S - E[S]}{\sqrt{Var[S]}} \sim \mathrm{N}(0,1)$$

Soit :

$$\frac{c \times E[S]}{\sqrt{Var[S]}} = \frac{c\mu n}{\sqrt{n(\sigma^2 + \mu^2)}} = z_{\epsilon/2}$$

Ainsi le nombre de sinistres attendus pour la crédibilité totale permettant l'application de la crédibilité totale est :

$$\mathrm{n}_0 = \left(\frac{z_{\epsilon/2}}{c}\right)^2 \left(1 + \frac{\sigma^2}{\mu^2}\right)$$

3.1.4 Crédibilité partielle

La crédibilité totale détermine le nombre de sinistres minimal attendu de la part d'un assuré pour lui construire sa prime uniquement sur son expérience individuelle.

Lorsqu'un assuré ne dispose pas du volume de risque n_0, la théorie de la fluctuation limitée permet de déterminer une prime qui sera un barycentre de la prime collective et de la prime d'expérience.

Soit N le nombre de sinistres d'un assuré suivant une loi de $Poisson$ de paramètre n.

Z le facteur de crédibilité partielle qui détermine le niveau de confiance que l'organisme d'assurance accorde à l'expérience individuelle de l'assuré.

$$\mathbb{P}\left[\frac{|S - E[S]|}{E[S]}\ \mathrm{Z} \geq\ \mathrm{c}\right] \leq\ \epsilon$$

On déduit donc : $Z\ = \frac{\sqrt{\mathrm{n}}}{\left(\frac{z_{\epsilon/2}}{c}\right)\sqrt{\left(1+\frac{\sigma^2}{\mu^2}\right)}}\ = \sqrt{\frac{\mathrm{n}}{\mathrm{n}_0}}$

De plus, si $n\ >\ \mathrm{n}_0$ alors $Z\ =\ 1$ par convention et la condition de crédibilité totale est retrouvée. Z peut alors se réécrire :

$$\mathrm{Z}\ = \min\left(\sqrt{\frac{\mathrm{n}}{\mathrm{n}_0}}, 1\right)$$

Plus le nombre de sinistre sera important, plus le facteur de crédibilité de l'entreprise sera important.

Pour appliquer la théorie de la fluctuation limitée, AXA France souhaite déterminer un facteur de crédibilité partielle pour la boulangerie. Un calcul analogue à la crédibilité totale a donné à AXA France comme référence $\mathrm{n}_0\ =\ 6$ incendie minimum pour accorder la crédibilité totale. De plus, l'entreprise assurée présente dans son historique de sinistres $n\ =\ 12$ incendies. Alors $Z\ = \sqrt{\frac{\mathrm{n}}{\mathrm{n}_0}}\ =\ 71\ \%$ de crédibilité lui sera accordé. Sa prime finale sera donc :

$$\mathrm{Prime}_{\mathrm{BoulangerieE}} = 0{,}71 * \mathrm{Prime}_{\text{exp. ind. Boulangerie}} +\ 0{,}29 * \mathrm{Prime}_{\text{exp. coll. ptf}}$$

3.1.5 Limites

Si le principal intérêt de la fluctuation limitée réside dans sa simplicité de mise en œuvre, cette théorie souffre d'inconvénients qui en font un outil peu utilisé en pratique :

- le choix délicat des paramètres c, la probabilité de déviation, et ϵ le seuil de déviation, qui ne dispose a priori d'aucune justification mathématique rationnelle.
- l'approximation centrale limite ;
- une approche paradoxale : pour obtenir le facteur de crédibilité Z il est nécessaire de faire l'hypothèse que n le paramètre de $poisson$ et μ, σ les paramètres $normaux$ sont connus. En d'autres termes il est possible de calculer la prime pure d'un assuré considéré sans passer par cette méthode.

3.2 FORMALISATION MATHEMATIQUE

Après les premiers modèles de la fluctuation limitée, Hans Buhlmann définit le problème de la tarification sous une nouvelle formalisation bayésienne plus sophistiquée et plus robuste.

Il prend en considération les profils de risque des assurés. L'objectif est identique que la fluctuation limitée c'est-à-dire la recherche de la prime la plus juste possible compte tenu de l'expérience collective du portefeuille de référence, et du niveau de confiance accordé à l'expérience individuelle des assurés.

3.2.1 Risque individuel et collectif

Pour un assuré qui a des montants agrégés de sinistres $\left(\mathrm{X_j}\right)_{\mathrm{j=1\ldots n}}$ où $\mathrm{X_j}$ est le montant de sinistre correspondant à la période j.

A partir de ce cadre, le problème de la tarification est la détermination de la meilleure prime pure $\mathbb{E}[\mathrm{X_{n+1}}]$ pour la période suivante $n+1$ selon l'hypothèse que les $\mathrm{X_j}$ sont indépendant, identiquement distribués, et de fonction de répartition F.

En pratique, F est inconnue et varie d'un risque à l'autre. Soit donc $\theta \in \Theta$ définit comme le profil de risque de l'assuré et $\mathrm{F_\theta}$ la fonction de répartition correspondante.

La prime individuelle correcte pour un assuré de profil de risque $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ est définit par :

$$\mathbf{P^{ind}}(\boldsymbol{\theta}) = \mathbb{E}[\mathbf{X_{n+1}} \mid \boldsymbol{\theta}] = \boldsymbol{\mu}(\boldsymbol{\theta})$$

En pratique, le profil de risque θ n'est pas connu. En effet, les assurés ont des profils de risque θ_i avec $i \in \{1, \dots, \mathrm{l}\}$ inconnus de l'organisme d'assurance. Déterminer $\mu(\theta)$ directement n'est donc pas possible. Il est alors nécessaire de construire un estimateur $\widehat{\mu(\theta)}$.

En assurance industrielle, sur le risque incendie, il est possible de distinguer deux profils de risque : les haut de niveau de protection (extincteur, murs anti-incendie..) et les mauvais élèves. Alors :

$$\Theta = \{\text{Bonne Protection}, \text{Mauvais élève}\}$$

Dans le cas d'un portefeuille réellement homogène, Θ est réduit à un singleton.

L'assureur dispose cependant d'informations sur la structure collective du risque (ex : la plupart des entreprises sont des bons risques qui ont rarement un incendie (risque d'intensité en non de fréquence)). Cette information peut être résumée par la distribution de probabilité U(θ) sur Θ. La distribution $\mathbf{U}(\boldsymbol{\theta})$ est appelée fonction de structure du portefeuille. La prime collective est alors donnée par :

$$P^{coll} = \int_\Theta \mathrm{b}\mu(\theta) dU(\theta) = \mu_0$$

Cette prime correspond simplement au montant moyen de sinistres sur l'ensemble du portefeuille.

La prime la plus compétitive est la prime individuelle correcte. Sinon il y a des opportunités d'arbitrage.

3.2.2 Formulation bayésienne

Le problème de la tarification précédent se décrit en termes bayésiens :

- Θ est une variable aléatoire de fonction de répartition U ;
- conditionnellement à $\Theta = \theta$ les variables aléatoires $\mathrm{X_1}, \mathrm{X_2}, \dots$ précédentes sont indépendantes et identiquement distribuées de fonction de répartition $\mathrm{F_\theta}$.

Les différentes primes qui rentrent alors en jeu se résument ainsi :

- $\mathrm{P^{ind}} = \mathbb{E}[\mathrm{X_{n+1}} \mid \Theta] = \mu(\Theta)$, la prime individuelle
- $P^{coll} = \int_\Theta \mu(\theta) dU(\theta) = \mu_0 = E[X_{n+1}]$, la prime collective
- Les X_1, X_2 ne sont indépendants que conditionnellement à $\Theta = \theta$. Dans le cas contraire $Cov[X_1, X_2] = Var[\mu(\Theta)]$

L'objectif de la formulation bayésienne est d'estimer pour chaque assuré la prime individuelle $\mu(\Theta)$ aussi précisément que possible compte-tenu de l'historique de sinistres $X = (X_1, \dots, X_n)$ disponible pour chacun.

La meilleure prime d'expérience ou prime de Bayes est définie par :

$$\mathbf{P^{Bayes}} = \widetilde{\boldsymbol{\mu(\Theta)}} = \mathbb{E}[\boldsymbol{\mu(\Theta)} \mid \mathbf{X}]$$

Il s'agit d'une prime certaine contrairement à la prime individuelle. C'est le meilleur estimateur en vertu du critère de l'erreur quadratique**.**

L'erreur quadratique moyenne de la prime de Bayes vaut :

$$\mathbb{E}\left[\left(\widetilde{\boldsymbol{\mu(\Theta)}} - \boldsymbol{\mu(\Theta)}\right)^2\right] = \mathbb{E}\left[\mathbb{V}[\boldsymbol{\mu(\Theta)}] \mid \mathbf{X}\right]$$

Pour déterminer une telle prime, plusieurs éléments doivent alors être spécifiés :

- la fonction de structure $U(\theta)$, autrement appelée la distribution a priori de Θ ;
- la famille de distribution conditionnelle $F := \{ F_\theta(x) : \theta \in \Theta \}$, autrement appelée la distribution a posteriori de Θ.

3.2.3 Prime de Bayes

Pour déterminer la prime de Bayes, deux familles de distribution a priori et a posteriori de Θ doivent être nécessairement définies. Dans certains cas, lorsque ces distributions sont bien connues et conjuguées, la prime de Bayes devient linéaire en fonction des observations. Elle est alors appelée « prime de crédibilité ». Dans ce cas précis et souhaitable, la prime de Bayes d'un assuré se simplifie et s'écrit alors :

$$P^{Bayes} = \widetilde{\mu(\Theta)} = \alpha\overline{X} + (1 - \alpha)P^{Coll}$$

avec :

- $\alpha = \frac{n}{n+k}$ le facteur de crédibilité accordé à l'assuré ;
- n le nombre d'années d'exposition ;
- k le coefficient de crédibilité qui dépend des lois a priori et a posteriori de Θ ;
- $\overline{X}$ l'estimation de la prime individuelle de l'assurée ;
- P^{Coll} l'estimation de la prime collective.

Il est dès lors très important de noter le point suivant.

La prime de Bayes est la meilleure prime qui puisse être réclamée à un assuré compte tenu de son profil de risque et de son historique de sinistres. Toutefois, pour avoir une forme analytique simple comme explicité plus haut, il est nécessaire de connaître les lois a priori et a posteriori de Θ ce qui n'est en pratique presque jamais le cas.

Ainsi des modèles plus simples sont développés avec une application concrète plus facilement réalisables.

3.3 MODELE DE BÜHLMANN

Afin de déterminer la meilleur prime individuelle $\mu(\Theta)$ d'un assuré de profil de risque $\theta \in \Theta$ inconnu de l'assureur, c'est-à-dire la prime la plus compétitive et la plus équitable possible, l'étude « A course in Credibility Theory and its Applications » (Hans Bühlmann, 2005) propose de se restreindre au cas des estimateurs linéaires en les observations $X = (X_1, \dots, X_n)$ de l'assuré.

3.3.1 Hypothèses et formalisme

- Les observations de sinistres d'un assuré $(X_j)_{j=1\dots n}$ sont indépendantes conditionnellement à $\Theta = \theta$, de même distribution et de fonction de répartition F_θ munies des deux premiers moments :

$$\mu(\theta) = \mathbb{E}[X_j \mid \Theta = \theta]$$

$$\sigma^2(\theta) = \mathbb{V}[X_j \mid \Theta = \theta]$$

- Θ est une variable aléatoire de distribution $U(\theta)$;

- La formule de décomposition de la variance donne :

$$\mathbb{V}[X_j] = \mathbb{E}(\mathbb{V}[X_j \mid \Theta = \theta]) + \mathbb{V}(\mathbb{E}[X_j \mid \Theta = \theta]) = \mathbb{E}(\sigma^2(\theta)) + \mathbb{V}(\mu(\theta))$$

d'ou : $\mathbb{V}[X_j] = \sigma^2 + \tau^2$ avec σ^2 la mesure de volatilité interne et τ^2 la mesure de l'hétérogénéité du portefeuille. Σ^2 et τ^2 sont également appelés paramètres de structure du portefeuille.

3.3.2 Problème d'optimisation

L'estimateur $\mu(\Theta)$ linéaire s'écrit alors de la manière suivante :

$$\widehat{\widehat{\mu(\Theta)}} = \widehat{a_0} + \sum_{j=1}^{n} \widehat{a_j} X_j$$

où les coefficients $\widehat{a_0}, \widehat{a_1}, \dots, \widehat{a_n}$ sont solutions du problème d'optimisation suivant :

$$\min_{\widehat{a_0}, \widehat{a_1}, \dots, \widehat{a_n} \in \mathbb{R}} \mathbb{E}\left[\left(\mu(\Theta) - \widehat{\widehat{\mu(\Theta)}}\right)^2\right] = \min_{\widehat{a_0}, \widehat{a_1}, \dots, \widehat{a_n} \in \mathbb{R}} \mathbb{E}\left[\left(\mu(\Theta) - \widehat{a_0} - \sum_{j=1}^{n} \widehat{a_j} X_j\right)^2\right]$$

La distribution des $(X_j)_{j=1\dots n}$ est invariante par permutation car ces derniers sont indépendants et identiquement distribués conditionnellement à Θ par hypothèse. Il en résulte que $\widehat{a_1} = \dots = \widehat{a_n}$ et le modèle se réécrit tel que :

$$\widehat{\widehat{\mu(\Theta)}} = \hat{a} + \hat{b}\overline{X}, \quad avec\ \overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$$

d'où :

$$(\hat{a}, \hat{b}) = \operatorname{argmin}_{(\hat{a}, \hat{b}) \in R} \mathrm{E}\left[\left(\mu(\Theta) - \hat{a} - \hat{b}\overline{X}\right)^2\right]$$

Ainsi après calculs :

$$\hat{b} = \frac{n}{n + \left(\frac{\sigma}{\tau}\right)^2} \quad \text{et} \quad \hat{a} = (1 - \hat{b})\mu_0$$

avec $\tau^2 = \mathbb{V}(\mu(\Theta))$ l'hétérogénéité du portefeuille et $\sigma^2 = \mathbb{E}[\sigma^2(\Theta)]$ sa volatilité interne.

L'estimateur de crédibilité sous les hypothèses du modèle est donné par :

$$\widehat{\widehat{\boldsymbol{\mu(\Theta)}}} = \boldsymbol{\alpha\overline{X} + (1 - \alpha)\mu_0}$$

avec $\mu_0 = \mathbb{E}[\mu(\Theta)]$ l'estimateur de la prime collective et $\alpha = \frac{n}{n + \left(\frac{\sigma}{\tau}\right)^2}$ le facteur de crédibilité accordé à l'assuré.

Quelques remarques :

- $P^{Cred} = \widehat{\widehat{\mu(\Theta)}}$ est une moyenne pondérée de la prime collective P^{Coll} et de la prime individuelle $\overline{X}$, cette décomposition a déjà été rencontrée dans le modèle de fluctuation limitée ;
- $\left(\frac{\sigma}{\tau}\right)^2$ est le coefficient de crédibilité, il représente le quotient de l'écart type espéré du risque interne au portefeuille σ^2 par sa mesure d'hétérogénéité τ^2 ;
- α est le facteur de crédibilité, croît avec le nombre d'années d'observations n, croît avec l'hétérogénéité du portefeuille et décroît avec la variabilité interne au risque.
- Si P^{Bayes} coïncide avec la prime P^{Cred} alors il est question de crédibilité exacte.

3.3.3 Interprétation générale de la prime de crédibilité

P^{Cred} est une moyenne pondérée de P^{Coll} et de $\overline{X}$

$P^{Coll} = \mu_0$ est le meilleur estimateur basé a priori de l'assuré. Son erreur quadratique vaut :

$$\mathbb{E}\left[(\mu_0 - \mu(\Theta))^2\right] = \mathbb{V}(\mu(\Theta)) = \tau^2$$

Les informations a priori contiennent donc les informations relatives au collectif et le meilleur estimateur qu'il est possible d'en tirer est μ_0 : l'erreur est τ^2 est ainsi commise.

$\overline{X}$ est le meilleur estimateur linéaire et individuellement non biaisé basé sur l'expérience individuelle. Son erreur quadratique vaut :

$$\mathbb{E}\left[(\overline{X} - \mu(\Theta))^2\right] = \mathbb{E}\left[\frac{\mathbb{E}[\sigma^2(\Theta)]}{n}\right] = \frac{\sigma^2}{n}$$

Les observations contiennent elles les informations sur le risque individuel et le profil de risque associé Θ. Le meilleur estimateur que l'on peut en tirer est $\overline{X}$; et l'erreur commise en prenant $\widehat{\widehat{\mu(\Theta)}} = \overline{X}$ est $\frac{\sigma^2}{n}$.

Finalement, la prime de crédibilité P^{Cred} est pondérée par ces deux estimateurs qui proviennent de deux sources d'information différentes. La valeur de ces pondérations est donnée par le facteur de crédibilité α, qui explicite mathématiquement la proportion de chaque estimateur μ_0 et $\overline{X}$ nécessaire pour calculer la meilleure prime possible compte tenu des observations de l'assuré.

L'erreur quadratique de l'estimateur de crédibilité vaut :

$$\mathbb{E}\left[\left(\widehat{\widehat{\boldsymbol{\mu(\Theta)}}} - \boldsymbol{\mu(\Theta)}\right)^2\right] = \boldsymbol{(1 - \alpha)\tau^2} = \boldsymbol{\alpha\frac{\sigma^2}{n}}$$

L'erreur quadratique de $\mathrm{P}^{\mathrm{Cred}}$, qui s'exprime en fonction des erreurs quadratique de $\mathrm{P}^{\mathrm{Coll}}$ et $\mathrm{P}^{\mathrm{Ind}}$ respectivement, est toujours inférieure à ces deux dernières puisque le facteur de crédibilité $\alpha \in [0,1]$. Son principal avantage par rapport à $\mathrm{P}^{\mathrm{Bayes}}$ est son écriture analytique simple. Toutefois, l'expression de $\mathrm{P}^{\mathrm{Cred}}$ explicitée précédemment ne dépend que des observations d'un assuré. Or dans la réalité, un modèle de tarification se construit à partir d'un portefeuille. Le modèle suivant vient donc solutionner ce problème.

3.3.4 Modèle de Bühlmann Simple

Soient des observations de sinistres relatives à l'ensemble d'un portefeuille d'assurés de risques similaires $i = (1, \dots, \mathrm{I})$. Il est défini alors le vecteur d'observation d'un assuré i comme $\mathrm{X_i} = (\mathrm{X_{i1}}, \dots, \mathrm{X_{in}})$ et Θ_i son profil de risque. Pour chaque assuré i de ce portefeuille homogène a priori, $\mathrm{X_i}$ et Θ_i suivent les mêmes hypothèses que la partie précédente :

- Les $\mathrm{X_{ij}}$ sont indépendants conditionnellement à $\Theta_\mathrm{i} = \theta$, de même distribution $\mathrm{F_\theta}$ dont les deux premiers moments sont :

$$\mu(\theta) = \mathbb{E}\left[\mathrm{X_{ij}} \mid \Theta_\mathrm{i} = \theta\right]$$

$$\sigma^2(\theta) = \mathbb{V}\left[\mathrm{X_{ij}} \mid \Theta_\mathrm{i} = \theta\right]$$

- Les couples de variables aléatoires $(\Theta_1, \mathrm{X_1}), \dots, (\Theta_\mathrm{n}, \mathrm{X_n})$ sont indépendants et identiquement distribués ;

Comme énoncé plut tôt, il n'est plus question ici de l'estimateur $\widehat{\widehat{\mu(\Theta)}}$ mais bien de l'estimateur $\widehat{\widehat{\mu(\Theta_\mathrm{i})}}$ qui dépend de toutes les observations du portefeuille.

A partir du même problème d'optimisation sous contrainte, une démonstration analogue à la partie précédente permet d'expliciter de nouveaux estimateurs.

Les estimateurs non homogène et homogène de crédibilité du modèle de Bühlmann Simple sont donnés par :

$$\boldsymbol{\widehat{\widehat{\mu(\Theta_i)}} = \alpha\bar{X}_i + (1 - \alpha)\mu_0}$$

$$\boldsymbol{\widehat{\widehat{\mu(\Theta_i)}}^{hom} = \alpha\bar{X}_i + (1 - \alpha)\bar{X}}$$

Avec

$\mu_0 = \mathbb{E}[\mu(\Theta)]$ l'estimateur de la prime collective

$\alpha = \frac{\mathrm{n}}{\mathrm{n} + \left(\frac{\sigma}{\tau}\right)^2}$ le facteur de crédibilité

$\bar{\mathrm{X}}_\mathrm{i} = \frac{1}{\mathrm{n}}\sum_{\mathrm{j}=1}^{\mathrm{n}} \mathrm{X_{ij}}$ L'estimateur de la prime individuelle du risque i.

La différence théorique entre l'estimateur homogène et non homogène vient du fait que l'estimateur homogène a une contrainte supplémentaire dans le problème d'optimisation. Il lui est imposé d'être non biaisé c'est-à-dire que $\mathbb{E}\left[\widehat{\widehat{\mu(\Theta_\mathrm{i})}}\right] = \mathbb{E}[\mu(\Theta_\mathrm{i})] = \mu_0$. $\bar{\mathrm{X}}$ la moyenne de sinistres sur l'ensemble du portefeuille est alors substituée à μ_0, le meilleur estimateur de μ_0, avec :

$$\bar{X} = \frac{1}{I}\sum_{i=1}^{I}\frac{1}{n}\sum_{j=1}^{n}X_{ij}$$

À ce niveau théorique, plusieurs éléments relatifs aux estimateurs de crédibilité sont à expliciter :

- la quantité à estimer est $\mu(\Theta_i) = \mathbb{E}[X_{in+1} \mid \Theta_i]$ la prime individuelle associée au risque i ;
- la statistique sous-jacente sera X le vecteur d'observations du risque étudié issu d'un portefeuille homogène a priori ;
- les estimateurs de crédibilité sont toujours les meilleurs dans une classe donnée a priori ;
- les paramètres d'hétérogénéité du portefeuille τ^2 et de variance interne du portefeuille σ^2 peuvent être déterminés soit par construction si les distributions a priori et a posteriori du portefeuille sont connues, soit empiriquement grâce aux formules suivantes :

$$\widehat{\sigma^2} = \frac{1}{I(n-1)}\sum_{i=1}^{I}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2$$

$$\widehat{\tau^2} = \frac{1}{I-1}\sum_{i=1}^{I}(\bar{X}_i - \bar{X}) - \frac{\sigma^2}{n}$$

AXA souhaite maintenant tarifer les stades de Saint Etienne, PSG et de Marseille (TE904 : Complexe sportif) contre les incendies causés par les fumigènes à l'aide d'un modèle de Bühlmann Simple. Leurs historiques de sinistres respectifs couvrent 6 saisons. Les montants de sinistres sont les suivants :

	2016 - 2017	2017 - 2018	2018 - 2019	2019 - 2020	2020 - 2021	2021 - 2022	Total
PSG	0	1 000	2 000	1 000	2 000	0	6 000
Marseille	3 000	4 000	2 000	2 000	2 000	5 000	18 000
ASSE	3 000	3 000	2 000	1 000	2 000	1 000	12 000

Tableau 8 - Tableau sinistres - Bühlman Simple

Le but est de définir quelle prime sera réclamée à ces deux contrats pour la saison à venir. En utilisant le modèle de Bühlmann Simple :

$$\widehat{\mu(\Theta_i)}^{hom} = \alpha\bar{X}_i + 1-\alpha X$$

pour $i \in \{\text{PSG}, \text{OM}, \text{ASSE}\}$ puisqu'il y a seulement 3 risques. D'après les formules empiriques :

- $\bar{X}_{PSG} = \frac{\text{Total Sinitres PSG}}{\text{Nombre de saison}} = 6\,000/6 = 1000\ \bar{X}_{OM} = 3\,000\ et\ \bar{X}_{ASSE} = 2\,000$
- $\bar{X} = \frac{\bar{X}_{OM}+\bar{X}_{ASSE}+\bar{X}_{PSG}}{\text{Nombre d'équipe}} = 2000$
- $\sigma^2 = \frac{1}{3(6-1)}\sum_{i=1}^{3}\sum_{j=1}^{6}(X_{ij} - \bar{X}_i)^2 = \frac{3\,200\,000}{3}$
- $\tau^2 = \frac{1}{3-1}\sum_{i=1}^{3}(\bar{X}_i - \bar{X}) - \frac{\sigma^2}{3} = \frac{7\,400\,000}{9}$
- $\alpha = \frac{n}{n + \left(\frac{\sigma}{\tau}\right)^2} = \frac{5}{6}$

Les primes de crédibilité homogènes correspondantes à la saison pour Paris, Saint Etienne et Marseille peuvent être ainsi déduites selon :

$$\widehat{\widehat{\boldsymbol{\mu(\Theta_{PSG})}}}^{\mathbf{hom}} = \boldsymbol{\alpha\bar{X}_{PSG} + (1 - \alpha)\bar{X} = 1\,178}$$

$$\widehat{\widehat{\boldsymbol{\mu(\Theta_{OM})}}}^{\mathbf{hom}} = \boldsymbol{\alpha\bar{X}_{OM} + (1 - \alpha)\bar{X} = 2\,822}$$

$$\widehat{\widehat{\boldsymbol{\mu(\Theta_{ASSE})}}}^{\mathbf{hom}} = \boldsymbol{\alpha\bar{X}_{ASSE} + (1 - \alpha)\bar{X} = 2\,000}$$

3.4 MODELE DE BÜHLMANN-STRAUB

Le modèle de Bühlmann-Straub approfondit le modèle de Bühlmann en introduisant la notion de poids très importante en assurance.

L'idée de normaliser et d'introduire un poids dans le modèle est de pouvoir comparer l'expérience de sinistres de différents assurés afin qu'ils soient du même ordre de grandeur

Le but de cette nouvelle variable est d'aider à différencier un peu mieux chaque assuré du portefeuille. En effet, dans le modèle de Bühlmann Simple, aucune distinction n'est faite entre les différents assurés relativement à l'exposition au risque. Il en résulte que α le facteur de crédibilité est le même pour tout le portefeuille. Cette exposition au risque peut varier grandement d'un risque à l'autre selon plusieurs critères.

Pour remédier à cela, la variance des observations $\sigma^2(\Theta_i) = \mathbb{V}[X_{ij} \mid \Theta_i]$ doit dépendre du volume de risque du contrat i. En introduisant donc la variable poids w_{ij} , les hypothèses du modèle sont :

- Les X_{ij} sont indépendants conditionnellement à $\Theta_i = \theta$, de même distribution F_θ dont les deux premiers moments sont :

$$\mu(\Theta_i) = \mathbb{E}[X_{ij} \mid \Theta_i]$$

$$\sigma^2(\Theta_i) = w_{ij}\mathbb{V}[X_{ij} \mid \Theta_i] \text{ , avec } w_{ij} \text{ le poids}$$

- Les couples de variables aléatoires $(\Theta_1, X_1), \dots, (\Theta_n, X_n)$ sont indépendants et identiquement distribués.

L'objectif est toujours le même : trouver la meilleure prime linéaire en les observations, mais cette fois-ci avec l'information additionnelle donnée par les poids w_{ij}.

Une autre démonstration analogue aux précédentes permet une nouvelle fois de calculer les estimateurs de crédibilité sous le modèle de Bühlmann-Straub :

Les estimateurs non homogène et homogène de crédibilité du modèle de Bühlmann Straub sont donnés par :

$$\widehat{\widehat{\boldsymbol{\mu(\Theta_i)}}} = \boldsymbol{\alpha_i\bar{X}_i + (1 - \alpha_i)\mu_0}$$

$$\widehat{\widehat{\boldsymbol{\mu(\Theta_i)}}}^{\mathbf{hom}} = \boldsymbol{\alpha_i\bar{X}_i + (1 - \alpha_i)\bar{X}}$$

avec :

- $\mu_0 = \mathbb{E}[\mu(\Theta)]$ l'estimateur de la prime collective ;

- $\alpha_i = \frac{w_i^{\blacksquare}}{w_i^{\blacksquare} + \left(\frac{\sigma}{\tau}\right)^2}$ le facteur de crédibilité à l'assuré i;
- $\bar{X}_i = \frac{1}{w_i^{\blacksquare}} \sum_{j=1}^{n} w_{ij} X_{ij}$ la prime individuelle ;
- $X_{ij} = \frac{Y_{ij}}{W_{ij}}$ la sinistralité normalisée de l'assuré i pour sa période d'assurance tt.
-
- $w_i^{\blacksquare} = \sum_{i=1}^{n} w_{ij}$ la somme totale des poids d'un contrat ;
- $\bar{X} = \frac{1}{\alpha_{\blacksquare}} \sum_i \alpha_i \bar{X}_i$ la prime collective sur le portefeuille ;
- $\alpha_{\blacksquare} = \sum_i \alpha_i$ la somme totale des facteurs de crédibilité du portefeuille ;
- $w_{\blacksquare\blacksquare} = \sum_i w_i^{\blacksquare}$, la somme des poids totale.

Le facteur de crédibilité α_i dépend maintenant de l'exposition au risque du contrat i. Mathématiquement, plus le poids w_{ij} augmente, plus l'exposition au risque augmente également, et plus de confiance sera accordée à l'exposition propre de l'assuré i. Et de même que dans le modèle de Bühlmann-Simple, plus le coefficient de crédibilité $k = \left(\frac{\sigma}{\tau}\right)^2$ diminue, plus le facteur de crédibilité α_i va augmenter. Ce qui, encore une fois, traduit le fait que plus de confiance sera accordée aux assurés si la variance interne du portefeuille est petite et son hétérogénéité est grande.

L'erreur quadratique de l'estimateur inhomogène de crédibilité vaut :

$$\mathbb{E}\left[\left(\widehat{\boldsymbol{\mu}(\boldsymbol{\Theta}_i)} - \boldsymbol{\mu}(\boldsymbol{\Theta}_i)\right)^2\right] = (\mathbf{1} - \boldsymbol{\alpha}_i)\boldsymbol{\tau}^2 = \boldsymbol{\alpha}_i \frac{\boldsymbol{\sigma}^2}{\boldsymbol{\omega}_i^{\blacksquare}}$$

À noter que :

- τ^2 est l'erreur quadratique de la prime collective μ_0. En utilisant l'estimateur de crédibilité, cela réduit cette erreur de $(1 - \alpha_i)$;
- $\frac{\sigma^2}{\omega_i^{\blacksquare}}$ est l'erreur quadratique du vecteur d'observations $\bar{X}_i$, qui est le meilleur estimateur linéaire basé sur les données d'expérience. Lorsque l'estimateur de crédibilité est utilisé à sa place, l'erreur est réduite de α_i.

L'erreur quadratique de l'estimateur homogène de crédibilité vaut :

$$\mathbb{E}\left[\left(\widehat{\boldsymbol{\mu}(\boldsymbol{\Theta}_i)}^{\mathbf{hom}} - \boldsymbol{\mu}(\boldsymbol{\Theta}_i)\right)^2\right] = (\mathbf{1} - \boldsymbol{\alpha}_i)\left(\mathbf{1} + \frac{\mathbf{1} - \boldsymbol{\alpha}_i}{\boldsymbol{\alpha}_{\blacksquare}}\right)\boldsymbol{\tau}^2$$

Finalement, pour déterminer l'estimateur de crédibilité de Bühlmann-Straub homogène, il est nécessaire de définir correctement les observations Y_{ij}, les poids w_{ij}, de calculer les variables de Bühlmann $X_{ij} = \frac{Y_{ij}}{w_{ij}}$, $\bar{X}_i$, $\bar{X}$ et les paramètres de structures μ_0, σ^2, et τ^2. Ces derniers étant déterminés :

- soit par les connaissances a priori et a posteriori des profils de risques, c'est alors une procédure bayésienne comme vue précédemment ;

- soit par les observations du collectif, c'est alors une procédure empirique mais avec cette fois-ci les formules suivantes :

$$\widehat{\sigma^2} = \frac{1}{I(n-1)} \sum_{i=1}^{I} \sum_{j=1}^{n} w_{ij} \left(X_{ij} - \overline{X}_i\right)^2$$

$$\widehat{\tau^2} = \max(0, T)$$

avec :

$$T = c * \left\{ \frac{I}{I-1} \sum_{i=1}^{I} \frac{w_i^{\blacksquare}}{w_{\blacksquare\blacksquare}} (\overline{X}_i - \overline{X})^2 - \frac{I\widehat{\sigma^2}}{w_{\blacksquare\blacksquare}} \right\}$$

$$\text{et } c = \frac{I-1}{I} * \left\{ \sum_{i=1}^{I} \frac{w_i^{\blacksquare}}{w_{\blacksquare\blacksquare}} \left(1 - \frac{w_i^{\blacksquare}}{w_{\blacksquare\blacksquare}}\right) \right\}^{-1}$$

Une autre notation courante est utilisée:

$$T = \frac{w_{\blacksquare\blacksquare}}{w_{\blacksquare\blacksquare}^2 - \sum_{i=1}^{I} w_{i\blacksquare}^2} \left(\sum_{i=1}^{I} w_{i\blacksquare} (\overline{X}_i - \overline{X})^2 - (I-1) \times \overline{\sigma^2} \right)$$

<u>Plusieurs remarques</u> :

1. $c = 1$ si $w_{1\blacksquare} = w_{2\blacksquare} = \cdots = w_{I\blacksquare}$ et $c > 1$ sinon ;

2. Il faut faire attention à la disposition du facteur $\frac{1}{n-1}$ dans la formule de $\widehat{\sigma^2}$ qui n'est valable que si tous les assurés disposent des observations Y_{ij} pour tout $j \in \{1, \ldots, n\}$. En pratique, et c'est le cas dans le présent mémoire, le nombre d'années d'exposition n varie d'un assuré à l'autre. En conséquence, la formule précédente se réécrie :

$$\widehat{\sigma^2} = \frac{1}{I} \sum_{i=1}^{I} \frac{1}{n_i - 1} \sum_{j=1}^{n} w_{ij} \left(X_{ij} - \overline{X}_i\right)^2$$

3. Il est également important de faire la différence entre le $\overline{X}$ qui apparait dans l'expression de T et celui dans l'estimateur homogène de crédibilité. En effet, dans l'expression de T, les facteurs de crédibilités ne sont pas encore déterminés, et donc :

$$\overline{X} = \frac{1}{w_{\blacksquare\blacksquare}} \sum_{i} w_i^{\blacksquare} \overline{X}_i$$

Mais dans l'expression de $\widehat{\overline{\mu(\Theta_i)}}^{hom}$, $\overline{X} = \frac{1}{\alpha_{\blacksquare}} \sum_i \alpha_i \overline{X}_i$. Cela vient du fait que dans le modèle de Bühlmann-Straub, le meilleur estimateur de μ_0 n'est pas la moyenne observée du portefeuille, mais bien la moyenne pondérée de crédibilité.

4. Lorsque l'on utilise l'approche empirique pour déterminer les coefficients $\widehat{\sigma^2}$ et $\widehat{\tau^2}$, l'estimateur est appelé alors estimateur homogène empirique de Bühlmann-Straub.

L'estimateur homogène de crédibilité de Bühlmann-Straub doit vérifier la propriété « d'équilibre » :

$$\sum_{ij} \mathbf{w_{ij}} \widehat{\mu(\Theta_i)}^{\mathbf{hom}} = \sum_{ij} \mathbf{w_{ij}} \mathbf{X_{ij}}$$

avec à gauche le terme relatif à la prime de crédibilité du portefeuille sur la durée d'exposition totale, et à droite la somme des montants de sinistres agrégés sur cette même période.

Cette propriété signifie simplement qu'un modèle de Bühlmann-Straub est correctement calibré sur un portefeuille de référence si la prime de crédibilité totale qui résulte des calculs et le montant total des sinistres pondérés sont égaux.

Il est recommandé de s'en servir pour vérifier que tout modèle de Bühlmann est correctement implémenté.

AXA qui souhaite maintenant tarifer les trois stades des équipes de Saint Etienne, PSG et $Marseille$ qui constituent son portefeuille avec le modèle de Bühlmann Straub. Elle dispose pour cela des sommes agrégées de sinistres annuelles pour chaque stade ainsi que le nombre de groupe de supporter au stade assuré, chaque stade représentant un risque différent :

	2019-2020	2020-2021	2021-2022	2022-2023
PSG (Y_{ij})	1 000	2 000	0	
groupe (w_{ij})	7	10	15	8
OM	2 000	2 000	5 000	
# groupe	12	10	8	10
Saint Etienne	1 000	2 000	1 000	
# groupe	5	6	6	6
Bordeaux	2 000	5 000	1 000	
# groupe	4	5	4	10

Tableau 9 - Tableau des poids et sinistres - Bühlman-Straub

L'idée est la suivante : plus une équipe possède de groupe de supporter à assurer, plus les probabilités qu'un sinistre ait lieu sont grandes, et donc plus ses données sont fiables. Par ailleurs, un modèle de Bühlmann n'est en soit pas généralisable. Il résulte du choix réfléchit de chaque variable pour les données d'observations et de chaque poids pour un contexte de tarification particulier.

Pour prendre en compte les poids dans l'exemple, il est nécessaire de travailler sur la variable $X_{ij} = \frac{somme\ agrégée}{poids} = \frac{Y_{ij}}{w_{ij}}$ (la prime pure).

La table des X_{ij} suivante est ainsi obtenue :

	2019-2020	2020-2021	2021-2022
PSG	143	200	-
Marseille	167	200	625
ASSE	200	333	167
FCGB	500	1 000	250

Tableau 10- X_{ij}- Bühlman-Straub

Pour calculer la prime de chaque contrat pour l'année 4, les estimateurs homogènes empiriques de Bühlmann-Straub pour tout $i \in \{1,2\}$ sur la variable X_{ij} sont calculés tel que :

$$\widehat{\mu(\Theta_i)}^{hom} = \alpha_i \bar{X}_{\ i} + (1 - \alpha_i)\bar{X}$$

D'après les formules empiriques, il vient alors :

$\bar{X}_{ij}$	2019-2020	2020-2021	2021-2022	$\bar{X}_i$
PSG	143	200	-	93,75
Marseille	167	200	625	300,00
ASSE	200	333	167	235,29
FCGB	500	1 000	250	615,38

w_{ij}	2019-2020	2020-2021	2021-2022	2022-2023	w_i	${w_i}^2$
PSG	7	10	15	8	32	1 024
Marseille	12	10	8	10	30	900
ASSE	5	6	6	6	17	289
FCGB	4	5	4	10	13	169

$$w^{\blacksquare\blacksquare} = {w_1}^{\blacksquare} + w2^{\blacksquare} + w2^{\blacksquare} = 92$$

$$\bar{X} = 260{,}87$$

$$\widehat{\sigma^2} = \frac{1}{4(4-1)} \sum_{i=1}^{4} \sum_{j=1}^{4} w_{ij} \left(X_{ij} - \bar{X}_i\right)^2 = 473\,170$$

$$c = \frac{4-1}{4} * \left\{ \sum_{i=1}^{4} \frac{w_i^{\blacksquare}}{92} \left(1 - \frac{w_i^{\blacksquare}}{92}\right) \right\}^{-1} = 1{,}04$$

$$\widehat{\tau^2} = 1{,}04 \times \left\{ \frac{4}{4-1} \sum_{i=1}^{3} \frac{w_i^{\blacksquare}}{92} (\bar{X}_i - 260{,}87)^2 - \frac{4 \text{ x } 473\,170}{92} \right\} = 17\,624$$

$$k = \frac{\widehat{\sigma^2}}{\widehat{\tau^2}} = 26{,}85$$

Ce qui permet de calculer les $\boldsymbol{\alpha_i}$

	α_i
PSG	54%
OM	53%
ASSE	39%
FCGB	33%

Avec :

$$\bar{X} = \frac{1}{\alpha_{PSG} + \alpha_{ASSE} + \alpha_{OM} + \alpha_{FCGB}} \sum_{i=1}^{3} \alpha_i X_i = 206$$

Les primes unitaires par groupe et par club sont données par le tableau suivant avec la prime globale de chaque club en fonction des poids respectifs.

	Unitaire	Globale
PSG	179	1 433
OM	291	2 909
ASSE	263	1 579
FCGB	390	3 899

Tableau 11 - Résultat de l'exemple BS

Approche itérative :

L'estimateur T peut être négatif. En pratique, le choix est $\widehat{\tau^2} = max(T, 0)$, qui est alors biaisé. Un estimateur alternatif du paramètre T est, quant à lui, toujours positif :

$$\alpha = \frac{1}{I-1}\sum_{i=1}^{I} z_i(\bar{X}_i - \bar{X})^2$$

Cet estimateur, dit de Bichsel–Straub, est en fait un pseudo-estimateur de Vylder explicité dans « Regression model with scalar credibility weights » (Regression model with scalar credibility weights, 1981) dans la mesure où il dépend de paramètres inconnus. Il est évalué itérativement par la méthode du point fixe. Il est alors possible de démontrer que si $T < 0$, alors α converge vers 0.

Enfin, il n'est pas difficile de vérifier que lorsque tous les poids sont égaux et que le nombre d'années d'expérience est le même pour tous les assurés, alors le modèle de Bühlmann–Straub est en tous points équivalent à celui de Bühlmann.

3.5 MODELE DE JEWELL OU CREDIBILITE HIERARCHIQUE.

Les bases de données en assurance présentent souvent des structures hiérarchiques. En effet, les risques peuvent se regrouper ou se décliner en plusieurs sous-groupes, dans lesquels ces mêmes risques se mesureront de manière plus « homogène » et affinée.

Il devient alors possible de déterminer les primes d'assurance à partir de structures dites en arborescence, avec des approches pyramidales pour calculer les primes redistribuées au fur et à mesure dans chaque niveau inférieur plus homogène.

Dans « Application du modèle de crédibilité hiérarchique à la modélisation des taux de mortalité de plusieurs populations » (Cary Chi-Liang Tsai), un modèle hiérarchique traditionnellement utilisé en assurance est construit pour modéliser les taux de mortalité de plusieurs populations. La figure suivante présente l'arborescence de leur modèle :

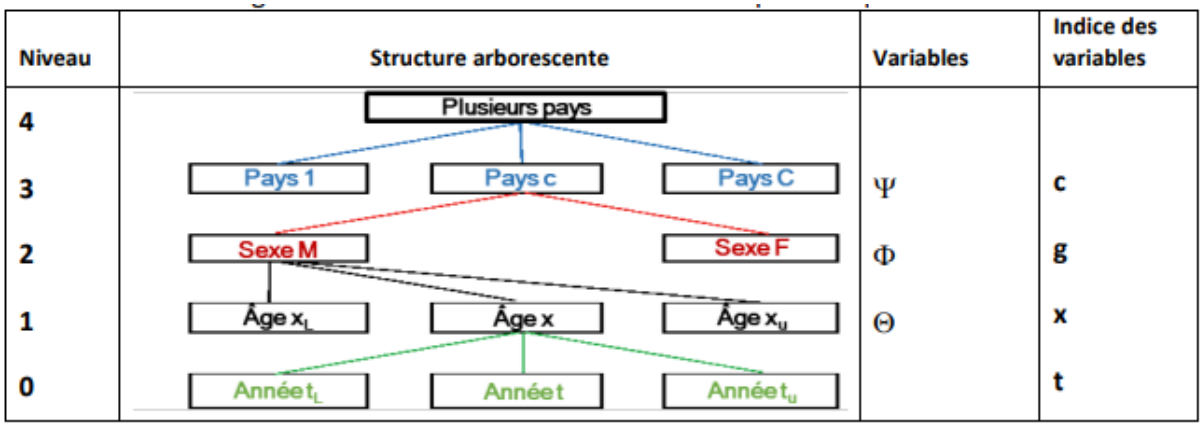


Figure 23 - Exemple de modelé hiérarchique - Tarification

Dans cette réprésentation, une structure à 4 niveaux est explicitée :

- Niveau 4 : le sommet de l'arbre, il correspond au niveau d'agrégation le plus haut du portefeuille ;
- Niveau 3 : chaque variable $\psi_1, \dots, \psi_p$ décrit le risque au niveau d'agrégation d'un seul pays ;
- Niveau 2 : chaque variable Φ_M, Φ_F décrit le risque au niveau d'agrégation d'un sexe, dans un seul pays ;
- Niveau 1 : chaque variable $x_1, \dots, x_l$ décrit le risque au niveau d'agrégation d'un âge, pour un sexe et un pays ;
- Niveau 0 : c'est le niveau d'agrégation minimal des données, là où l'on dispose des valeurs des données pour chacune des années constituant la période totale, pour un âge, un sexe, et un pays.

Ces modèles hiérarchiques présentent l'intérêt de répartir les données entre elles selon la charge de risque dans le collectif. Ainsi, si les différents niveaux sont judicieusement choisis en amont, les risques « homogènes » vont avoir tendance à être regroupés entre eux à mesure que l'on descend dans l'arborescence. La forme en arborescence peut paraitre complexe au premier abord, mais elle n'est en définitif qu'une généralisation des modèles de Bühlmann présentés dans les parties précédentes. En effet, tout modèle de Bühlmann est articulé selon une arborescence à deux niveaux :

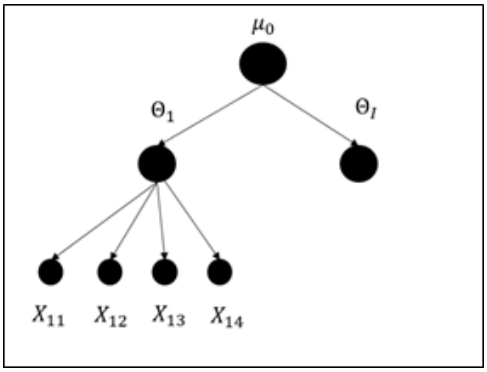


Figure 24- Arborescence des modèles de Bühlmann

Un modèle de crédibilité hiérarchique à quatre niveaux est présenté dans « A course in Credibility Theory and its Applications » (Hans Bühlmann, 2005) résume les travaux de Bühlmann & Jewell de 1987 et permet de fixer les hypothèses et fondements préalables au calcul des estimateurs de crédibilité hiérarchique. Il est important de noter que ces modèles se généralisent finalement au nombre de niveaux souhaités. Toutefois, à partir d'un certain seuil, les calculs peuvent devenir très laborieux. Il convient donc de choisir ce seuil avec parcimonie. La figure suivante présente l'arborescence à quatre niveaux présentée dans « A course in Credibility Theory and its Applications. » (Hans Bühlmann, 2005) :

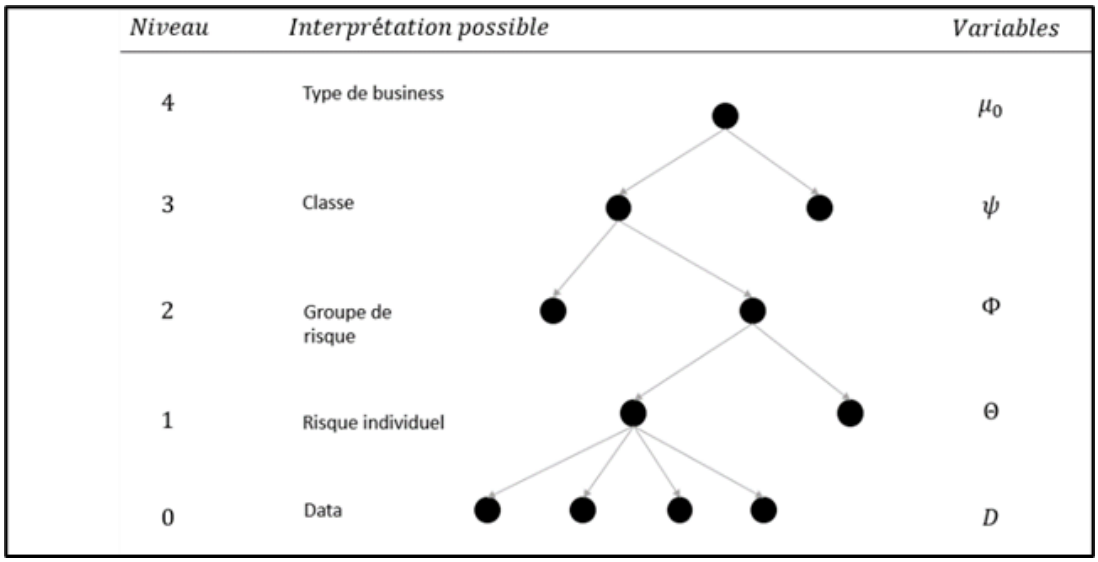


Figure 25 - Crédibilité hiérarchique générale à quatre niveaux

3.5.1 Hypothèses du modèle :

L'objectif du modèle hiérarchique est toujours le même : déterminer un estimateur de la prime individuelle correcte $\mu(\Theta_i)$ mais en utilisant différents niveaux d'agrégation. Pour cela, il est possible de définir les paramètres et variables nécessaires pour chaque niveau, au vu de l'étude la présentation est réduite à 3 niveaux :

- Niveau 3 (Portefeuille):

  $\mu_3 = \mathbb{E}[\mu_2(\Phi_k)] = \mathbb{E}[X_{ji}]$ est la prime collective.

  ${\sigma_3}^2 = \mathbb{V}[\mu_2(\Phi_k)] = \mathbb{E}\left\{\left[\mu_2(\Phi_k) - \mu_3\right]^2\right\}$ est le paramètre de structure du collectif.

- Niveau 2 (Classe k):

  Φ_k sont des variables indépendantes et identiquement distribuées ;

  $\mu_2(\Phi_k) = \mathbb{E}[\mu_1(\Theta_{j,k}) \mid \Phi_k] = \mathbb{E}[X_{ji} \mid \Phi_k]$ est la prime individuelle du niveau 2.

  ${\sigma_2}^2 = \mathbb{E}[{\sigma_2}^2(\Phi_k)] = \mathbb{E}\left[\mathbb{V}[\mu_1(\Theta_{j,k}) \mid \Phi_k]\right]$ est le paramètre de structure du niveau 2.

- Niveau 1 (Individu j) :

  $\Theta_{j,k} \mid \Phi_{k=1\ldots,K}$ sont des variables indépendantes et identiquement distribuées ;

  $\mu_1(\Theta_{J,k}) = \mathbb{E}[X_{i,j,k} \mid \Theta_{j,k}]$ est la prime individuelle du niveau 1.

  ${\sigma_1}^2 = \mathbb{E}[\sigma_1^2(\Theta_{j,k})] = \mathbb{E}\left[w_{ji}\mathbb{V}[X_{i,j,k} \mid \Theta_{jk}]\right]$, avec w_{ji} le poids connu

- Niveau 0 (data, par exemple à l'année i) :

  $X_{i,j,k} \mid \Theta_{j,k=(1,1)\ldots,(J,K)}$ sont des variables indépendantes et identiquement distribuées. Chaque $X_{i,j,k}$ représente une observation du risque pour les niveaux (j, k) et l'année i. Ce sont les données les plus granulaires disponibles.

A noter que d'après le théorème de la variance totale :

$$\mathbb{V}[\mu_1(\Theta_{j,k})] = \mathbb{E}\left[\mathbb{V}[\mu_1(\Theta_{j,k}) \mid \Phi_k]\right] + \mathbb{V}\left[\mathbb{E}[\mu_1(\Theta_{j,k}) \mid \Phi_k]\right]$$

$$= {\sigma_2}^2 + \mathbb{V}[\mu_2(\Phi_k)]$$

$$= {\sigma_2}^2 + {\sigma_3}^2$$

Le modèle est soumis aux hypothèses suivantes :

H0 : les v.a. $\mu_2(\Phi_k), \mu_1(\Theta_{j,k}), X_{ijk}$ sont de carré intégrable.

H1 : les classes sont indépendantes deux à deux. C'est-à-dire que pour $k \neq p$, les suites de v.a. $(\Phi_k, \Theta_{jk}, X_{ijk}, j = 1, \dots, n_p ; i \geq 1)$et $(\Phi_k, \Theta_{jk}, X_{ijk}, j = 1, \dots, n_p; i \geq 1)$ sont indépendantes en probabilité.

H2 : Pour k ∈ K, conditionnellement à Φ_k, les individus (k,1), …, (k,n_k) sont indépendants. C'est-à-dire que les n_k suites de v.a. $(\Theta_{jk}, X_{ijk}, j = 1, \dots, n_k ; i \geq 1)$ sont indépendantes en probabilité.

H3 : Pour k ∈ K, et j fixé dans $\{1, \dots, n_k\}$, conditionnellement à $\Theta_{j,k}$,les v.a. X_{ijk} ; $i \geq 1$ sont indépendantes en probabilité.

H4 : Les couples de variables aléatoires $(\Phi_k, \Theta_{j,k})$ ont tous la même loi de probabilité.

H5 : $\mathrm{E}(X_{kji} \mid \Phi_k \vartheta_{kj}) = \mu(\Phi_k, \vartheta_{kj})$

$\mathrm{Cov}(X_{i,j,k}, X_{l,j,k} \mid \Phi_k \vartheta_{kj}) = \frac{\sigma(\Phi_k \vartheta_{kj})}{\omega_{kji}} \delta_{il}$,

où δ_{il} est le symbole de Kronecker : $\delta_{il} = 0$ si $i \neq l$, $\delta_{il} = 1$ si $i = l$.

3.5.2 Estimation des paramètres

Dans le modèle hiérarchique de crédibilité, les données agrégées et les facteurs de crédibilités sont calculés des niveaux inférieurs vers les niveaux supérieurs (approche « bottom-up »). Ensuite, les estimateurs de crédibilités sont calculés des niveaux supérieurs vers les niveaux inférieurs (approche « Top-down») jusqu'à avoir les primes de Bühlmann $\widehat{\mu_1(\Theta_{j,k})}$ au niveau du risque individuel. Ainsi, pour avoir l'estimateur de crédibilité $\mu_1(\widehat{\Theta_{j,k}})_{1+1}$ correspondant aux niveaux (i, h) et à l'année $i + 1$, il est nécessaire de déterminer les estimateurs de crédibilités du niveau supérieur $\widehat{\mu_2(\Phi_h)}$

Le meilleur estimateur de crédibilité peut être calculé des niveaux inférieurs vers les niveaux supérieurs comme suit :

- X_{kji} est le taux de prime pure du niveau 0.
- $X_{kj\cdot} = \sum_i \frac{\omega_{kji}}{\omega_{kj\cdot}} X_{kji}$ est le taux de prime pure moyen du niveau 1
- $X_{k\cdot\cdot} = \sum_j \frac{\omega_{kj\cdot}}{\omega_{k\cdot\cdot}} X_{kj\cdot} = \sum_j \sum_i \frac{\omega_{kji}}{\omega_{k\cdot\cdot}} X_{kji}$ est le taux de prime pure du niveau 2.
- ω_{kji} est le poids du niveau 0.
- $\omega_{kj\cdot} = \sum_i \omega_{kji}$ est le poids niveau 1.

- $\omega_{k\cdot\cdot} = \sum_j \omega_{kj\cdot} = \sum_j \sum_i \omega_{kji}$ est le niveau 2.

- $\omega_{\cdots} = \sum_k \sum_j \sum_i \omega_{kji}$ est le poids du portefeuille (niveau 3).

- $z_{kj} = \frac{a\omega_{kj}}{S^2 + a\omega_{kj}}$ est le coefficient de crédibilité niveau 1. Le coefficient de crédibilité z_{kj} dépend de la variabilité des observations individuelles du niveau 0 dans le niveau 1 . Plus la variabilité est grande, plus le coefficient de crédibilité est faible.

- $z_{k\cdot} = \sum_j z_{kj}$

- $z_k = \frac{bz_{k\cdot}}{a + bz_{k\cdot}}$ est le coefficient de crédibilité au niveau 2.

- $z = \sum_k z_k$

- $\varepsilon_{k\cdot\cdot} = \sum_j \frac{z_{kj}}{z_{k\cdot}} X_{kj\cdot}$ est le taux de prime pure moyen au niveau 2 pondéré par les coefficients de crédibilité de chaque niveau 1.

- $\varepsilon = \sum_k \frac{z_k}{z} X_{k\cdot\cdot}$ est le taux de prime pure pour le portefeuille (niveau 3).

Les estimateurs hétérogènes de crédibilité hiérarchique peuvent être déterminés des niveaux supérieurs vers les niveaux inférieurs de l'arborescence comme suit :

$$\boldsymbol{\mu_2}\widehat{\widehat{(\boldsymbol{\Phi_k})}} = z_k \times \varepsilon_{k\cdot\cdot} + (\mathbf{1} - z_k)\boldsymbol{\mu_3}$$

$$\boldsymbol{\mu_1}\widehat{\widehat{(\boldsymbol{\Theta_{j,k}})}} = z_{kj} \times X_{kj\cdot} + (\mathbf{1} - z_{kj})\boldsymbol{\mu_2}\widehat{\widehat{(\boldsymbol{\Phi_{j,k}})}}$$

Une correspondance itérative entre $\mu_2\widehat{\widehat{(\boldsymbol{\Phi_k})}}$ et $\mu_1\widehat{\widehat{(\Theta_{j,k})}}$ s'explicite ainsi. De plus, si μ_0 est remplacé par ε, les estimateurs inhomogènes sont substitués par les estimateurs homogènes de crédibilité hiérarchique.

Les estimateurs homogènes de crédibilité hiérarchique doivent vérifier la propriété d'« équilibre » :

$$\sum_{jihg} \mathbf{w_{i,j,k}} \boldsymbol{\mu}\widehat{\widehat{(\boldsymbol{\Theta_{j,k}})}}^{\mathbf{hom}} = \sum_{jihg} \mathbf{w_{i,j,k}} \mathbf{X_{i,j,k}}$$

Ce résultat et son interprétation sont totalement analogues au modèle de Bühlmann-Straub.

Les erreurs quadratiques sont calculées des niveaux inférieurs vers les niveaux supérieurs. Les interprétations demeurent ainsi similaires à ce qui a été vu précédemment, mais l'on calcule ici une erreur quadratique par niveau.

3.5.3 Estimation des paramètres de structure

Comme dans le modèle de Bühlmann-Straub, à chaque niveau de l'arborescence se définit un paramètre de structure décrivant la variance du sous-groupe. Chaque paramètre de structure peut être définit soit de manière bayésienne si les lois dérivant les groupes sont connues, soit de manière empirique en utilisant les formules suivantes :

- $\widehat{\sigma_1^2} = \frac{1}{JK} \sum_{k=1}^{K} \sum_{j=1}^{J} \frac{1}{t_{kj} - 1} \sum_{i=1}^{T_{k,j}} w_{i,j,k} \left[X_{i,j,k} - X_{kj\cdot} \right]^2$

- $\widehat{\sigma_2^2} = \frac{1}{K}\sum_{k=1}^{K} \max\left\{0, \widehat{T_k}^{(1)}\right\}$

$$\widehat{\mathbf{T_k}}^{(1)} = \mathbf{c_k^{(1)}} \cdot \left\{ \frac{\mathbf{J}}{\mathbf{J-1}} \sum_{\mathbf{j=1}}^{\mathbf{J}} \frac{\boldsymbol{\omega}_{\mathbf{kj\cdot}}}{\boldsymbol{\omega}_{\mathbf{k\cdot\cdot}}} \left[\mathbf{X_{kj\cdot} - X_{k\cdot\cdot}}\right]^2 - \frac{\mathbf{J}\widehat{\boldsymbol{\sigma}_1^2}}{\boldsymbol{\omega}_{\mathbf{k\cdot\cdot}}} \right\}$$

$$\mathbf{c_k^{(1)}} = \frac{\mathbf{J-1}}{\mathbf{J}} \cdot \left\{ \sum_{\mathbf{j=1}}^{\mathbf{J}} \frac{\boldsymbol{\omega}_{\mathbf{kj\cdot}}}{\boldsymbol{\omega}_{\mathbf{k\cdot\cdot}}} \left[\mathbf{1} - \frac{\boldsymbol{\omega}_{\mathbf{kj\cdot}}}{\boldsymbol{\omega}_{\mathbf{k\cdot\cdot}}}\right] \right\}^{-1}$$

- $\widehat{\sigma_3^2} = \max\left\{0, \widehat{T^{(3)}}\right\}$

$$\widehat{\mathbf{T_k}}^{(2)} = \mathbf{c_k^{(2)}} \cdot \left\{ \frac{\mathbf{K}}{\mathbf{K-1}} \sum_{\mathbf{k=1}}^{\mathbf{K}} \frac{\mathbf{z_{k\cdot}}}{\mathbf{z_k}} \left[\boldsymbol{\varepsilon}_{\mathbf{k\cdot\cdot}} - \boldsymbol{\varepsilon}\right]^2 - \frac{\mathbf{K}\widehat{\boldsymbol{\sigma}_2^2}}{\mathbf{z_k}} \right\}$$

$$\mathbf{c_k^{(2)}} = \frac{\mathbf{K-1}}{\mathbf{K}} \cdot \left\{ \sum_{\mathbf{k=1}}^{\mathbf{K}} \frac{\mathbf{z_{k\cdot}}}{\mathbf{z_k}} \left[\mathbf{1} - \frac{\mathbf{z_k}}{\mathbf{z_{k\cdot}}}\right] \right\}^{-1}$$

Afin de rendre les formules plus attractives, une application simple permet de mieux appréhender la progression indicielle qui suit l'arborescence de l'arbre. Pour calculer les primes de crédibilité hiérarchique pour chaque risque des niveaux supérieurs aux niveaux inférieurs, il est nécessaire d'évaluer toutes les grandeurs et de tous les paramètres de structures empiriques.

AXA souhaite finalement tarifer les quatre stades des équipes de football de son nouveau portefeuille avec la méthode hiérarchique de $Jewell$.

En particulier, elle choisit d'utiliser un modèle à deux niveaux (pour faciliter les calculs) avec l'utilisation de l'information du niveau de football avec :

- **niveau 1** : Les équipes $\{PSG, OM, Saint\ Etienne, Bordeaux\}$ à tarifer ;
- **niveau 2** : Les Ligues (Ligue 1 pour Paris et Marseille, et Ligue 2 pour Bordeaux et Saint Etienne)

Il sera noté L_k la ligue avec $L_1 = Ligue$ et $L_2 = Ligue\ 2$, et $E_{j,k}$ les équipes $j \in \{PSG, OM, Saint\ Etienne, Bordeaux\}$ de la ligue $k \in \{1\ ;\ 2\}$.

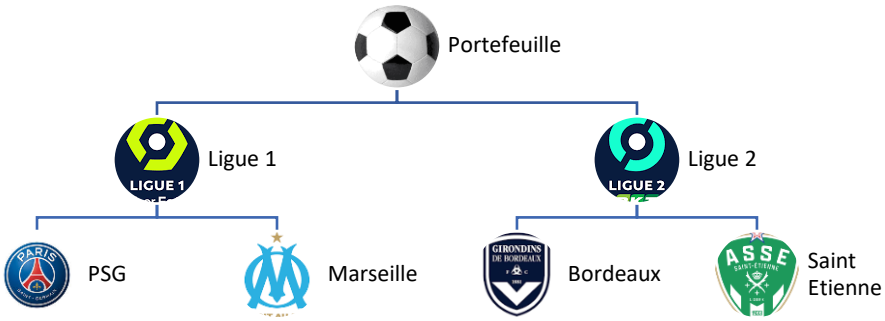


Figure 26 - Illustration de la hiérarchie de Jewell pour l'exemple

Pour effectuer une telle tarification, elle dispose des données suivantes, constituées des montants de sinistres agrégés et des poids qui leurs sont associés comme dans l'exemple de Bühlmann Straub :

		2019-2020	2020-2021	2021-2022	2022-2023
Ligue 1	PSG (Y_{ji})	1 000	2 000	0	
	# groupe (w_{ji})	7	10	15	8
	OM	2 000	2 000	5 000	
	# groupe	12	10	8	10
Ligue 2	Saint Etienne	1 000	2 000	1 000	
	# groupe	5	6	6	6
	Bordeaux	2 000	5 000	1 000	
	# groupe	4	5	4	10

Tableau 12 – Exemple de portefeuille- Jewell

Pour calculer les primes de chaque équipe pour la saison à venir, il est donc possible de construire les estimateurs de crédibilités empiriques des niveaux supérieurs vers les niveaux inférieurs comme explicité précédemment selon :

- la prime de crédibilité de la Ligue L_k :

$$\widehat{\gamma_2(L_r)} = z_k X_k^{(2)} + (1 - z_k)\widehat{\widehat{\gamma_3}}$$

- la prime de crédibilité de l'equipe $E_{r,i}$ appartenant à la Ligue r :

$$\widehat{\gamma_1(E_{r,i})} = z_{k,j} X_{k,j}^{(1)} + (1 - z_{k,j})\widehat{\gamma_2(L_r)}$$

$$= z_{k,j} X_{r,i}^{(1)} + (1 - z_{k,j}) z_k X_r^{(2)} + (1 - z_{k,j})(1 - z_k)\widehat{\widehat{\gamma_3}}$$

Comme dans l'exemple de Bühlmann Straub, il est nécessaire de calculer les quotients des variables d'observation Y_{ij} par leur poids respectif w_{ij} et de travailler ensuite sur la variable $X_{ji} = \frac{Y_{ji}}{w_{ji}}$ comme suit :

X_{ji}	2019-2020	2020-2021	2021-2022
PSG	143	200	0
OM	167	200	625
Saint Etienne	200	333	167
Bordeaux	500	1 000	250

Figure 27 - Tableau des poids de Jewell

S'en suit le calcul des variables de Bühlmann des niveaux inférieurs vers les niveaux supérieurs.

Les données agrégées du niveau 1 : $X_{k,j}^{(1)} = \sum_j \frac{w_{k,j,i}}{w_{k,j}^{(1)}} X_{k,j,i}$, soit au niveau de l'équipe j appartenant à la Ligue k, pour toutes les années i, avec $w_{k,j}^{(1)} = \sum_i w_{k,j,i}$ la somme des poids. Cela donne :

	$X_{k,j}^{(1)}$
PSG	93,75
OM	300,00
Saint Etienne	235,29
Bordeaux	615,38

Le paramètre de structure du niveau 1 : $\widehat{\sigma_1^2} = \frac{1}{JK} \sum_{k=1}^{K} \sum_{j=1}^{J} \frac{1}{t_{kj}-1} \sum_{i=1}^{T_{k,j}} w_{i,j,k} \left[X_{i,j,k} - X_{kj\cdot}\right]^2$

avec n_r le nombre d'équipe dans la Ligue r, n_i le nombre d'années d'exposition de l'équipe i appartenant à la Ligue k, et K le nombre de Ligue distinctes.

Dans l'exemple $t_{k,j} = n_{saison} = 3, J = n_{Equipe_ligue1} = n_{Equipe_ligue2} = 2, K = 2.$

Il vient :

$$\widehat{\sigma_1^2} = 354\,877{,}55$$

Le paramètre de structure du niveau 2 : $\widehat{\sigma_2^2} = \frac{1}{K} \sum_{r=1}^{2} \max\left\{0, \widehat{T_k}^{(1)}\right\}$

Avec les étapes suivantes :

$\omega_{kj\cdot} = \sum_i \omega_{kji}$	$w_{11\cdot} = 32$ $w_{21\cdot} = 17$	$w_{12\cdot} = 30$ $w_{22\cdot} = 13$
$\omega_{k\cdot\cdot} = \sum_j \omega_{kj\cdot}$	$\omega_{1\cdot\cdot} = 62$	$\omega_{2\cdot\cdot} = 30$
$X_{k\cdot\cdot} = \sum_j \frac{\omega_{kj\cdot}}{\omega_{k\cdot\cdot}} X_{kj\cdot} = \sum_j \sum_i \frac{\omega_{kji}}{\omega_{k\cdot\cdot}} X_{kji}$	$X_{1\cdot\cdot} = 193{,}55$	$X_{2\cdot\cdot} = 400$
$c_k^{(1)} = \frac{J-1}{J} \cdot \left\{ \sum_{i=1}^{J} \frac{\omega_{kj\cdot}}{\omega_{k\cdot\cdot}} \left[1 - \frac{\omega_{kj\cdot}}{\omega_{k\cdot\cdot}}\right] \right\}^{-1}$	$c_1^{(1)} = 1{,}00$	$c_2^{(1)} = 1{,}02$

$$\widehat{T_k}^{(1)} = c_k^{(1)} \cdot \left\{ \frac{J}{J-1} \sum_{j=1}^{J} \frac{\omega_{kj\cdot}}{\omega_{k\cdot\cdot}} \left[X_{kj\cdot} - X_{k\cdot\cdot} \right]^2 - \frac{J\widehat{\sigma_1^2}}{\omega_{k\cdot\cdot}} \right\} \qquad \widehat{T_1}^{(1)} = 9\,810 \qquad \widehat{T_2}^{(1)} = 48\,147$$

$$\widehat{\sigma_2^2} = \frac{1}{K} \sum_{k=1}^{2} \max\left\{0, \widehat{T_k}^{(1)}\right\} \qquad \widehat{\sigma_2^2} = \frac{\widehat{T_1}^{(1)} + \widehat{T_2}^{(1)}}{2} = 28\,979$$

Les coefficients de crédibilités du niveau 1 peuvent alors être déduits $z_{kj} = \frac{\widehat{\sigma_2^2}\omega_{kj}}{\widehat{\sigma_1^2} + \widehat{\sigma_2^2}\omega_{kj}}$:

$z_{1,1}$ PSG	0,72
$z_{1,2}$ OM	0,71
$z_{1,2}$ ASSE	0,58
$z_{2,2}$ FCGB	0,51

Tableau 13- Coefficients de crédibilités des équipes

Pour toutes les équipes $j \in J_k$ avec J_k l'ensemble des équipes appartenant à la ligue k.

Les données agrégées au niveau 2 des Ligues : $\varepsilon_{k\cdot\cdot} = \sum_j \frac{z_{kj}}{z_{k\cdot}} X_{kj\cdot}$

La somme des facteurs de crédibilité de la ligue k : $z_{k\cdot} = \sum_j z_{kj}$

$$\varepsilon_{1\cdot\cdot} = 185{,}93 \; et \; \varepsilon_{2\cdot\cdot} = 413{,}84$$

$$z_{1\cdot} = 1{,}43 \; et \; z_{2\cdot} = 1{,}10$$

Le paramètre de structure de l'ensemble du portefeuille : $\widehat{\sigma_3^2} = max\left\{0, \widehat{T}^{(2)}\right\}$

Avec les étapes suivantes :

$$z_{\cdot\cdot} = \sum_k z_{k\cdot} = 2{,}53$$

$$\varepsilon = \sum_k \frac{z_k}{z} X_{k\cdot\cdot} = 290{,}36$$

$$c^{(2)} = \frac{K-1}{K} \cdot \left\{ \sum_{k=1}^{K} \frac{z_{k\cdot}}{z_{\cdot\cdot}} \left[1 - \frac{z_{k\cdot}}{z_{\cdot\cdot}} \right] \right\}^{-1} = 1{,}02$$

$$\widehat{T}^{(2)} = c^{(2)} \cdot \left\{ \frac{K}{K-1} \sum_{k=1}^{K} \frac{z_{k\cdot}}{z_{\cdot\cdot}} \left[\varepsilon_{k\cdot\cdot} - \varepsilon \right]^2 - \frac{K\widehat{\sigma_2^2}}{z_{\cdot\cdot}} \right\} = 414{,}84$$

Les facteurs de crédibilité du niveau 2 $z_k = \frac{\widehat{\sigma_2^2} z_{k\cdot}}{\widehat{\sigma_2^2} + \widehat{\sigma_3^2} z_{k\cdot}}$, correspondant à la Ligue k tout entière :

$$z_1 = 0.25 \;\; z_2 = 0.24$$

❖ L'ajustement collectif du portefeuille $\varepsilon = \sum_k \frac{z_k}{z} X_{k..}$, autrement appelé moyenne de crédibilité, calculé sur la Ligue 1 et la Ligue 2, avec $z = \sum_k z_k$ la somme des deux facteurs de crédibilité.

Il reste finalement à construire les estimateurs de crédibilités empiriques. À partir des formules explicitées au début de l'exemple :

$$\widehat{\widehat{\gamma_3}} = 290$$

$$\widehat{\widehat{\gamma_2(L_1)}} = 288.71, \qquad \widehat{\widehat{\gamma_2(L_2)}} = 292.71$$

$$\widehat{\widehat{\gamma_1(PSG)}} = 147.71, \qquad \widehat{\widehat{\gamma_1(OM)}} = 296.73, \qquad \widehat{\widehat{\gamma_1(ASSE)}} = 259.26, \qquad \widehat{\widehat{\gamma_1(FCGB)}} = 458.77$$

Avant de conclure, il est possible de vérifier la propriété d'équilibre de Bühlmann permettant de s'assurer que le modèle a été correctement implémenté :

$$\sum_{i,j,k} w_{i,j,k} \widehat{\widehat{\gamma_1(E_{r,1})}} = 24\,000$$

$$\sum_{i,j,k} w_{i,j,k} X_{i,j,k} = 24\,000$$

Pour rappel, Cette propriété signifie simplement qu'un modèle de Bühlmann-Straub est correctement calibré sur un portefeuille de référence si la prime de crédibilité totale qui résulte des calculs et le montant total des sinistres pondérés sont égaux.

Finalement, il reste à déduire les primes réclamées aux quatre équipes du portefeuille en multipliant les $\widehat{\widehat{\gamma_1(E_{j,k,,4})}}$ par leur poids respectif de l'année 4 :

$$Y_{j,k,4} = \widehat{\widehat{\gamma_1(E_{j,k,,4})}} * w_{j,k,4}$$

	Prime saison 2022-2023
Y_{PSG}	1 181,68
Y_{OM}	2 967,28
Y_{ASSE}	1 411,76
Y_{FCGB}	4 587,73

Tableau 14- Prime finale par équipe pour la saison

Approche itérative

Dans la même logique que la méthode itérative explicité dans le modèle de Bülhmann-Straub, les estimateur $\widehat{\mathbf{T}}^{(\mathbf{h})}$, avec $\boldsymbol{h}$ indiquant le niveau, peuvent être négatifs. Il est donc pratique de passer par des estimateurs alternatifs dit pseudo-estimateur pour $\widehat{\sigma_2^2}$ et $\widehat{\sigma_3^2}$.

Par exemple dans le cas où le PSG aurait eu 1000 € de dégâts sur la saison 2021-2022 $\widehat{\sigma_3^2}$ aurait une valeur nulle. Ainsi pour l'étude la méthode utilisée sera l'approche itérative.

Pour chacune des rubriques d'entreprise (TRE), le portefeuille dispose d'observations pendant une durée de t années, avec $1 \leq t \leq 10$. Ainsi pour la rubrique d'entreprise (TRE) n°j du niveau hiérarchique n°k, il est observé les données $\mathrm{X_{kji}}, 1 \leq \mathrm{i} \leq \mathrm{t}$. Il n'est pas nécessaire de disposer pour chaque classe d'exactement t observations. Il peut donc être défini pour la rubrique n°j de la classe n°k la valeur T_{kj} telle que $\mathrm{i} \in \mathrm{T_{kj}} \subset \{1, \dots, \mathrm{t}\}$.

Les K classes sont différenciées par un paramètre aléatoire $\Phi_{\mathrm{g=1,\dots,k}}$. La caractéristique du risque au niveau de la classe n°k est $\mu_0(\Phi_{\mathrm{k}})$. Elle a pour fonction de structure U et pour moyenne $m = E\left[\mu_0\left(\Phi_g\right)\right]$.

Les différents rubriques TRE contenus dans la classe n°k sont différenciées par le paramètre aléatoire ϑ_{kj}. Ainsi il en ressort deux paramètres : l'un représentatif du niveau hiérarchique (classe), l'autre de la rubrique d'activité (TRE). La caractéristique du risque au niveau de la rubrique est donc : $\mu(\vartheta_{\mathrm{k}}, \vartheta_{\mathrm{kj}})$. Elle a pour fonction de structure U_2 qui est la loi conjointe de $(\vartheta_{\mathrm{k}}, \vartheta_{\mathrm{kj}})$.

Le rubrique d'entreprise (TRE) n°j de la classe n°k est donc décrit par : $\left(\vartheta_{\mathrm{kj}}\ ,\ \mathrm{X_{kji}}\ ;\ \mathrm{i} \geq 1\ ;\mathrm{j} = 1, \dots, \mathrm{n_k}\right)$. Le niveau hiérarchique n°k est lui décrit par : $(\vartheta_{\mathrm{k}}\ ,\ \vartheta_{\mathrm{kj}}\ ,\ \mathrm{X_{kji}}\ ,\mathrm{i} \geq 1\ ;\mathrm{j} = 1, \dots, \mathrm{n_k})$.

Pour résumer, il est défini :

- Au niveau du portefeuille : $\mathrm{E}\left(\mathrm{X_{kji}}\right) = \mathrm{m}$
- Au niveau de la classe n°k : $\mathrm{E}\left(\mathrm{X_{kji}}|\ \vartheta_{\mathrm{k}}\right) = \mu_0(\vartheta_{\mathrm{k}})$.
- Au niveau de la rubrique d'entreprise TRE n°j de la classe n°k : $\mathrm{E}\left(\mathrm{X_{kji}}|\ \vartheta_{\mathrm{k}}\vartheta_{\mathrm{kj}}\right) = \mu\left(\vartheta_{\mathrm{k}}, \vartheta_{\mathrm{kj}}\right)$.

A partir l'estimateur sans biais $\widehat{\sigma_1^2} = \frac{1}{\mathrm{n}}\sum_{\mathrm{k=1}}^{\mathrm{n}} \frac{1}{\mathrm{n_r}} \sum_{\mathrm{j=1}}^{\mathrm{n_k}} \frac{1}{\mathrm{n_i - 1}} \sum_{\mathrm{i=1}}^{T_{kj}} \mathrm{w_{r,i,j}} \left[\mathrm{X_{k,j,i}} - \mathrm{X_{k,j}^{(1)}}\right]^2$, il est possible de calculer les estimateurs proposés par Goovaerts pour respectivement $\widehat{\sigma_2^2}$ et $\widehat{\sigma_3^2}$:

- $\alpha = \frac{1}{\sum_{\mathrm{k=1}}^{\mathrm{n}}(\mathrm{n_k}-1)} \sum_{\mathrm{k=1}}^{\mathrm{n}} \sum_{\mathrm{j=1}}^{\mathrm{n_k}} \mathrm{z_{kj}} (\mathrm{X_{kj\cdot}} - \varepsilon_{\mathrm{k\cdot\cdot}})^2$ est un pseudo-estimateur sans biais de a.

- $\beta = \frac{1}{(\mathrm{n}-1)} \sum_{\mathrm{k=1}}^{\mathrm{n}} \mathrm{z_k} (\varepsilon_{\mathrm{k\cdot\cdot}} - \varepsilon)^2$ est un pseudo-estimateur sans biais de b

On obtient donc les formules de crédibilité suivantes :

- Pour le groupe de risque k : $\widehat{\boldsymbol{\mu}}(\boldsymbol{\vartheta_k}) = (\mathbf{1} - \hat{\mathbf{z}}_\mathbf{k})\hat{\boldsymbol{\varepsilon}} + \hat{\mathbf{z}}_\mathbf{k}\hat{\boldsymbol{\varepsilon}}_{\mathbf{k\cdot\cdot}}$

- Pour le code TRE n°j du Fascicule n°k : $\widehat{\boldsymbol{\mu}}\left(\boldsymbol{\vartheta_k}, \boldsymbol{\vartheta_{kj}}\right) = \left(\mathbf{1} - \hat{\mathbf{z}}_{\mathbf{kj}}\right)\hat{\boldsymbol{\varepsilon}}_{\mathbf{k\cdot\cdot}} + \hat{\mathbf{z}}_{\mathbf{kj}}\mathbf{X_{kj}}$.

- Les coefficients de crédibilité étant donnés par :

  - $\hat{\mathbf{z}}_{\mathbf{kj}} = \frac{\boldsymbol{\alpha\omega_{kj\cdot}}}{\widehat{\sigma_1^2} + \boldsymbol{\alpha\omega_{kj\cdot}}}$, il faut noter que plus $\widehat{\boldsymbol{\sigma_1^2}}$ est élevé, plus $\hat{\mathbf{z}}_{\mathbf{kj}}$ sera faible.
  - $\hat{\mathbf{z}}_{\mathbf{k}} = \frac{\boldsymbol{\beta}\hat{\mathbf{z}}_{\mathbf{k\cdot}}}{\boldsymbol{\alpha} + \boldsymbol{\beta}\hat{\mathbf{z}}_{\mathbf{k\cdot}}}$
  - $\hat{\mathbf{z}}_{\mathbf{k\cdot}} = \sum_{\mathbf{j=1}}^{\mathbf{n}} \hat{\mathbf{z}}_{\mathbf{kj}}$

En pratique, $\widehat{\sigma_1^2}$ et X_{kj} se calculent aisément. En revanche, les autres grandeurs dépendent des paramètres de structure et sont interdépendants, il est donc nécessaire de procéder au calcul itératif. Chaque itération se décompose en quatre étapes :

- La première étape consiste à calculer le $\widehat{\sigma_1^2}$.
- Ensuite il faut initialiser les valeurs de α et de β. En général l'initialisation de ces deux paramètres est à 1. Le paramètre α va permettre de calculer les z_{kj} et les z_k , le paramètre β va quant à lui permettre de calculer les z_k et le z.
- Par le calcul de $\varepsilon_{k..}$ l'algorithme trouve une nouvelle valeur pour le paramètre α et par le calcul de ε l'algorithme obtient une nouvelle valeur pour le paramètre β.

Ainsi il est calculé le taux d'évolution en valeur absolu entre la valeur finale des paramètres α et β et leur valeur de départ. En cas de convergence (taux d'évolutions < 0.001 pour α et pour β) l'algorithme peut calculer les estimateurs empiriques de crédibilité linéaire. Dans le cas contraire, l'algorithme effectue une itération supplémentaire et ainsi de suite jusqu'à obtenir la convergence.

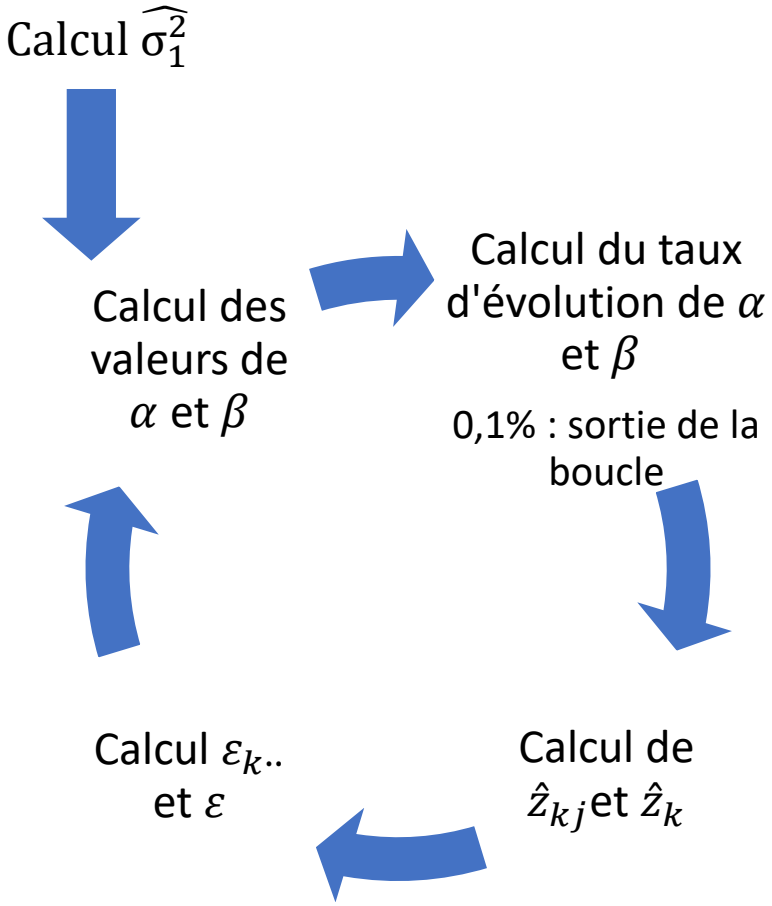


Figure 28 - Schéma calcul pseudo-estimateur

Le but est d'obtenir des valeurs $\hat{z}_{kj}$ suffisantes pour que le modèle prenne en compte suffisamment l'information individuelle de la rubrique.

3.6 CONCLUSIONS DES METHODES DE BÜHLMANN-STRAUB ET DE CREDIBILITE HIERARCHIQUE

Les modèles de Bühlmann-Straub et de crédibilité hiérarchique sont aujourd'hui les plus couramment utilisés dans le milieu de l'actuariat, et présentent les avantages suivants :

- **une certaine robustesse mathématique** : par leur construction, les estimateurs de crédibilité sont calculés mathématiquement à partir d'un problème d'optimisation sous contraintes bien définit. Les estimateurs qui en résultent ne sont donc pas parfaits mais approchent de manière très efficace les solutions recherchées.

- **la capacité à appréhender des portefeuilles hétérogènes** : il est courant que les portefeuilles en assurance soient hétérogènes devant un certain risque. Les modèles de crédibilité y performent très bien relativement en proposant des primes sur mesure qui dépendent du niveau de confiance que l'on accorde à un assuré, son historique de sinistres, et l'expérience collective du portefeuille.

- **la possibilité d'utiliser une approche empirique** : c'est certainement le côté le plus pratique de ces modèles. L'approche qui consiste à estimer les paramètres de structures et les primes de crédibilités de manière empirique permet de s'affranchir des hypothèses de loi a priori et a posteriori des profils de risque des assurés. Ce qui constitue un avantage non négligeable.

La méthode de crédibilité hiérarchique permet la tarification du risque incendie, la suite de l'étude se concentre maintenant sur le choix des niveaux hiérarchique à partir de méthode actuarielle et sur la recherche des seuils de sinistralité afin de préparer la base de données pour le modèle.

4 MUTUALISATION-ECRETEMENT DES SINISTRES EXCEPTIONNELS

4.1 OBJECTIF DES SEUILS

En assurance, l'évènement est défini comme extrême s'il génère une perte financière dite extraordinaire.

L'exemple le plus connu est en 1976, un assuré de la MAIF se retrouve coincé sur une voie ferrée très fréquentée, à la suite de deux facteurs, le mauvais état de la chaussée et de la voiture et une vitesse excessive. La Citroën traction se retrouve bloquée sur les voies ferrées et malgré l'alerte le train de PARIS-STRASBOURG percute la voiture. Le train déraille est renverse une grande partie de sa marchandise, dont une partie va s'écouler dans un étang de pêche (paquet de knorr et biere).

L'événement n'engendre que des pertes matérielles :

- Au niveau du particulier : la voiture
- Au niveau de la SNCF : la locomotive et les wagons ainsi que la perte liée aux dégagements des rails
- La perte denrées transportées
- Au niveau de l'association de pêche, il est évoqué 100 kg de poissons morts.

Le conducteur pourtant à 0,6 grammes par litre n'a enfreint ni le taux d'alcool, ni la vitesse maximale, seul le mauvais état de la voiture lui est reproché. Le coût total de l'accident de 3 milliards de Francs est donc entièrement à la charge de l'assurance.

Un tel enchainement d'événements est exceptionnel cependant, si un tel évènement se produit, **la théorie des valeurs extrêmes** fournit le cadre mathématique probabiliste rigoureux pour évaluer les pertes financières qu'il implique :

- Estimer un quantile extrême
- Estimer la probabilité d'occurrence d'un événement qui n'a pas (encore) été observé.

Deux points de vue sont donc importants dans la théorie des valeurs extrêmes :

- L'étude des lois limites du maximum d'un échantillon
- L'étude des excès par rapport à un seuil u.

L'application de cette théorie permettra donc de déterminer les seuils S1, S2 et S3 évoqués dans la partie précédente. Par définition, ces seuils extrêmes doivent dépasser la gravité d'un événement propre à un TRE. Au sein d'AXA, le seuil des graves pour un contrat est défini à 150 0000 €. L'étude se fera donc sur l'ensemble des sinistres au-delà de ce seuil.

4.2 METHODE UTILISEE : LA THEORIE DES VALEURS EXTREMES

Tout d'abord, afin d'appréhender la notion de valeurs extrêmes, il est utile d'étudier la loi du maximum de la distribution. Soit $X1, \dots, Xn$ une suite de variables aléatoires indépendantes et identiquement distribuées,$Mn = max[X1, \dots, Xn]$ et $FX(x) = P(X \leq x)$. Il vient alors que la loi de Mn est :

$$P(\mathrm{M_n} \leq \mathrm{x}) = P(\mathrm{X_1} \leq \mathrm{x}, \dots, \mathrm{X_n} \leq \mathrm{x}) = P(\mathrm{X_1} \leq \mathrm{x}) \dots P(\mathrm{X_n} \leq \mathrm{x}) = [\mathrm{F_X(x)}]^n$$

Cependant, F n'est pas connue, et de ce fait, cette relation n'est pas utilisable. Rappelons que l'intérêt est d'étudier le comportement asymptotique de la distribution. Notons x^F , le point extrêmal de F défini par :

$$x^F = \sup\{ x \in R: F_X(x) < 1\}$$

Le support de F peut-être borné ($x^F < \infty$) ou infini ($x^F < \infty$). $\mathrm{M_n} \underset{n \to \infty}{\longrightarrow} \mathrm{x^F}$ autrement dit, la distribution asymptotique de M_n est dégénérée. Néanmoins, le théorème de Fisher-Tippet permet de trouver une loi non dégénérée.

Théorème : Fisher et Tippett (1928), Gnedenko (1953)

S'il existe deux suites $an > 0$ et bn et une loi non-dégénérée G telles que :

$$\lim_{n\to\infty} P\left(\frac{M_n - b_n}{a_n} \leq x\right) = F^n(a_n x + b_n) \xrightarrow{d} G(x),\ \forall x \in R$$

Alors G est du type GEV (Distribution des extrêmes généralisée) est nécessairement du type

$$G_{\mu,\sigma,\gamma}(x) = \begin{cases} \exp\left(-\left(1+\gamma\ \frac{x-\mu}{\sigma}\right)_+^{-\frac{1}{\gamma}}\right) si\ \gamma \neq 0 \\ \exp\left(-\exp\left(-\frac{x-\mu}{\sigma}\right)\right) si\ \gamma = 0 \end{cases}, \forall x \in R$$

Si F vérifie les hypothèses ci-dessus, alors F appartient au max-domaine d'attraction de $G_\gamma, F\ \in DA(\gamma)$.

En pratique, l'estimation des suites(a_n) et (b_n) est difficile, le théorème suggère que pour n grand

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) \approx\ G_{\mu,\sigma,\gamma}(x) \Leftrightarrow\ P(M_n \leq x) \approx\ G_{\mu,\sigma,\gamma}(a_n x + b_n) = G^*_{\mu,\sigma,\gamma}(x)$$

Ou $G^*_{\mu,\sigma,\gamma}$ est aussi une GEV.

Cela introduit les lois de valeurs extrêmes généralisées (GEV) avec γ le paramètre de forme (aussi appelé paramètre de queue), μ le paramètre de position, et σ le paramètre d'échelle. Plus γ est grand, plus le poids des extrêmes dans la distribution est important. La fonction GEV est dans le domaine d'attraction de :

- Fréchet si la queue de distribution est épaisse ($\gamma > 0$) ;
- Gumbel si la queue de distribution est fine ($\gamma = 0$) ;
- Weibull si la queue de distribution est finie à droite ($\gamma < 0$)

Domaine d'attraction	**Gumbel** $\gamma=0$	**Fréchet** $\gamma>0$	**Weibull** $\gamma<0$
Loi	Normale	Cauchy	Uniforme
Loi	Exponentielle	Pareto	Beta
Loi	Lognormale	Student	
Loi	Gamma	Burr	
Loi	Weibull		

Figure 29 - Domaine d'attraction par queue de distribution

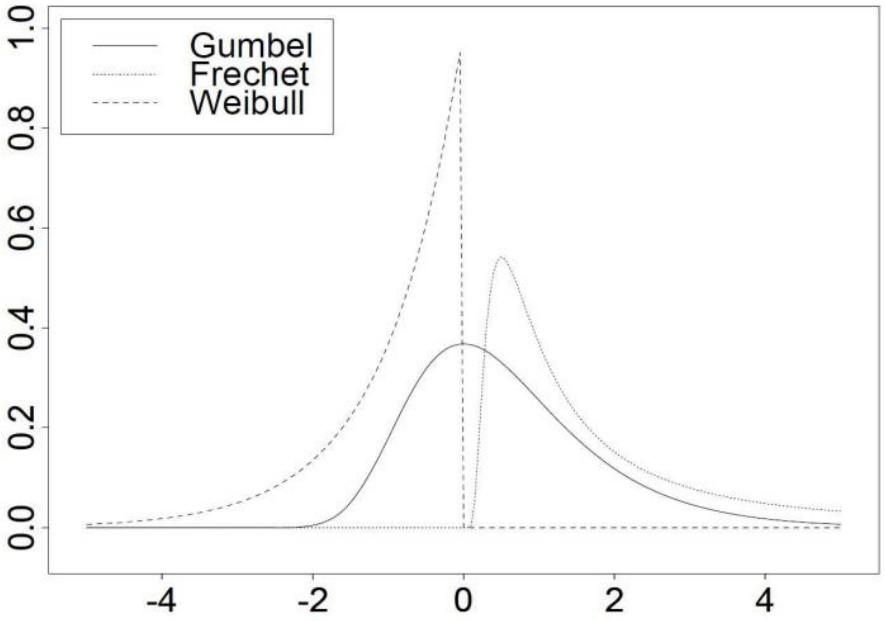


Figure 30 - Densités des distributions de GUMBEL, FRECHET et WEIBULL

Dans ce mémoire, l'étude se focalise sur les sinistres X_i dépassant un seuil u défini, c'est-à-dire les sinistres telles que $(X_i - u)$ soient strictement positifs. Ces sinistres sont caractérisés par des lois de GPD (Distribution de Pareto Généralisée) définies par

$$G^{p}_{\sigma,\gamma}(x) = \begin{cases} 1 - \left(1 + \gamma \frac{x}{\sigma}\right)^{-\frac{1}{\gamma}} \; si\; \gamma \neq 0 \\ 1 - \exp\left(-\frac{x}{\sigma}\right) \; si\; \gamma = 0 \end{cases}$$

$$\forall x \in R^{+}\; si\; \gamma \in R^{+}$$

$$0 \leq x \leq \frac{-\sigma}{\gamma} \; si\; \gamma \in R^{*-}$$

Il existe une relation entre les lois GEV et GPD. En effet, les propositions suivantes sont équivalentes :

- Il existe deux suites $a_n > 0$ et b_n telles que $F(a_n x + b_n)^n \rightarrow G_{\mu,\sigma,\gamma}(x)$
- Il existe une fonction $a(.)$ telle que et b_n

$$\lim_{n \to x^F} \frac{\overline{F}(u + xa(u))}{\overline{F}(u)} = \begin{cases} (1 + \gamma\, x)^{-\frac{1}{\gamma}} \; si\; \gamma \neq 0 \\ \exp(-x)\; si\; \gamma = 0 \end{cases}$$

Pour compléter la distribution GPD est une distribution seuil stable c'est-à-dire que pour $X \sim G\mathrm{PD}(\,\sigma, \gamma\,)$ et u le seuil, alors

$$P(X - u > x | X > x) = \frac{\left(1 + \gamma \frac{x+u}{\sigma}\right)^{-\frac{1}{\gamma}}}{\left(1 + \gamma \frac{x}{\sigma}\right)^{-\frac{1}{\gamma}}} = 1 + \frac{1 + \gamma u}{\sigma + \gamma x}$$

Donc pour $X - u \big| X \sim G\mathrm{PD}(\,\sigma + \gamma x, \gamma\,)$,n le paramètre γ est le même pour tout u.

4.3 DETERMINATION DE SEUILS DE SINISTRALITE

Pour déterminer les seuils, l'étude exploitera certaines propriétés des distributions de Pareto généralisée. Les méthodes utilisées sont la fonction de dépassement moyen des excès, l'estimateur de Hill et Picklands. Toutes ces méthodes sont présentées dans cette section.

Visuellement chaque méthode se lira graphiquement, pour déterminer les seuils il convient de trouver les plus petits seuils pour lequel le graphique de l'estimateur se stabilise. Ainsi les paliers de chaque graphique peuvent être extrapolé comme un seuil potentiel.

L'étude des seuils se base sur les travaux de François Longin (Journal de la société statistique de Paris, 1995)

4.3.1 Méthode des excès au-delà d'un seuil (ou Peak Over Threshold, POT).

Cette méthode développée pour les distributions de Pareto généralisée (Models for exceedances over high thresholds, 1990) permet de sélectionner un seuil optimal. Supposons qu'une $GPD(\sigma, \gamma)$ soit un modèle approprié pour les charges de sinistres excédant un certain seuil donné u, alors la fonction de dépassement moyen des excès est définie par :

$$[e(v) = E[X - v | X > v] = \frac{\sigma + \gamma(v-u)}{1-\gamma}, \forall\, v > u \; et \; \gamma < 1.]$$

Pour γ la fonction est infinie. L'estimateur de cette fonction est quant à lui défini par :

$$e_n(v) = \frac{1}{N_v} \sum_{i=1}^{N_v} (X_i - v)_+$$

Avec N_v le nombre de sinistres supérieurs à v.

Pour la distribution des données au-dessus d'un certain seuil u, une distribution de Pareto généralisée est une approximation robuste. Ainsi, le seuil à partir duquel les sinistres peuvent être considérés comme graves correspond au montant de sinistres à partir duquel le graphique est linéaire en v (avec une pente $\frac{\gamma}{1-\gamma}$). En effet, cela signifierait que les montants correspondant aux sinistres graves suivent une loi de Pareto Généralisée.

Graphiquement le résultat est le suivant :

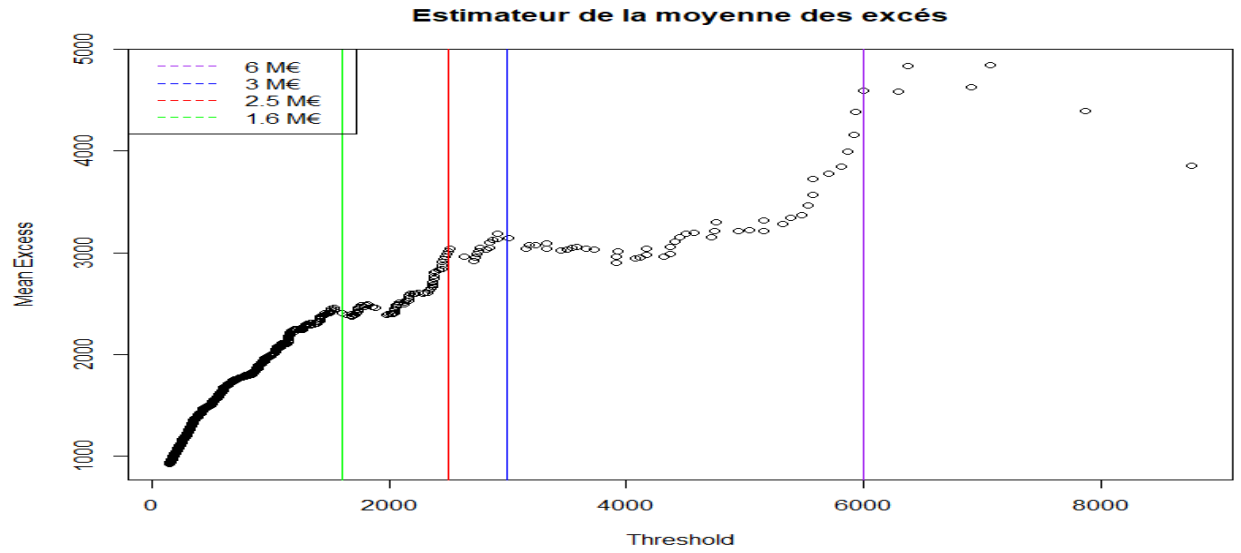


Figure 31 - Estimateur de la moyenne des excès

Les seuils pouvant être retenus sont 1.6 M€, 3 M€ et 6M€.

Cette méthode de l'étude de la fonction de dépassement moyen des excès est très commune car pratique. Elle a pour limite qu'il n'est pas évident de trouver un seuil précis. Les grandes valeurs perturbent le graphique.

4.3.2 Méthode de la Stabilité des paramètres

Dans le cas d'une GPD, les distributions sont dites à « seuil stable ». La propriété de « seuil stable » donne que pour tout dépassement $X - u$ au delà d'un seuil u suit une $GPD(\beta_u, \gamma)$, alors pour tout seuil $v \geq u$, tout dépassement $X - v$ suit également une $GPD(\beta_u, \gamma)$. Notons que γ (paramètre de forme) ne dépend pas du seuil u et que le paramètre d'échelle est une fonction linéaire du seuil : $\beta_v = \beta_u + \gamma(v - u)$.

Graphiquement le choix du seuil se traduit en analysant la stabilité du paramètre d'échelle (Par manque de facilité de lecture du paramètre de forme (γ)). Le graphique permet d'estimer de façon plus explicite ce paramètre. Une approche de γ est donc représentée pour plusieurs seuils, avec les IC (Intervalle de confiance) à 95%. L'objectif est de maintenir un seuil minimum tel que la stabilité du paramètre d'échelle soit garantie, compte tenu de l'incertitude de la mesure compte tenu de l'amplitude de l'intervalle de confiance. Graphiquement cette stabilité dans le cadre de l'étude donne le résultat suivant :

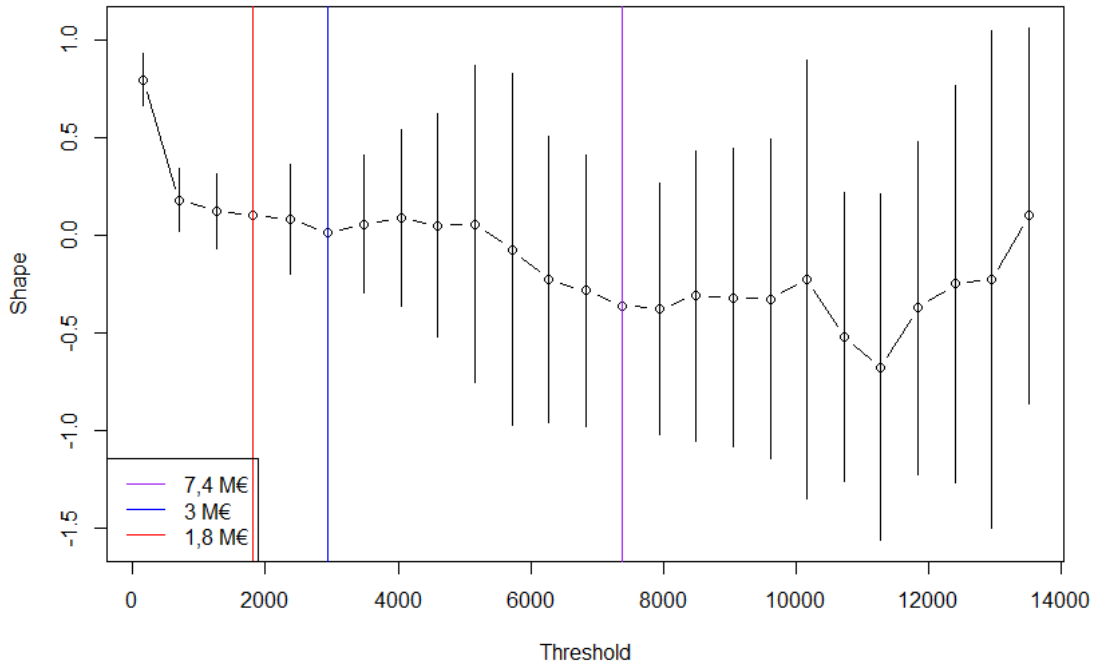


Figure 32 - Stabilité des paramètres

Les seuils pouvant être retenus est uniquement 3 M€ et 7M€.

4.3.3 Estimateur de Pickands

Il est défini par la statistique :

$$\hat{\gamma}^{\mathrm{P}}_{k,n} = \frac{1}{\ln 2} \ln\left(\frac{X_{\mathrm{k,n}} - X_{\mathrm{2k,n}}}{X_{\mathrm{2k,n}} - X_{\mathrm{4k,n}}}\right)$$

Il présente l'intérêt d'être valable quelle que soit la distribution des extrêmes (Gumbel, Weibull ou Fréchet).

La représentation graphique de cet estimateur en fonction du nombre k d'observations considérées montre un comportement en général très volatil au départ, ce qui nuit à la lisibilité du graphique. De plus,

cet estimateur est très sensible à la taille de l'échantillon sélectionné, ce qui le rend peu robuste. Il est donc d'un maniement délicat. Il est alors asymptotiquement normal, avec :

$$\sqrt{k} \times \frac{\hat{\gamma}_{k,n}^{\mathrm{P}} - \gamma}{\sigma(\gamma)} \longrightarrow \mathbb{N}(0,1)$$

lorsque $k \longrightarrow +\infty$ la variance asymptotique étant donnée par:

$$\sigma(\gamma) = \gamma \frac{\sqrt{2^{2\gamma+1} + 1}}{2(2^{\gamma} - 1) \times ln\, 2}$$

Ainsi, le seuil optimal est choisi de telle sorte qu'au-delà de ce seuil, la régression linéaire soit la "meilleure" possible :

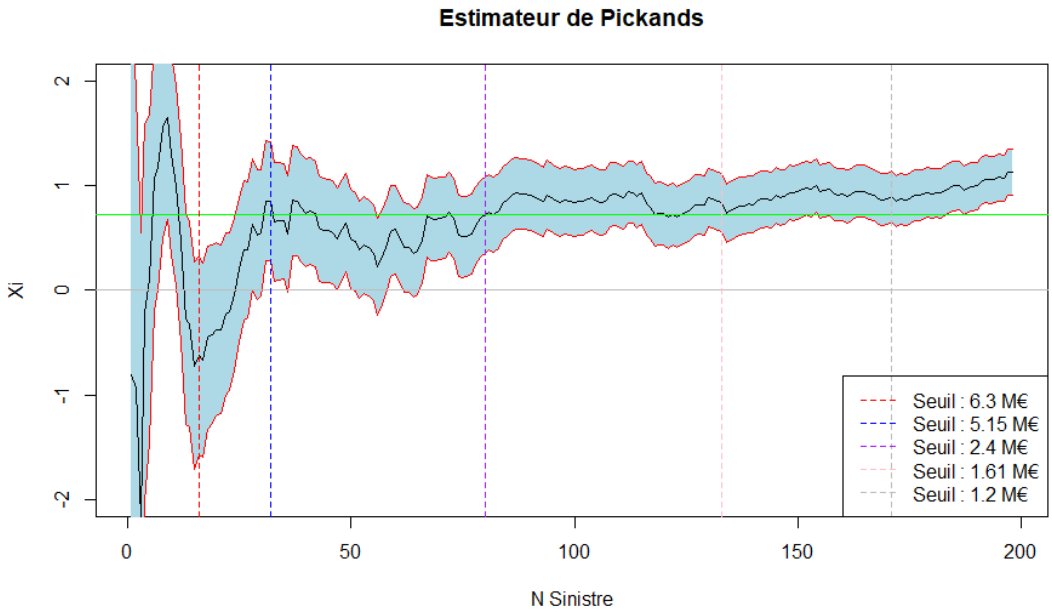


Figure 33 - Estimateur de Pickands

L'objectif est de trouver une zone, appelée plateau, où l'estimateur semble robuste, permettant de déduire les seuils de sinistralités

Les seuils pouvant être retenus sont 1.6 M€, 2.4 M€ et 5,2 M€.

On note que l'estimation $\hat{\gamma}_{k,n}^{P} > 0.$

4.3.4 L'estimateur de Hill

L'estimateur de Hill est le plus usuellement considéré lorsque $\gamma > 0.$ Son principal avantage est l'équilibre biais-variance. Mathématiquement il se définit comme suit :

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

lorsque $k, n \longrightarrow +\infty$ de sorte que lorsque $\frac{k}{n} \longrightarrow 0$ alors la limite $\lim_{k\to\infty} \hat{\gamma}_{k,n}^{H} = \gamma$ et l'estimateur de Hill est le plus asymptotiquement normal:

$$\sqrt{k} \times \frac{\hat{\gamma}_{k,n}^{\mathrm{H}} - \gamma}{\gamma} \longrightarrow \mathbb{N}(0,1)$$

la convergence étant en loi. Cet estimateur est l'estimateur du maximum de vraisemblance dans le cas particulier du modèle $S(x) = 1 - F(x) = Cx^{-1/\gamma}$, soit une distribution de Pareto d'indice $\alpha = \frac{1}{\gamma}$.

Dans le cas général du domaine de Fréchet, la fonction de survie est de la forme $S(x) = 1 - F(x) = x^{-1/\gamma}L(x)$ avec L une fonction à variation lente. Cela induit un biais important sur l'estimateur de Hill, qui est donc en pratique d'un maniement délicat. Dans le cas général, la fonction L apparaît comme un paramètre de nuisance de dimension infinie, qui complique l'estimation.

Le graphique de l'estimateur de Hill représente la valeur de l'estimateur en fonction de l'indice k de la statistique d'ordre, soit l'estimateur construit à partir des observations supérieures ou égales à $X_{k,n}$. Chaque statistique d'ordre peut être alors reliée à un seuil. L'objectif comme pour l'estimateur de Pickands est de trouver une zone, appelée plateau, où l'estimateur semble robuste. Le plus petit seuil u appartenant à cette zone est défini comme optimal. La méthode est illustrée par le graphique ci-dessous :

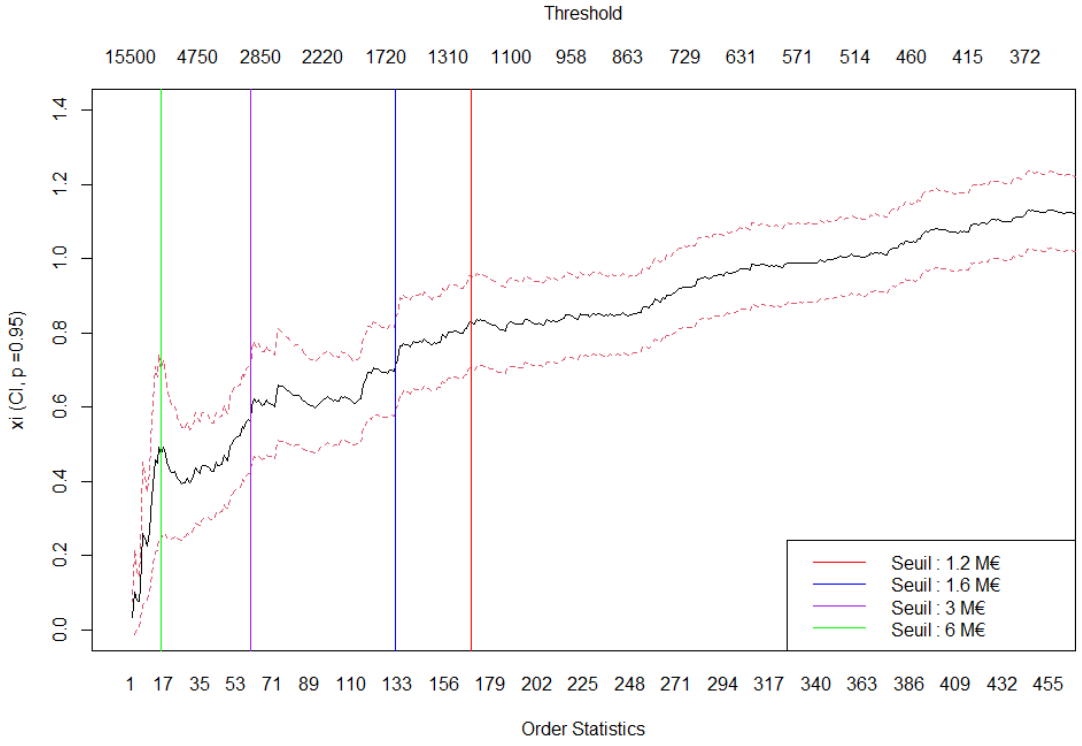


Figure 34 - Estimateur de Hill

Les seuils pouvant être retenu sont 1.2 M€, 3M€ et 6 M€.

4.3.5 Estimateur DEdH (Dekkers, Eimahl et de Haan)

Cet estimateur de l'indice de queue est celui proposé par Dekkers, Einmahl et de Hann. Cet estimateur est une généralisation de l'estimateur de Hill, valable pour tous les domaines d'attraction. Il est défini par

$$\hat{\gamma}_{k,n}^{DEdH} = M_{k,n}^{(1)} + 1 - \frac{1}{2}\left(1 - \frac{\left(M_{k,n}^{(1)}\right)^2}{M_{k,n}^{(2)}}\right)^{-1}$$

Avec $M_{k,n}^{(\mathrm{r})} = \frac{1}{k}\sum_{j=1}^{k}\left(\log\left(X_{n-j+1,n}\right) - \log\left(X_{n-k,n}\right)\right)^r$. La valeur de $M_{k,n}^{(\mathrm{r})}$ correspond à l'estimateur de Hill.

On retrouve donc la convergence suivante :

$$\sqrt{k} \times \frac{\hat{\gamma}_{k,n}^{DEdH} - \gamma}{\sigma(\gamma)} \longrightarrow \mathbb{N}(0,1)$$

Avec

$$\sigma(\gamma)^2 = \begin{cases} 1 - \gamma^2 si \gamma \geq 0 \\ (1-\gamma^2)(1-2\gamma)\left(4 - 8\frac{(1-2\gamma)}{(1-3\gamma)} + \frac{(5-11\gamma)}{(1-3\gamma)} \times \frac{(1-2\gamma)}{(1-4\gamma)}\right) si\ \gamma = 0 \end{cases}$$

En pratique, il n'est pas facile de classer ces estimateurs. Cependant, l'estimateur de Hill présente une variance asymptotique plus faible. C'est pourquoi c'est ce dernier qui est utilisé dans la suite. L'estimateur de Hill étant valable uniquement pour les indices de queue γ > 0, il convient de s'assurer de cette hypothèse.

Graphiquement l'estimateur est le suivant, l'estimation de γ est donc positive :

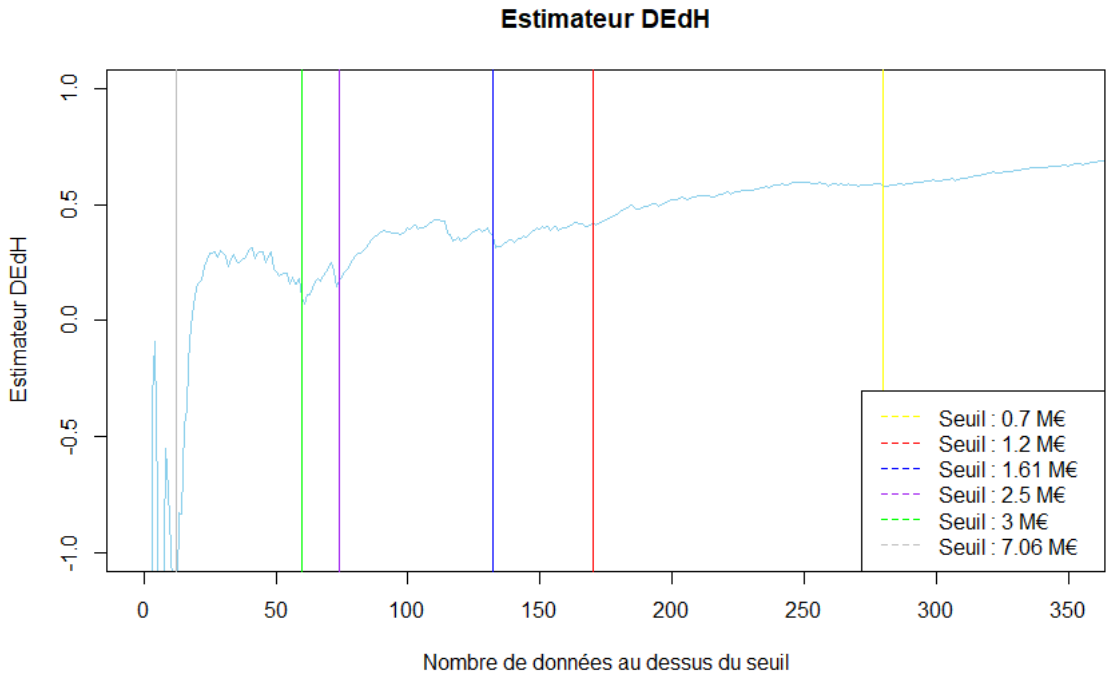


Figure 35 - Estimateur DEdH (Dekkers, Eimahl et de Haan)

L'objectif reste de trouver une zone où l'estimateur semble robuste, permettant de déduire les seuils de sinistralités. **Les seuils sélectionnés à partir de cet estimateur sont 1,6M€, 2,5 M€ et 7 M€.**

4.3.6 Seuils sélectionnés

Voici les seuils par méthode :

Méthode	0,7 M€	1,2 M€	1,6 M€	2,5 M€	3 M€	5,5 M€	6 M€	7 M
Exces			X		X		X	
Stabilité					X			X
Pickands			X	X		X		
Hill		X			X		X	
DEdH			X	X				X

Les seuils sélectionnés pour les tests de l'étude sont les suivants :

- 1,2 M€ pour le seuil 1
  Ce seuil de 1,2 M€ est le seuil utilisé dans l'ensemble des travaux d'AXA France sur la sinistralité « extrême », cela permet de lier l'étude aux travaux AXA.
- 3 M€ pour le seuil 2
  Ce seuil est légèrement supérieur au seuil qui ressort assez fortement de 2,5M€, il permet de garder un fort poids sur le poids individuel.
- 6 M€ pour le seuil 3
  Ce seuil est compris entre les 2 principaux seuils maximum observés (5 M€ et 7 M€).

Figure 36 - Sinistralité par année

Sur l'ensemble de la sinistralité observée pour l'étude, la distribution est la suivante :

Quantile	Montant
100%	50 000 000 €
99%	3 112 850,0 €
95%	662 000,0 €
90%	254 515,0 €
Q3	41 416,0 €
Median	7 025,8 €
Q1	1 680,0 €
10%	408,0 €
5%	270,0 €
1%	171,9 €
0%	6,0 €

Tableau 15 - Quantile de la charge

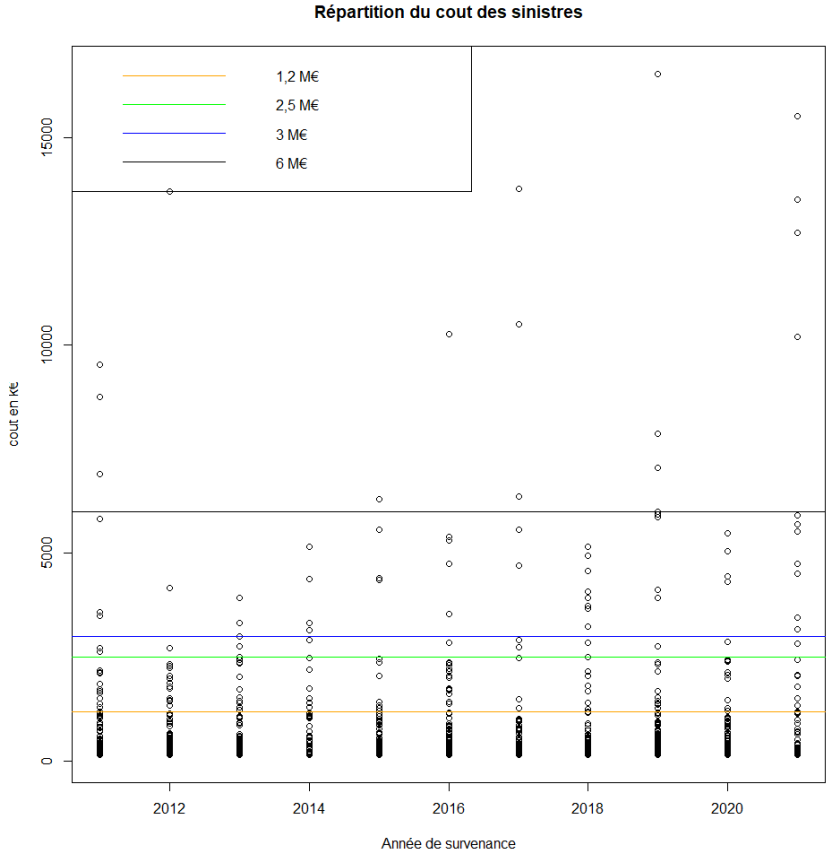
Le seuil de 3M€ permet de conserver 99% des sinistres au niveau de la rubrique d'entreprise.

L'étude de la sinistralité écrêtée à partir du quantile à 99 % donne le tableau suivant :

Seuil	Classe	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	Total
150 k€	Inf au seuil	21,31%	22,50%	32,35%	27,89%	26,19%	20,79%	21,31%	21,89%	18,37%	23,54%	12,66%	78,72%
	Sup au seuil	78,69%	77,50%	67,65%	72,11%	73,81%	79,21%	78,69%	78,11%	81,63%	76,46%	87,34%	
650 k€	Inf au seuil	44,60%	50,49%	61,80%	55,70%	56,26%	47,70%	43,21%	46,01%	41,63%	52,74%	26,87%	54,60%
	Sup au seuil	55,40%	49,51%	38,20%	44,30%	43,74%	52,30%	56,79%	53,99%	58,37%	47,26%	73,13%	
1 200 k€	Inf au seuil	58,34%	65,23%	78,37%	71,86%	71,40%	62,33%	52,73%	60,38%	53,41%	68,92%	35,77%	41,42%
	Sup au seuil	41,66%	34,77%	21,63%	28,14%	28,60%	37,67%	47,27%	39,62%	46,59%	31,08%	64,23%	
1 600 k€	Inf au seuil	65,31%	72,82%	86,10%	78,44%	76,12%	70,23%	56,95%	68,16%	58,99%	75,29%	40,61%	35,18%
	Sup au seuil	34,69%	27,18%	13,90%	21,56%	23,88%	29,77%	43,05%	31,84%	41,01%	24,71%	59,39%	
2 500 k€	Inf au seuil	75,20%	81,80%	97,29%	88,95%	84,44%	81,26%	65,48%	81,86%	68,30%	86,13%	49,75%	25,19%
	Sup au seuil	24,80%	18,20%	2,71%	11,05%	15,56%	18,74%	34,52%	18,14%	31,70%	13,87%	50,25%	
3 000 k€	Inf au seuil	78,75%	83,92%	99,44%	93,10%	87,36%	84,38%	69,23%	87,66%	71,90%	89,53%	53,97%	21,60%
	Sup au seuil	21,25%	16,08%	0,56%	6,90%	12,64%	15,62%	30,77%	12,34%	28,10%	10,47%	46,03%	
5 500 k€	Inf au seuil	90,46%	89,88%	100,00%	100,00%	98,73%	94,77%	83,14%	100,00%	86,28%	100,00%	69,91%	10,36%
	Sup au seuil	9,54%	10,12%	0,00%	0,00%	1,27%	5,23%	16,86%	0,00%	13,72%	0,00%	30,09%	
6 000 k€	Inf au seuil	92,38%	90,50%	100,00%	100,00%	99,57%	95,32%	85,01%	100,00%	88,64%	100,00%	72,01%	9,21%
	Sup au seuil	7,62%	9,50%	0,00%	0,00%	0,43%	4,68%	14,99%	0,00%	11,36%	0,00%	27,99%	
8 000 k€	Inf au seuil	97,57%	92,97%	100,00%	100,00%	100,00%	97,51%	90,19%	100,00%	92,80%	100,00%	78,69%	6,21%
	Sup au seuil	2,43%	7,03%	0,00%	0,00%	0,00%	2,49%	9,81%	0,00%	7,20%	0,00%	21,31%	

Tableau 16 - Répartition annuelle de sinistralité au-delà de différents seuils prédéfinis

Visuellement voici la sinistralité conservée par année, en accord avec les directions techniques, devant le caractère exceptionnel de l'évènement, le sinistre de 50 M€ a été écrêté à 16 M€ pour l'étude, un point représentant le montant final d'un sinistre :



4.4 SYNTHESE

L'écrêtement de la sinistralité permet de limiter l'impact d'intensité des sinistres. Un sinistre extraordinaire sur une rubrique d'activité ne vient pas biaiser sa tarification. Les différentes méthodes étudiées pour le choix des seuils de mutualisation convergent vers les 3 montants de 1,2 M€, 3 M€ et 6 M€.

La base de données et le choix de la méthode étant finalisés, la dernière étape de l'étude est donc la détermination du niveau hiérarchique appliqué à la crédibilité.

5 CHOIX DU NIVEAU HIERARCHIQUE ET APPLICATION

Le modèle de crédibilité hiérarchique se construit à partir de structures dites en arborescence, avec des approches pyramidales pour calculer les primes redistribuées au fur et à mesure dans chaque niveau inférieur plus homogène.

La détermination de ces classes homogènes est réalisée dans cette partie à partir de différentes méthodes actuarielles.

Il est important de noter que les rubrique d'activité (TRE) sont présentes chacune dans un unique fascicule, un seul NARI et un segment, cependant les combinaisons fascicule NARI et segment sont multiples. L'étude se portera donc sur un seul niveau hiérarchique. Le choix du niveau sera le critère le plus discriminant.

5.1 1ER METHODE : TEST D'HOMOGENEITE VIA L'ANOVA

5.1.1 Mise en évidence de l'hétérogénéité

On considère X_{jt} = coût pour AXA pour le TRE_j, lors de l'année t avec $j = 1, \dots, J, et\ t = 1, \dots, Tj$.

L'interrogation sur l'étude est de choisir la classe représentant le plus d'homogénéité. Cette méthode d'approche est abordée par Olivier Wintenberger dans son cours, conseillant l'ANOVA.

L'ANOVA est un test statistique permettant d'identifier la significativité de certains facteurs sur la variance de données.

L'analyse de la variance, également appelée ANOVA (correspondant à Analysis Of Variance en anglais) a été développée au début du XXè siècle par Ronald Fisher, un statisticien britannique. C'est un modèle statistique qui sert à démontrer l'existence de similitudes ou différences sur des aspects précis dans une population étudiée.

Dans l'ANOVA, il est étudié une variable quantitative à laquelle il est attribué une ou deux variables qualitatives : les variables catégorielles.

Dans le cas de l'étude, la variable quantitative sera la charge observée, et les variables catégorielles seront les 3 classes possibles (catégorie, Niveau d'acceptation du risque, fascicule)

Ces variables catégorielles sont appelées « facteurs » ou « facteurs de variabilité ». Si l'analyse de la variance se concentre sur un unique facteur, il s'agit alors analyse à un facteur ou One-way ANOVA. Si plusieurs facteurs entrent dans le test analytique, il s'agit d'analyse à deux facteurs, multifactorielle ou MANOVA pour Multivariate Analysis Of Variance.

L'ANOVA sert concrètement à mettre en lumière l'existence d'une interaction entre ces facteurs de variabilité et la variable quantitative principale étudiée, généralement une population divisée en 2 ou 3 groupes.

5.1.2 Formulation mathématique

On considère des variables X_{jt}avec $E[X_{jt}] = m_j$, deux à deux indépendantes.

Il faut donc tester

- $H0 :\ m_1 = \dots = m_j = \dots = m_J$,
- $H1 : \exists (j,k) : m_j \neq m_k$.

On note

$$\bar{X}_{jt} = \frac{1}{T_j}\sum_{t=0}^{T_j} X_{jt}$$

$$\bar{X} = \frac{1}{\sum_{t=0}^{J} T}\sum_{j,t} X_{jt}$$

La part de la variance totale SCE qui peut être expliquée par le $\text{SCE}_{\text{inter-classes}}$, aussi appelée variabilité inter-classe, SSB ou Sum of Square Between class) et la part de la variance totale SCE qui ne peut être expliquée par le $\text{SCE}_{\text{inter-classes}}$ aussi appelée variabilité aléatoire, variabilité intra-classe, bruit, SSW ou Sum of Square Within class) sont données par les formules :

$$\text{SCE}_{\text{total}} = \text{SCE}_{\text{inter-classes}} + \text{SCE}_{\text{intra-classes}}$$

$$\text{SCE}_{\text{inter-classes}} = \sum_{\text{j=1}}^{\text{ni}} \text{T}_\text{j}(\bar{X}_j - \bar{X})^2$$

$$\text{SCE}_{\text{intra-classes}} = \sum_{j=1}^{J}\sum_{t=1}^{\text{T}_\text{j}}(X_{jt-} \bar{X}_j)^2$$

Avec

$\text{SCE}_{\text{intra-classes}} \longrightarrow \chi^2_{j-1}$

$\text{SCE}_{\text{inter-classes}} \longrightarrow \chi^2_{n-j}$

Le test est le suivant :

$$F = \frac{\frac{\text{SCE inter} - \text{classes}}{J-1}}{\frac{\text{SCE intra} - \text{classes}}{T-J}}$$

Sous H0, il est attendu à ce que F soit petit.

La forme du test :

- Si F > s_α, l'hypothèse H0 est rejetée.
- Si F < s_α, l'hypothèse H0 est conservée.

s_α se calcule en faisant une hypothèse sur la loi des X_{jt}.

5.1.3 La loi statistique du test et p-value

On suppose que $X_{jt} \sim N(m_j, \sigma^2)$ (on rappelle qu'on avait supposé l'indépendance). Sous l'hypothèse H0,$F \sim F(J-1, T-J)$, où F désigne la loi de Fisher.

La « valeur p » en français, ou « p-value » est la probabilité pour un modèle statistique donné sous l'hypothèse nulle d'obtenir une valeur au moins aussi extrême que celle observée.

- p > 0,05 : très probablement dû au hasard, la différence entre les deux séries d'observation n'est pas significative, il n'est pas possible de conclure.
- 0,05 ≤ p < 0,01 : faiblement significatif
- 0,01 ≤ p < 0,005 : significatif
- p ≤ 0,005 : fortement significatif

5.1.4 Résultat de l'ANOVA

A partir du jeu de données les variables testées sont les charges incendies observées et les 3 classes précisées ci-dessus.

Facteur	Df	Sum Sq	Mean Sq	F value	p-value
NARI	5	3,63E+11	7,27E+10	5,75	0,00%
Fascicule	11	2,25E+11	2,04E+10	1,62	8,72%
Segment	3	9,59E+10	3,20E+10	2,53	5,53%
NARI x Fascicule	21	7,75E+11	3,69E+10	2,92	0,00%
NARI x Segment	8	9,08E+10	1,13E+10	0,90	51,70%
Fascicule x Segment	13	2,29E+11	1,76E+10	1,39	15,33%
Fascicule x Segment x NARI	1	2,96E+09	2,96E+09	0,23	62,85%
Résidus	410833	5,19E+15	1,26E+10		

Tableau 17 - Résultat ANOVA

Le test d'ANOVA montre la pertinence du choix du NARI, qui dans le cadre du NARI est proche de 0%.

5.2 2ND METHODE : MACHINE LEARNING

Afin de confirmer le choix de l'ANOVA, l'approche par deux modèles de Machine Learning est appliquée : Gradient Boosting Model et Random Forest.

Ces méthodes sont robustes et pratiques à utiliser, cependant l'effet « boite noire » n'est pas abordable d'un point de vue opérationnel. L'étude va se pencher sur ces deux modèles afin de confirmer la pertinence du choix de niveau hiérarchique.

Ces modèles ont l'avantage d'être non paramétriques, c'est-à-dire qu'il n'est pas nécessaire de les alimenter avec des hypothèses sur les données qui les alimentent. D'autre part, il n'est pas nécessaire de spécifier les liens entre les variables. Si des dépendances existent, les modèles les détecteront et l'appliqueront à la régression.

L'utilisation du Machine Learning dans le cadre de l'étude a une limite, les 3 choix possibles étant clivant : un secteur d'activité ayant un indicateur propre (un secteur d'activité a un seul NARI, Fascicule et segment)

5.2.1 Méthodologie

Le principe nécessite de travailler sous deux sous-groupes de la base de données :

- Une sous base de données d'apprentissage, dans le cas de l'étude il s'agit de 70% de la base d'origine
- Une sous base de données qui servira à tester la qualité de l'ajustement de chaque modèle et à comparer les modèles entre eux

Dans le même axe d'étude que pour l'ANOVA, la prédiction de la charge incendie est attendue par les modèles.

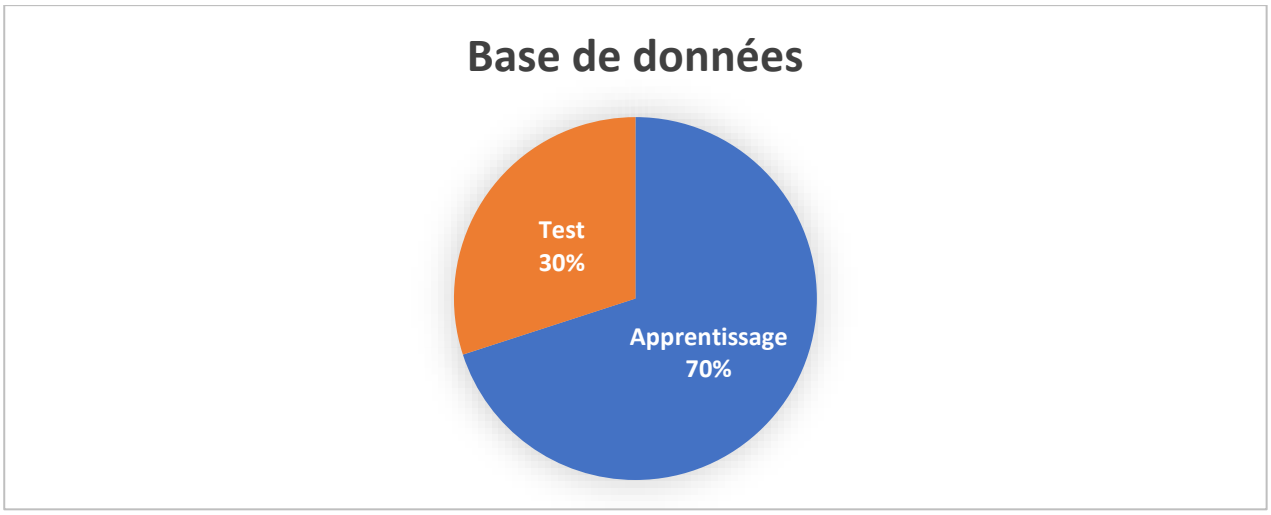


Figure 37 - Répartition entre la base d'entrainement et la base d'apprentissage

5.2.2 Prérequis

5.2.2.1 Mesure de l'erreur

L'erreur moyenne quadratique, notée :

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

Le modèle sélectionné sera celui qui minimise ce critère.

Dans la continuité, l'utilisation de la RMSE (racine de l'erreur quadratique moyenne) permet de s'exprimer dans la même unité que la variable réponse :

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

5.2.2.2 Mesure de la qualité d'ajustement

La courbe de Lorenz permet de juger du pouvoir de prédiction d'un modèle. L'AUC (Area Under Curve) qui lui est rattaché permet de comparer les modèles entre eux. L'indice de Gini est défini comme suit : $Gini = 2 * AUC - 1$.

Construction de la courbe de Lorenz :

- Soit $\hat{y}i$ la valeur prédite par le modèle pour l'individu i
- Soit yi la valeur observée de la variable réponse pour le même individu i
- Et soit$y(\widehat{1}) \geq y(\widehat{2}) \geq \cdots \geq y(\widehat{n})$la suite ordonnée des valeurs prédites par le modèle

La courbe est ainsi construite :

$$\frac{\sum_{k=1}^{i} y_k}{\sum_{k=1}^{n} y_k}$$

L'objectif dans le cadre d'une utilisation de la courbe de Lorenz sur une recherche de segmentation est d'avoir une courbe le plus éloignée possible de la droite d'égalité, qui représente le tarif moyen.

Une illustration simple est l'image des buteurs d'une équipe de football : il est facile de comprendre que les attaquants vont en moyenne plus marquer que les défenseurs. La courbe de fréquence de but va donc s'éloigner de l'axe d'égalité de but par joueur.

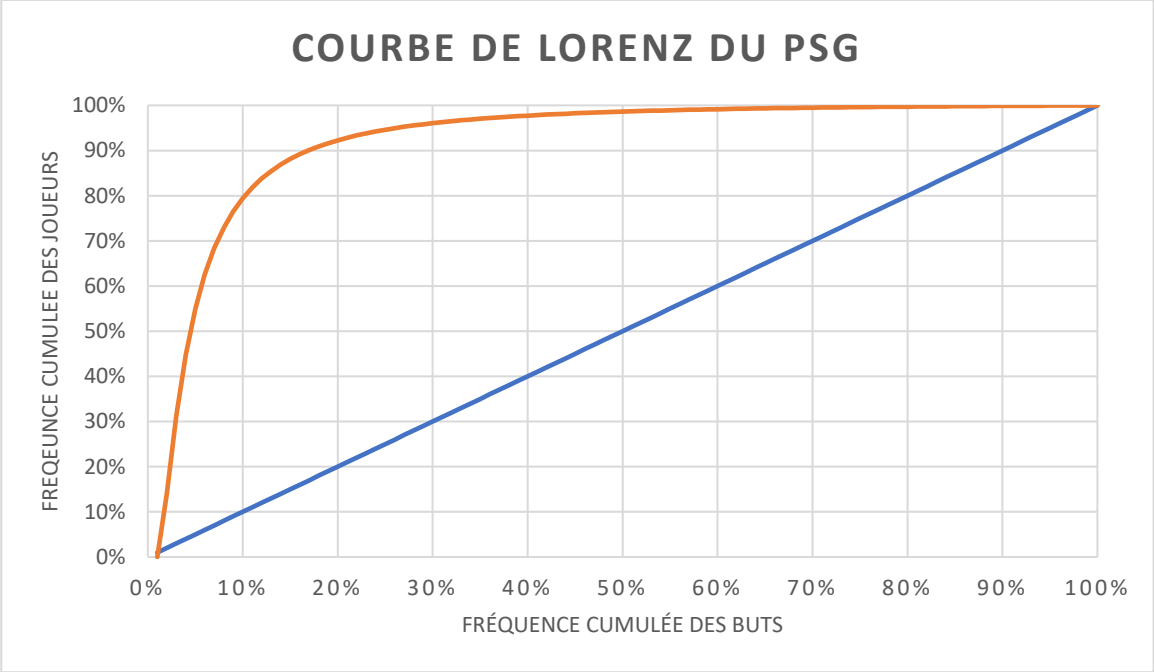


Figure 38 - Courbe de Lorenz du PSG

Dans le cadre d'une classification, l'objectif étant de déterminer le modèle le plus clivant, la courbe doit maximiser son AUC, c'est-à-dire son écart au-dessus de la courbe d'égalité.

5.2.2.3 Arbres de décision

Méthodes d'apprentissage supervisé, les arbres de décision se basent sur des données d'apprentissage et une variable réponse connue. Ils ont un fonctionnement simple : regrouper les individus de manière homogène par rapport à la variable à prédire.

Introduisons quelques notions de vocabulaire :

- Le noeud parent est le point de départ de l'arbre
- Chaque noeud se découpe en deux noeuds fils
- Les noeuds terminaux s'appellent des feuilles

Citons quelques avantages et inconvénients de ces méthodes :

Avantage	Inconvenient
– Modèles non linéaires, rapides à entrainer – Faciles à comprendre, interpréter et visualiser – Utiles pour la sélection et le découpage en tranches des variables – Robustes aux valeurs manquantes et aux corrélations entre variables – Peu de préparation des données en amont – Temps d'executions raisonnables	– Risque de sur-apprentissage – Peuvent être instables à cause de petites variations dans les données – Tendance à favoriser les variables qualitatives avec beaucoup de modalités (car elles offrent plus de souplesse dans la division) – Faible performance

Comme indiqué dans le tableau les arbres ont une tendance au sur-apprentissage : l'algorithme apprend avec tellement de précision les données d'entrainement qu'il ne parvient pas à généraliser un résultat satisfaisant sur de nouvelles données. Pour pallier cela, il est primordial d'utiliser des critères d'arrêt, l'élagage, les trois plus courants sont :

- L'amélioration de l'erreur doit être supérieure à un seuil fixé
- Le nombre de points dans une feuille ne peut être plus petit qu'un seuil fixé
- Le nombre de nœuds de l'arbre doit être inférieur à un nombre donné

L'AUC est un indicateur de performance permettant de sélectionner le modèle optimal.

5.2.2.4 L'arbre CART

L'algorithme CART (Classification And Regression Tree) est introduit par Breiman et col. en 1984, il a pour objectif d'améliorer les résultats obtenus par les arbres de décision simples.

A chaque nœud, l'arbre cherche à découper la population en deux groupes homogènes du point de vue de la variable à expliquer. Selon la nature de la variable explicative, le découpage sera de la forme ($X < s$, $X \geq s$) ou ($X = xi$, $X \neq xi$). Il faut réitérer plusieurs fois la segmentation jusqu'à ce que plus aucun découpage ne soit possible. La segmentation optimale sous-entend une mesure de distance à optimiser ou fonction d'hétérogénéité : le critère par défaut est l'erreur quadratique moyenne.

Le principe de CART est de construire l'arbre maximal et d'en déduire l'arbre optimal par élagage (ou pruning). L'algorithme construit une suite emboîtée de sous-arbres de l'arbre maximal puis choisit, seulement parmi cette suite, l'arbre optimal qui minimise le risque ou erreur de généralisation.

Cet algorithme sous le logiciel R offre plusieurs options pour les critères d'arrêt, par exemple :

- minsplit : si un nœud contient moins d'observations que minsplit, il devient une feuille
- minbucket : nombre minimum d'observations dans une feuille
- cp : paramètre de complexité. C'est la correction minimum d'erreur qui doit être apportée par un découpage
- weights : poids à affecter aux observations (ici : Capitaux Incendie)

L'arbre ci-dessous est une illustration d'arbre CART appliqué aux données.

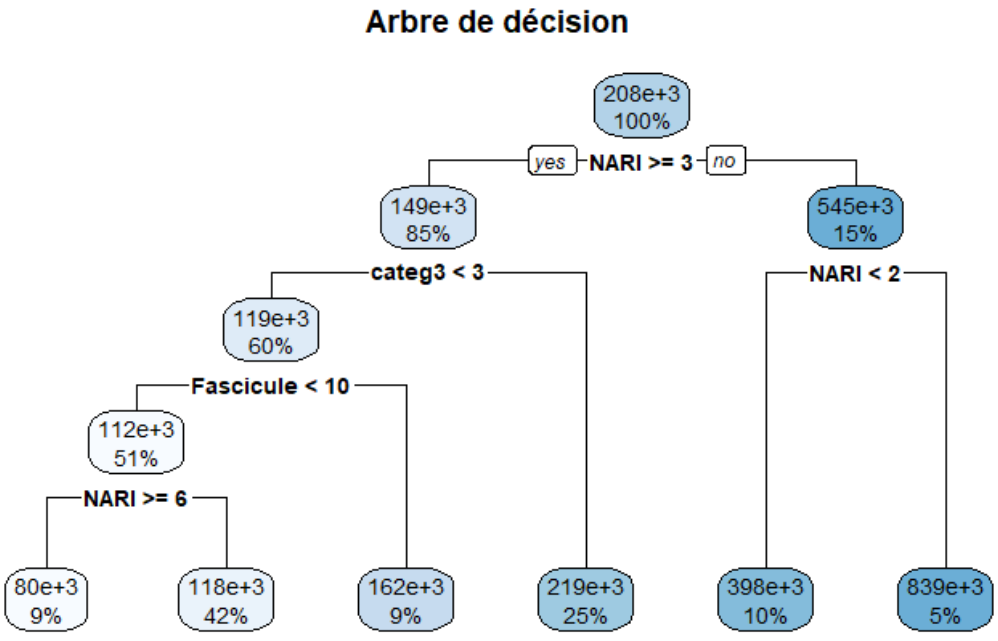


Figure 39 - Arbre de décision

5.2.2.5 *K-folds Cross Validation*

La validation croisée aide à l'évaluation des modèles d'apprentissage automatique. Cette méthode statistique aide à comparer et à sélectionner le modèle dans l'apprentissage automatique appliqué. La compréhension et la mise en œuvre de ce problème de modélisation prédictive sont faciles et directes. Cette technique présente un biais plus faible lors de l'estimation des compétences du modèle.

Avant de vérifier le modèle final sur la base de validation, la base de modélisation est partagée en plusieurs jeux de données équilibrés afin de vérifier les différentes modélisations par k-fold. Cette technique permet d'éviter l'over fitting, c'est-à-dire de s'assurer qu'il n'existe pas de cases tarifaires trop petites qui auraient tendance à biaiser les résultats en forçant le modèle à s'ajuster.

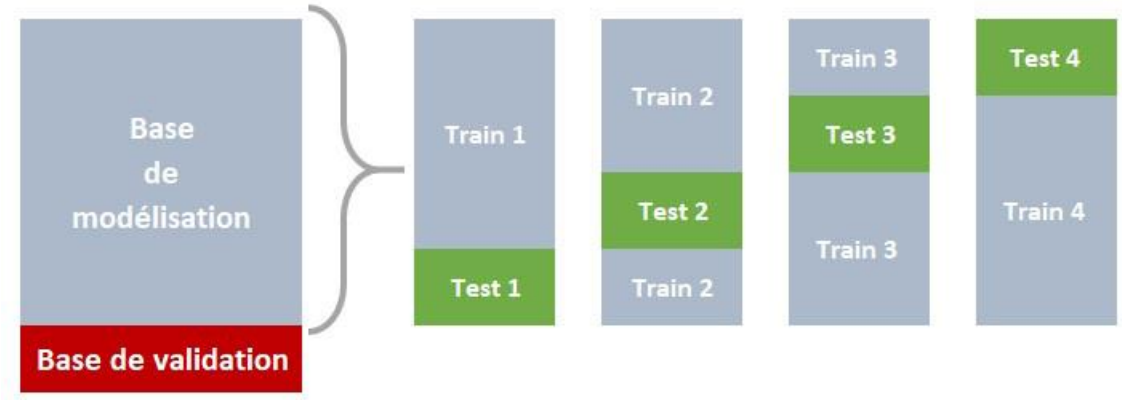


Figure 40 - Illustration K-Folds

La base de données D sous la forme D = (($x1$, $y1$), …, (xn, yn)) et soit C le nombre de modèles parmi testé. Alors pour m ∈ {1, …, C} le numéro du modèle. Soit L la fonction de perte et $\varepsilon_m = \mathbb{E}[L(Y, \widehat{Ym})]$ l'erreur du modèle m. L'objectif est donc d'avoir m qui minimise ε_m.

Principe de la méthode $K - folds$:

Pour K donné, la base D est échantillonnée en permutant les données de manière aléatoire puis découpée en K sous-échantillons de même taille.

Ainsi pour chaque modèle m, l'erreur d'estimation ε_m est calculée K fois, soit $\varepsilon_m(j)$ l'erreur d'estimation pour le modèle m à la $j^{\text{ième}}$ itération. Ainsi à la $j^{\text{ième}}$ itération, le $j^{\text{ième}}$ échantillon est la base de validation et les K − 1 restants représentent base d'apprentissage.

$$\hat{\varepsilon}_m = \frac{1}{K}\sum_{j=1}^{K} \hat{\varepsilon}_m(j)$$

5.2.2.6 *L'erreur Out Of Bag*

L'erreur Out Of Bag (OOB), également appelée estimation Out of Bag, est une méthode de mesure de l'erreur de prédiction des forêts aléatoires, des arbres de décision et d'autres modèles d'apprentissage automatique utilisant l'agrégation bootstrap (bagging). Le bagging utilise le sous-échantillonnage avec remplacement pour créer des échantillons d'apprentissage à partir desquels le modèle peut apprendre. L'erreur OOB est l'erreur de prédiction moyenne sur chaque échantillon d'apprentissage x_i, en utilisant uniquement les arbres qui n'avaient pas xi dans leur échantillon bootstrap.

Si (X_i, Y_i) est l'observation en question et Y_i la prédiction de Y_i en fonction de $X_i$en agrégeant uniquement les arbres ne contenant pas (X_i, Y_i) . Alors le taux d'erreur OOB est donné par :

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2$$

5.2.2.7 Méthodes ensemblistes

En pratique, ces arbres sont peu utilisés du fait de leur tendance au surapprentissage. Ils sont plutôt utilisés en tant que classifieurs faibles, ce qui permet une prédiction meilleure que l'aléatoire, à la base des méthodes ensemblistes. Le principe de ces méthodes est de considérer les résultats d'un ensemble de classifieurs et de les agréger pour en déduire une meilleure prédiction. Ils sont basés sur des stratégies aléatoires, en considérant des sous-ensembles de l'échantillon de départ et des variables explicatives (Random Forest), ou sur des stratégies adaptatives, en variant le poids des observations (GBM).

Le principe est simple : combiner les prévisions de plusieurs modèles permet de réduire la variance et donc l'erreur de prévision.

5.2.2.8 Bagging

Le mot Bagging est une contraction de Bootstrap Aggregation. Le bagging est une technique utilisée pour améliorer la classification notamment celle des arbres de décision, considérés comme des « classifieurs faibles », c'est-à-dire à peine plus efficaces qu'une classification aléatoire.

Soit Y la variable à expliquer (quantitative ou qualitative) et soient $X_1, \dots, X_p$ les variables explicatives et $f(x)$ un modèle fonction de$x = x_1, \dots, x_p$.

Soit n le nombre d'observations et $z = \left((x_1, y_1), \dots, (x_n, y_n)\right)$ un échantillon de loi F.

Avec m échantillons indépendants notés $\{z_i\}_{i=\{1,\dots,m\}}$, une prévision par agrégation est définie selon la nature de la variable à expliquer :

- Quantitative : $\hat{f}_m = \frac{1}{n}\sum_{i=1}^{m} \hat{f}_{zi}(.)$
- Qualitative : $\hat{f}_m = arg\ max_j card\{i | \hat{f}_{zi}(.) = j\}$

En vulgarisant : Soit une base d'entraînement D de taille n, le bagging génère m sous-bases d'entraînement D_i, chacun de taille n',en échantillonnant à partir de D uniformément et avec remise. En échantillonnant avec remise, certaines observations peuvent être répétées dans chaque D_i. Si $n' = n$, alors pour grand n l'ensemble D_i devrait avoir la fraction $(1 - 1/e)$ (≈63,2%) des exemples uniques de D, le reste étant des doublons.

Ce type d'échantillon est appelé échantillon bootstrap. L'échantillonnage avec remise garantit que chaque bootstrap (D_i) est indépendant de ses pairs, car il ne dépend pas des échantillons choisis précédemment lors de l'échantillonnage. Ensuite, m modèles sont ajustés à l'aide des m échantillons bootstrap ci-dessus et combinés en faisant la moyenne de la sortie (pour la régression) ou en votant (pour la classification).

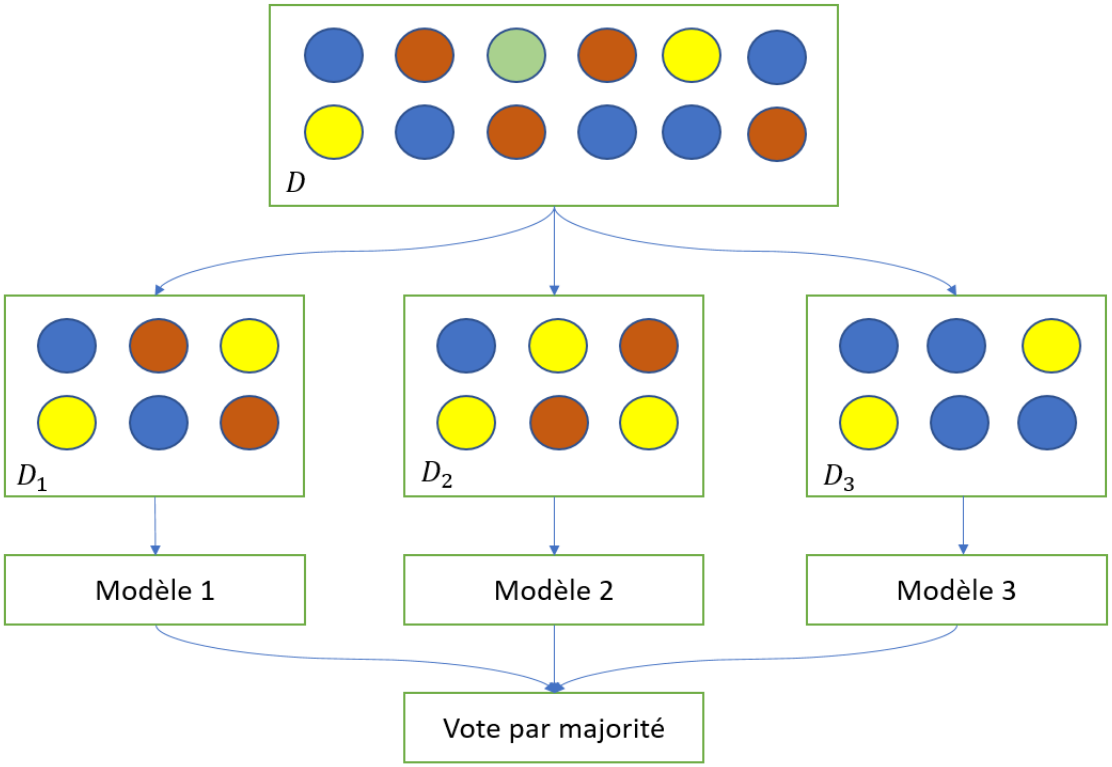


Figure 41 - Illustration du BAGGING

5.2.2.9 Le boosting

Le principe du boosting est quelque peu différent du bagging. Les différents classifieurs sont pondérés de manière qu'à chaque prédiction, les classifieurs ayant prédit correctement auront un poids plus fort que ceux dont la prédiction est incorrecte.

Le principe est issu de la combinaison de classifieurs (appelés également hypothèses). Par itérations successives, la connaissance d'un classifieur faible - weak classifier - est ajoutée au classifieur final - strong classifier. L'estimateur à l'étape k concentre donc ses efforts sur les observations mal ajustées à l'étape k-1.

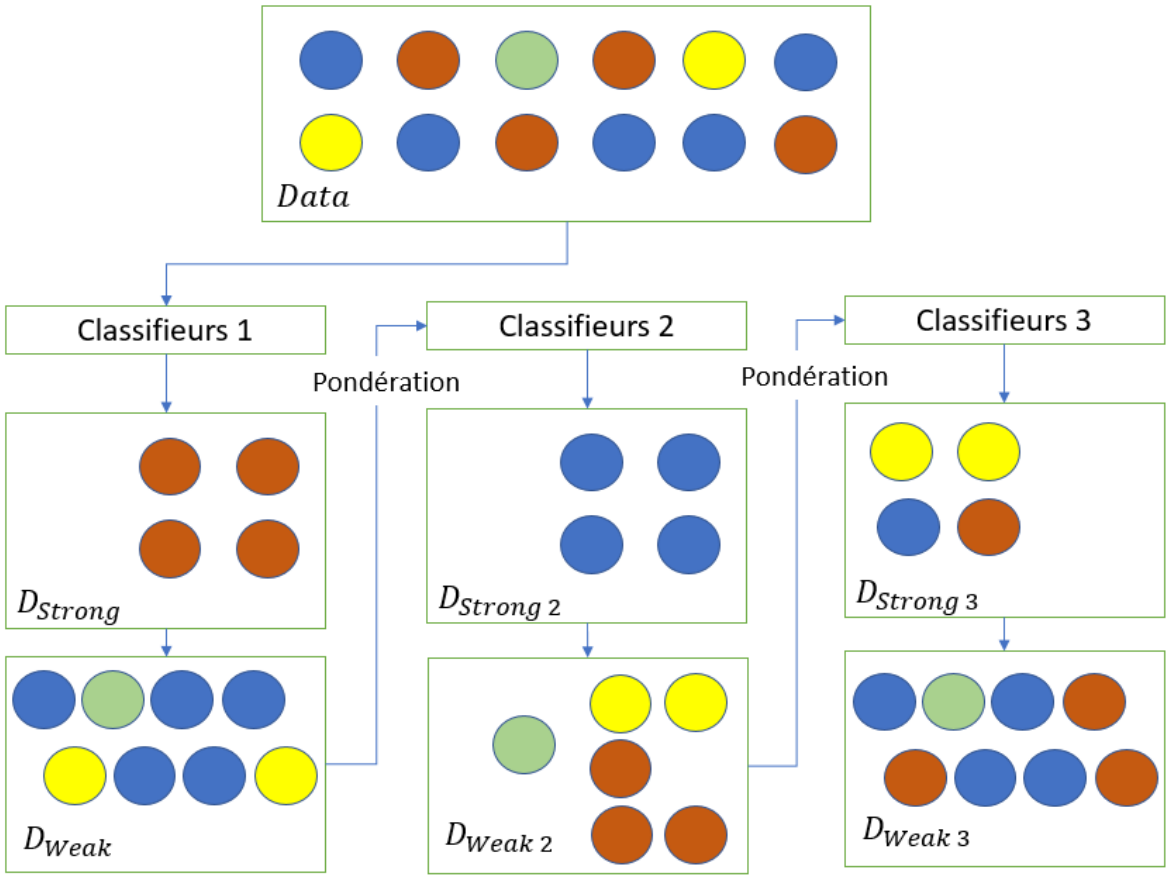


Figure 42 - Illustration du Boosting

Le bagging et le boosting, les deux étant les méthodes couramment utilisées, ont une similitude universelle d'être classés comme méthodes d'agrégation. Les similitudes sont les suivantes :

- Les deux sont des méthodes d'agrégation pour obtenir n sous bases D_i d'apprentissage à partir d'une base D.
- Les deux génèrent plusieurs ensembles de données d'entraînement par échantillonnage aléatoire.
- Les deux prennent la décision finale en faisant la moyenne de n apprentissage (ou en prenant la majorité d'entre eux, c'est-à-dire le vote majoritaire).
- Les deux sont efficaces pour réduire la variance et offrent une plus grande stabilité.

Les principales différences sont les suivantes :

- Modèle indépendant les uns des autres pour le bagging, et influences des modèles précédents pour le boosting
- Baisse de la variance et du sur-apprentissage pour le bagging contre le biais pour le boosting

Dans le cadre d'une variance importante des classificateurs il faut privilégier le bagging, dans le cadre d'un classificateur stable avec un biais élevé le boosting est recommandé.

5.2.3 Random forest

Les forêts d'arbres décisionnels (ou forêts aléatoires de l'anglais random forest classifier) ont été premièrement proposées par Ho en 1995 et ont été formellement proposées en 2001 par Leo Breiman et Adele Cutler. Elles font partie des techniques d'apprentissage automatique. Cet algorithme combine les concepts de sous-espaces aléatoires et de bagging

Les forêts aléatoires permettent de créer des arbres décorrélés tout en générant un grand nombre d'arbres par la méthode de rééchantillonnage Bootstrap. De plus, cette méthode d'agrégation d'arbres répond à la fois aux problèmes de régression et de classification.

Lorsque les arbres sont construits, à chaque fois qu'un noeud est considéré, un sous-ensemble aléatoire m des p variables explicatives est choisi comme candidat pour le découpage de l'espace des variables. Généralement, m est choisi de telle sorte que $\mathrm{m} = \sqrt{\mathrm{p}}$. En d'autres termes, lors de la construction d'une forêt aléatoire, à chaque noeud d'un arbre, l'ensemble des variables explicatives n'est pas considéré pour couper l'espace. Bien que contre-intuitif, il y a une explication rationnelle à cette méthode.

Ainsi lorsque dans la base de modélisation, une variable explicative expliquant très fortement la variable réponse, notée p_1 alors la majorité des arbres construits utiliseront p_1 pour découper l'espace dans le premier noeud. Ainsi, la plupart des arbres construits se ressembleront, et les prédictions de ces derniers seront fortement corrélées. C'est pourquoi les forêts aléatoires permettent de contourner ce problème en forçant chaque nœud à considérer seulement un sous ensemble de variables explicatives.

Avantages	Inconvénients
– Modèle non linéaire, rapide à entrainer – Robuste aux valeurs manquantes et aux corrélations entre variables – Peut gérer de très grands jeux de données – Réduit la variance en minimisant l'augmentation du biais	– Peu de contrôle sur ce que fait le modèle – Mauvais résultats si le problème sous-jacent est linéaire

Le Random Forest ne donne pas une équation et des coefficients permettant d'exprimer la variable réponse. Néanmoins des informations pertinentes sont obtenues par le calcul et la représentation graphique d'indices proportionnels à l'importance de chaque variable dans le modèle agrégé et donc de sa participation à la régression. Deux critères sont proposés pour évaluer l'importance des variables :

- %IncMSE : repose sur une permutation aléatoire des valeurs de cette variable. Plus la qualité de la prévision, en termes d'erreur OOB, est dégradée par la permutation, plus celle-ci est importante. Il s'agit d'une mesure globale mais indirecte de l'influence d'une variable sur la qualité des prévisions
- IncNodePurity : indicateur local, repose sur la décroissance de l'hétérogénéité. L'importance de la variable est alors la somme pondérée des décroissances d'hétérogénéité induites lorsqu'elle est utilisée pour définir la division associée à un nœud

Les résultats obtenus montrent la pertinence du niveau d'acceptation du risque :

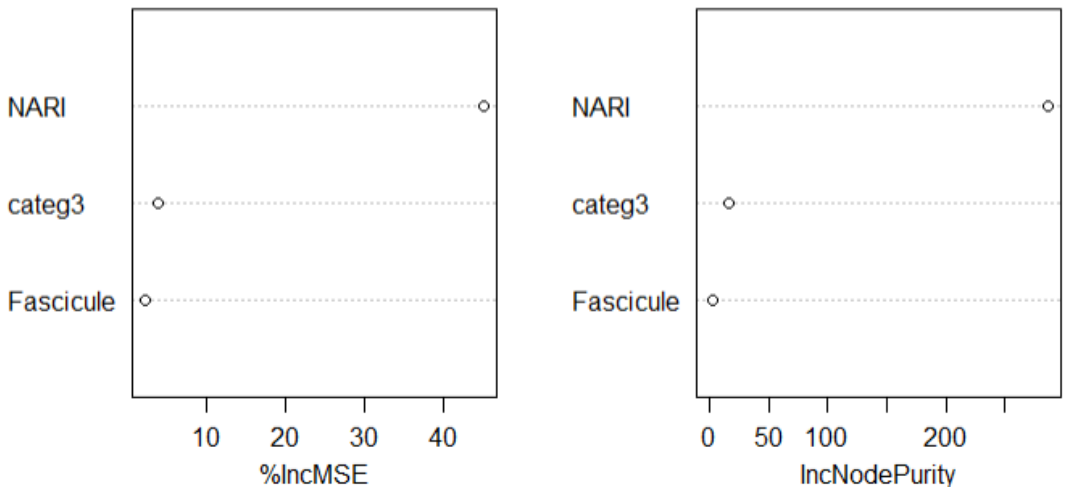


Figure 43 - Sortie du Random Forest

La qualité de l'ajustement est donnée par les paramètres suivants :

- AUC : 0,636
- RMSE : 0.000480

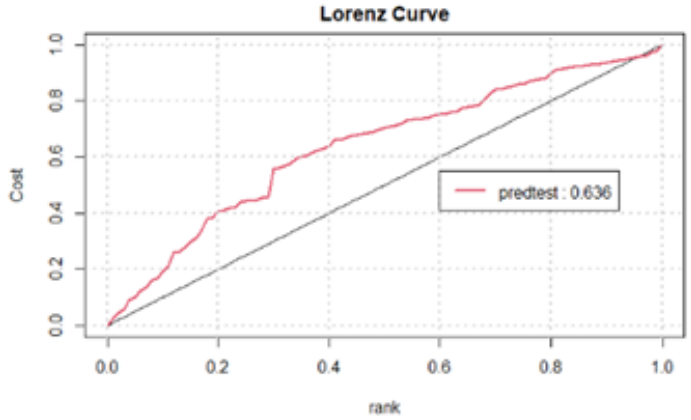


Figure 44 - Courbe de Lorenz du Random Forest

Le modèle n'est pas optimal mais confirme le choix du niveau d'acception du risque comme critère hiérarchique.

5.2.4 Le Gradient Boosting Machine (GBM)

5.2.4.1 La descente du gradient

Beaucoup d'algorithmes de Machine Learning obtiennent de très bonnes performances à partir de la multiplication des entraînements. Mais une majorité de ces entraînements repose sur l'optimisation d'une fonction de perte. Plus sa valeur est petite, meilleurs sont les résultats de l'algorithme. Ainsi, l'algorithme de descente de gradient est un des nombreux moyens qui permettent de trouver le minimum (ou maximum) d'une fonction. Le but est d'optimiser les fonctions de perte afin de réduire les erreurs liées aux prévisions

Une compréhension raisonnable de cette fonction dépend de divers facteurs comme l'objectif d'optimisation. Dans le cadre de l'étude sur la classification des coûts, la fonction de perte est une mesure permettant de classifier les TRE défavorables.

La descente du gradient est un des nombreux algorithmes dits de descente. La formule générale est la suivante :

$$x_{t+1} = x_t - \eta \Delta x_t$$

où η le taux d'apprentissage et Δx_t la direction de descente. Cette classe d'algorithme a pour but à chaque itération d'avoir $f(x_{t+1}) \leq f(x_t)$, avec f une fonction convexe que l'on souhaite minimiser. L'algorithme de descente du gradient décide de suivre comme direction de descente l'opposé du gradient d'une fonction convexe f, ie $-\Delta f$. Ainsi le gradient d'une fonction indique sa croissance maximale à partir d'un point. Alors choisir l'opposé revient à prendre la pente la plus abrupte, dans l'objectif de minimiser la valeur de cette fonction. Voici son fonctionnement :

- Soit un point d'initialisation x_0 appartenant au domaine de f
- Calcul de $f(x_t)$
- Mise à jour les coordonnées : $x_{t+1} = x_t - \eta \Delta f(x_t)$ (*)
- Répétition des deux étapes ci-dessus jusqu'au critère d'arrêt

5.2.4.2 AdaBoost et GBM

Adaboost a été utilisé pour la première fois par Yoav Freund et Robert Schapire, et a remporté le prix Gödel en 2003. Les « weak learners » d'AdaBoost sont des arbres décisionnels à seulement 2 branches et 2 feuilles (aussi appelés souches) mais il peut être utilisé d'autres types de classificateur.

Les étapes de construction du premier « weak learner » noté w_1 :

- le même poids est attribué à chaque ligne de la base de données
- w_1 est entrainé de manière à maximiser le nombre de bonnes réponses
- $w_1$est noté en fonction de ses performances

Cette notation permet qu'un bon « weak learner » soit plus écouté qu'un mauvais « weak learner » lors du vote final. Dans le cadre d'un GBM, le système de notation est écarté, sa particularité est qu'il essaye de prédire à chaque étape non pas les données elles-mêmes mais les résidus.

Avantages	Inconvénients
– Robuste – Souvent le meilleur modèle possible – Optimise directement la fonction de coût – Réduction du biais et de la variance – Différentes fonctions de coût	– Peut surapprendre – Difficile à paramétrer – Manque de transparence – Lourdeur et intensité des calculs

Le GBM a une tendance au surapprentissage plus importante que le RF, l'optimisation des paramètres est donc d'autant plus importante dans ce cas. Les principaux paramètres disponibles sont :

- Shrinkage (ou learning rate) : coefficient qui permet de ralentir la descente du gradient. S'il est trop grand, le risque est de passer à côté de l'optimum. Plus il est petit, plus l'apprentissage est long et plus il faut augmenter le nombre d'itérations. Généralement, il est conseillé d'utiliser un coefficient de l'ordre de 0.1 pour une base de plus de 10.000 entrées
- Interaction depth : nombre de niveaux dans l'arbre ou nombre d'interactions
- Min terminal node size : nombre d'observations minimum dans un nœud
- Bag.fraction : permet d'ajouter une composante aléatoire en ne sélectionnant qu'une partie des variables à chaque étape. Cela permet d'améliorer sensiblement les performances prédictives du modèle tout en réduisant le temps de traitement

En reprenant les paramètres imposés au RF, l'algorithme donne les résultats suivants, le niveau d'acception du risque est à nouveau prépondérant :

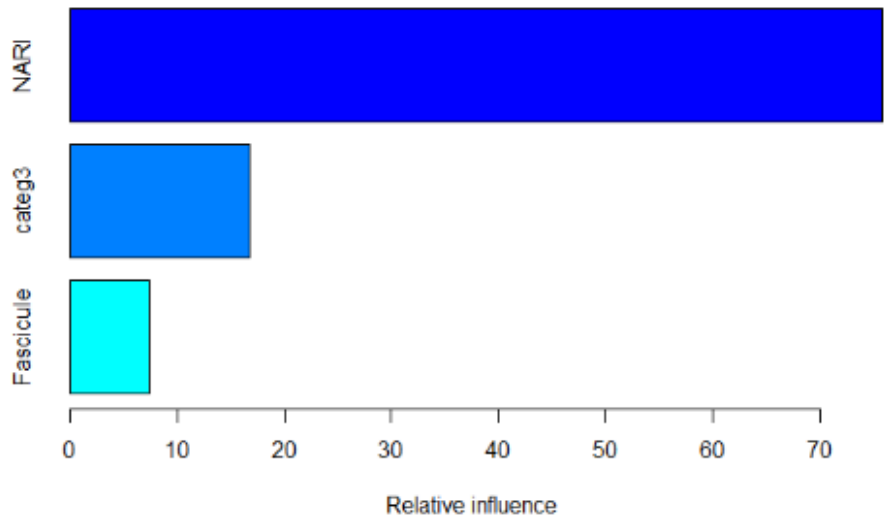


Figure 45 - Sortie du GBM

La qualité de l'ajustement est donnée par les paramètres suivants :

- AUC : 0,613
- RMSE : 0. 0.00018

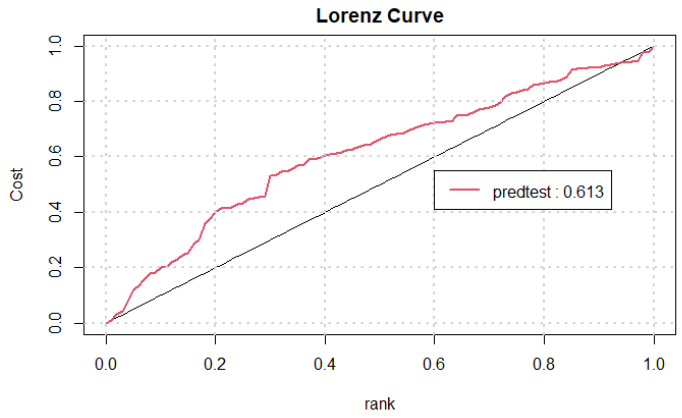


Figure 46 - Courbe de Lorenz du GBM

Le modèle n'est pas optimal mais confirme le choix du NARI.

5.3 L'AVANTAGE DES NIVEAU D'ACCEPTATION DU RISQUE INCENDIE

La segmentation utilisée pour son étude par France Assureurs est la segmentation « Fascicule », elle a pour pertinence de regrouper les activités par type d'activité, cependant le type d'activité n'est pas corrélé au niveau de risque. La variance au sein des groupes est très importante.

Les tableaux suivants présentent une comparaison des S/P ainsi que la dimension du portefeuille d'AXA entre une segmentation « Niveau d'Acceptation du Risque Incendie (NARI)» « catégorie » et « Fascicule » sur 10 ans.

5.3.1 Fascicule

Fascicule	ptf	SP
0	1 926	25,8%
1	5 187	36,7%
2	92 579	48,2%
3	16 344	61,7%
4	6 262	45,4%
5	9 335	55,6%
6	12 729	75,6%
7	20 252	69,1%
8	5 180	123,3%
9	174 589	47,3%
99	10 187	20,6%

Tableau 18 - S/P par fascicule

5.3.2 Niveau d'Acceptation du Risque Incendie (NARI)

NARI	ptf	SP
1	27 244	65%
2	36 573	63%
3	3 311	42%
4	101 972	55%
5	94 341	43%
6	91 384	43%

Tableau 19 - S/P par niveau d'acceptation du risque

5.3.3 Catégorie

Clé	ptf	SP
Cible	106 619	40%
Exclu	1 323	76%
Lourd	82 236	54%
LourdR	2 823	98%
Standard	109 404	53%
Supcible	52 485	48%

Tableau 20 - S/P par catégorie

L'utilisation des Niveau d'Acceptation du Risque Incendie (NARI) permet donc d'avoir un niveau hiérarchique discriminant et équilibrés en nombre de contrats.

Le choix du Niveau d'Acceptation du Risque Incendie (NARI) permet de conserver l'impact des fascicules (Groupe NARI 6 : les entrepôts) et l'impact des segments (Groupe NARI 1 : Les exclus).

5.4 APPLICATION ET ROBUSTESSE DES RESULTATS

5.4.1 Rappel de la méthode

La tarification du risque incendie des risques industriels est basée sur la théorie de la crédibilité, cette méthode est utilisée en tarification actuarielle pour estimer la prime dans un portefeuille hétérogène. Elle est construite sur le concept de facteur de crédibilité, qui mesure la crédibilité/confiance accordée à la prime individuelle à partir de l'historique de chaque contrat. Ce facteur de crédibilité va ainsi permettre de définir la part de d'information individuelle et la part d'information collective qui vont déterminer la prime pure du contrat.

Il existe différents modèles de crédibilité, l'étude se concentre en particulier sur les modèles de Buhlmann-Straub et de Jewell. Le modèle de Buhlmann-Straub introduit l'idée d'attribuer des poids aux observations des sinistres dans l'historique. Le modèle de Jewell ou modèle hiérarchique ajoute à cette idée une approche hiérarchique en appliquant la théorie de la crédibilité à chaque niveau de segmentation du portefeuille, la contrainte étant de segmenter le portefeuille en groupe de contrats similaires.

Le choix du niveau hiérarchique est réalisable soit par dire d'expert, qui présente la difficulté d'avoir une connaissance complète du portefeuille, ce qui à l'échelle d'un portefeuille comme celui d'AXA n'est pas réalisable, soit par méthodes actuarielles.

L'étude propose 3 approches, la première par ANOVA, complétée de 2 méthodes non paramétriques, le Random Forest de le Gradien Boosting. Les trois méthodes ont amené l'étude à orienter le choix du niveau hiérarchique vers le niveau de risque (NARI) qui est un indicateur de risque conçu par AXA.

Les risques industriels étant un secteur d'assurance marqué par des risques d'intensité, il est nécessaire de réaliser une politique d'écrêtement et de mutualisation des sinistres afin que les contrats avec des sinistres importants ne portent l'ensemble de l'information collective du portefeuille. La théorie des valeurs extrêmes permet de définir des seuils de sinistres qui correspondent à des niveaux de risques atypiques ou extraordinaires.

5.4.2 Choix du poids des contrats

Le principal élément manquant avant de lancer l'application avec les paramètres définis précédemment est le choix du poids pour le modèle de crédibilité.

Les travaux du mémoire « TARIFICATION DE L'INCENDIE DES RISQUES INDUSTRIELS » (Cohen Dupin et Levy, 1985) précise qu'il existe la possibilité de pondérer les données par le nombre de polices ou les capitaux assurés : *« Dans le premier cas, les rubriques constituées d'une multitude de risques à faibles capitaux sont surreprésentées (exemple: garages) tandis que des rubriques dont le tissu industriel est oligopolistique (exemple: construction automobile) deviennent négligeables. En conséquence les capitaux ont été retenus comme facteur de pondération, choix qui correspond d'ailleurs à celui des assureurs qui raisonnent d'ordinaire par rapport au franc assuré. »*

La pondération a été remise en cause dans les travaux « TARIFICATION DES RISQUES INDUSTRIELS » (DOUVILLE, 2004) , cette étude critique l'absence de la prise en compte des années de risques qui permet de mieux considérer les deux extrêmes (peu d'année et beaucoup de capitaux, peu de capitaux et beaucoup d'année police).

L'étude gardera donc la pondération proposée en 2004, la moyenne géométrique obtenue en faisant la racine carrée du produit des années-polices par les capitaux Incendie assurés.

5.4.3 Résultat sans écrêtement

Afin de montrer la pertinence de l'utilisation de la méthode d'écrêtement, l'étude présente ci-dessous le résultat sans l'utilisation de cette étape.

Les pseudo-estimateurs sont les suivants :

- $\widehat{\sigma_1^2}$ = 367,55
- α = 7,8%
- β = 2,8%

Ce résultat est obtenu après 81 itérations.

La valeur de l'estimateur $\widehat{\sigma_1^2}$ de la variabilité des observations individuelles est très élevée. La particularité du risque d'intensité du risque industriel entraine des années avec des poids très importantes sans écrêtement.
Ce résultat justifie la méthode d'écrêtement-mutualisation. L'information n'est pas bien répartie, le tarif est donc trop collectif et il n'engendre pas une bonne répartition des primes.

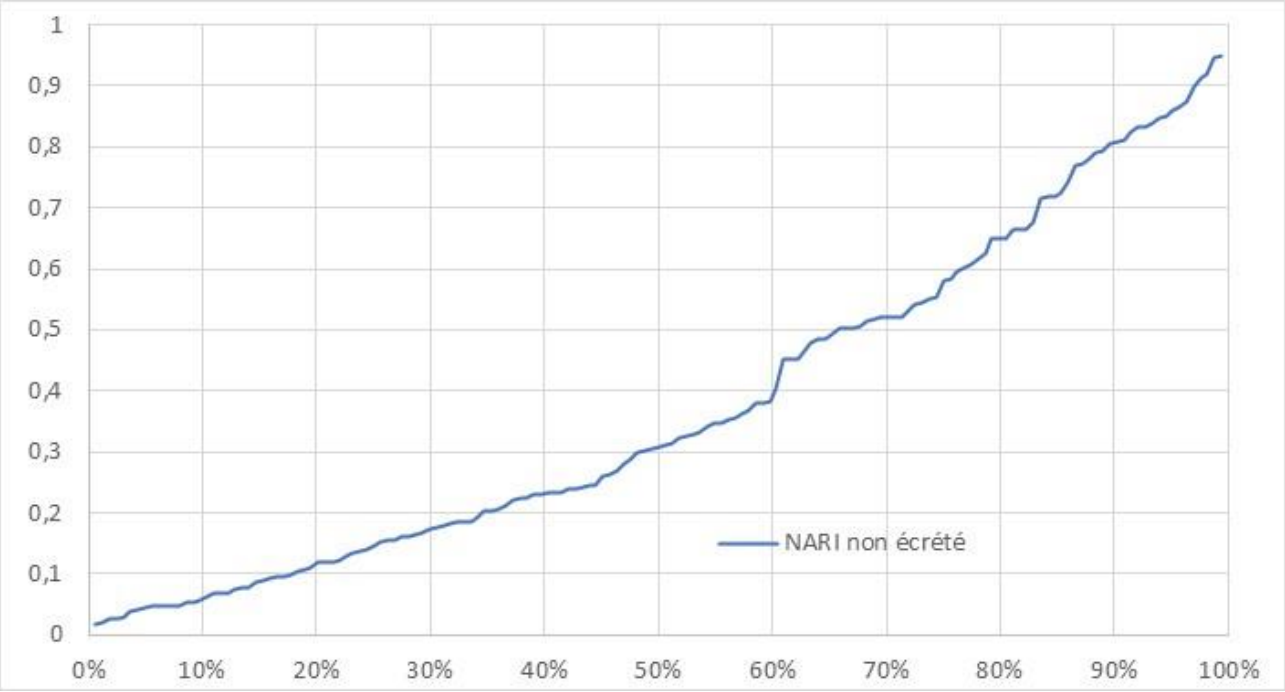


Figure 47 - distribution des ZI (non écrété)

Les Z_i sont faibles pour un grand nombre de TRE, le tarif porte donc trop d'information individuelle. La mutualisation du risque n'est pas efficace dans ce cas.

5.4.4 Résultat avec écrêtement- mutualisation.

A partir des seuils choisis pour donner suite à l'étude grâce à la théorie des valeurs extrêmes et aux travaux pour le choix du niveau hiérarchique, les estimateurs sont les suivants :

$\widehat{\sigma_1^2}$ = 90,52

α = 16,4%

β = 4%

Ce résultat est obtenu après 36 itérations.

La valeur de l'estimateur $\widehat{\sigma_1^2}$ de la variance des observations individuelles est inférieure par rapport à l'essai précédent. La méthode d'écrêtement-mutualisation apporte donc 2 avantages :

- Un nombre d'itération plus faible
- Un meilleur équilibre de l'information individuelle-collective

La distribution des Zi est la suivante :

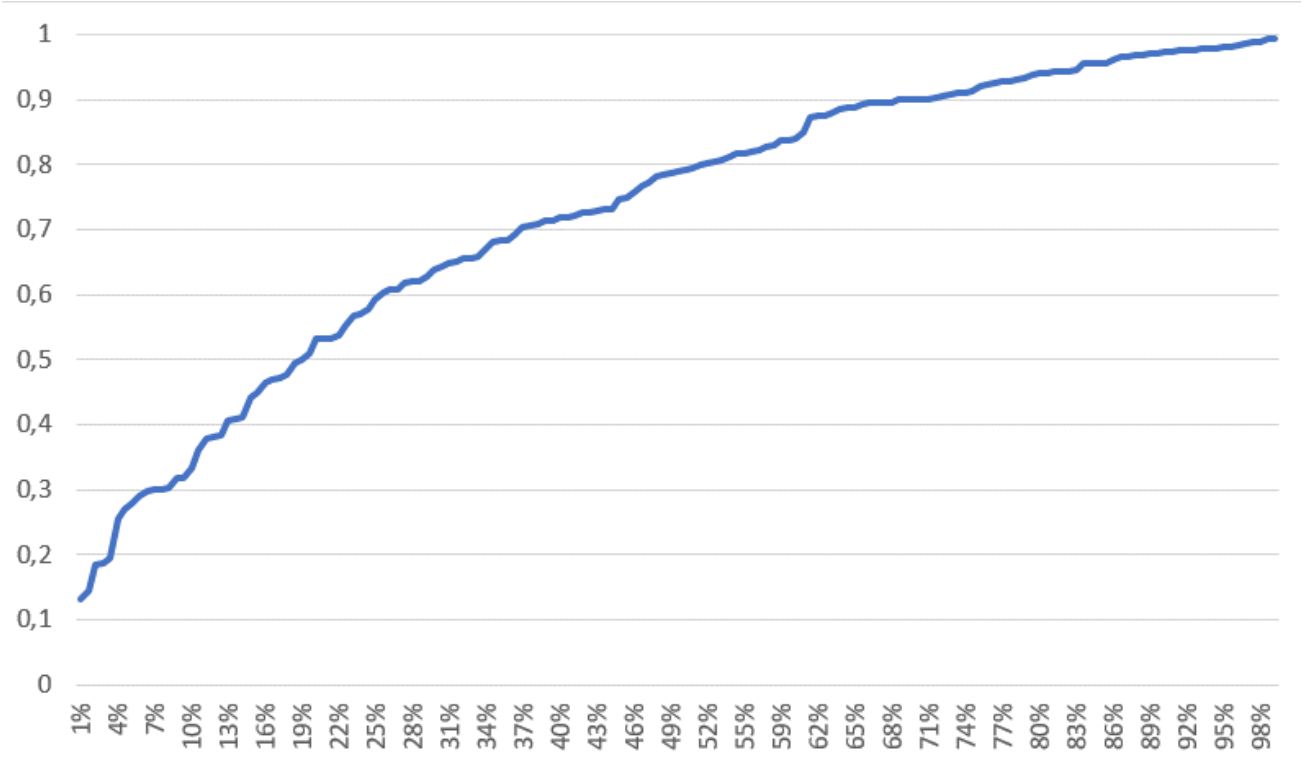


Figure 48 - distribution des Zi

Les coefficients de crédibilité Zi calculés sont mieux répartis et plus significatifs avec la méthode d'écrêtement-mutualisation.
Sans écrêtement 80% des Zi sont inférieurs à 0,7 contre 35% sinon. Le poids de l'information collective est donc mieux réparti.

5.4.5 Modélisation avec les différents seuils

Afin de vérifier la pertinence des choix, l'étude a été réalisée sur les seuils qui ressortaient.

Test	Seuil 1	Seuil 2	Seuil 3	$\widehat{\sigma_1^2}$	α	β
1	1 200 000	2 500 000	6 000 000	90,52	14,62%	4,01%
2	1 200 000	3 000 000	6 000 000	90,52	16,33%	4,00%
3	1 200 000	2 500 000	7 000 000	90,52	14,62%	4,08%
4	1 200 000	3 000 000	7 000 000	90,52	16,33%	4,08%
5	1 600 000	2 500 000	6 000 000	127,60	11,15%	3,71%
6	1 600 000	3 000 000	6 000 000	127,60	12,82%	3,74%
7	1 600 000	2 500 000	7 000 000	127,60	11,15%	3,79%
8	1 600 000	3 000 000	7 000 000	127,60	12,82%	3,82%

Le choix pour le troisième seuil entre 6 M€ et 7 M€ a peu d'impact, cependant le seuil 1 entre 1,2 M€ et 1,6 M€ a beaucoup d'impact sur la variance intra-cohorte.

Comme le confirme le graphique suivant, le choix du seuil 1 à 1,2 M€ permet de limiter l'impact des années exceptionnelles et donc de limiter l'impact individuel (A noter que le test 2 et 4 sont quasi-identique.) :

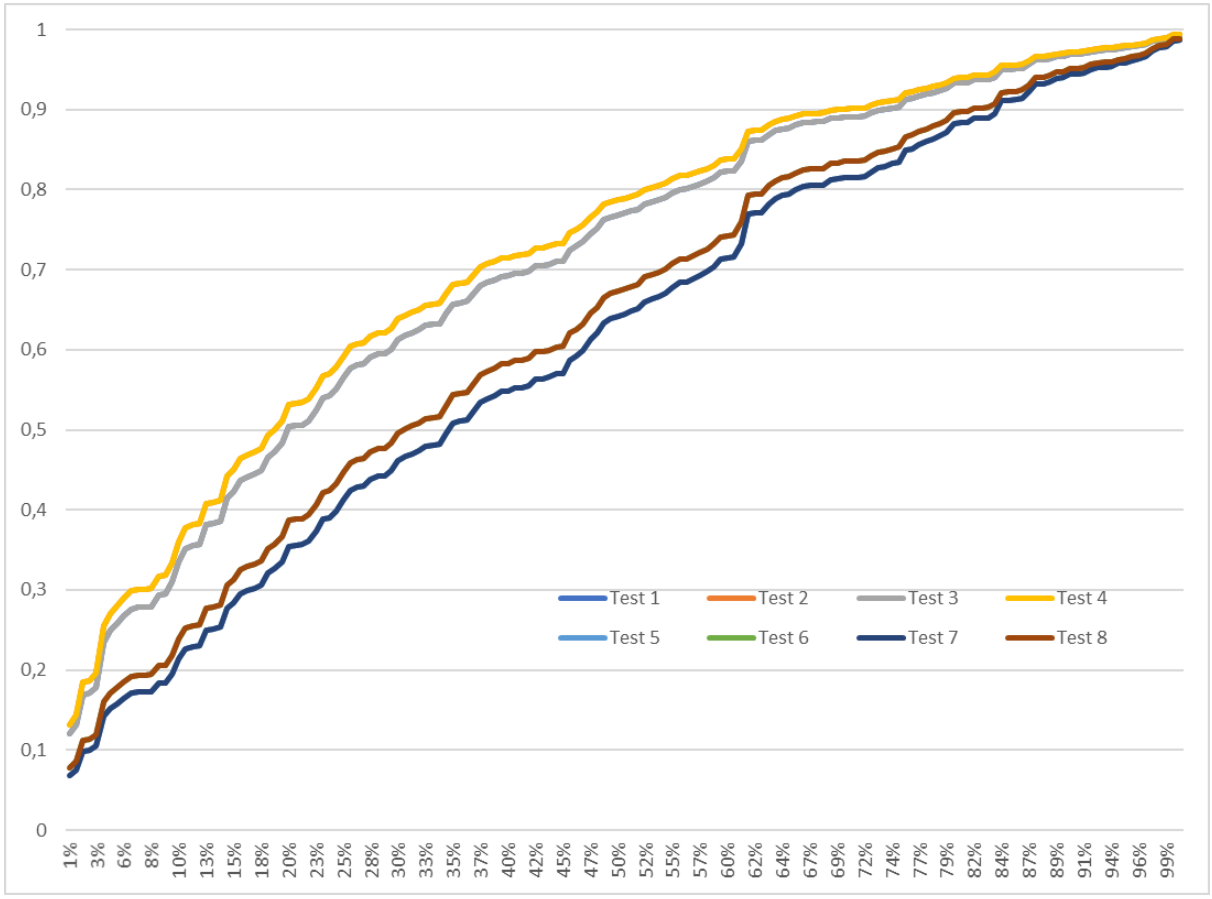


Figure 49 - Distribution des Zi par test de seuils

5.4.6 Robustesse des résultats

Dans ses travaux « TARIFICATION DES RISQUES INDUSTRIELS PAR LE MODELE DE CREDIBILITE » (DOUVILLE, 2004) », Cecile Douville propose de confirmer la robustesse des résultats en calculant pour chacune des rubriques 4 TPP crédibilisés sur des périodes glissantes de l'historique. Comme indicateur de cette stabilité, le coefficient de variation calculé sur ces 4 périodes permet d'identifier la stabilité des travaux.

Le coefficient de variation, également connu sous le nom de coefficient de variation de Pearson, est une mesure statistique qui renseigne sur la dispersion relative d'un ensemble de données, il se calcule par la formule suivante :

$$\sigma_{pearson} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

Notre étude va se baser sur 5 périodes glissantes de 5 ans $n \in \{1, 2, 3, 4, 5\}$ et $\bar{x}$ est la moyenne des 6 taux de primes pure crédibilisés pour une activité.

Inf à 5%	]5%;10%]	]10%;15%]	]15%;20%]	]20%;30%]	]30%;50%]	Sup à 50%
93	39	12	8	2	6	3
57,1%	23,9%	7,4%	4,9%	1,2%	3,7%	1,8%

Le test montre une stabilité importante ($\sigma_{pearson}$< 10%) pour 80% des rubriques. Cette stabilité est en forte corrélation avec les méthodes d'écrêtement-mutualisation qui permettent un lissage de la sinistralité.

5.4.7 Comparaison avec la tarification actuelle et appréciation.

Afin de modéliser l'impact de la nouvelle tarification, le graphique ci-dessous montre l'évolution entre les taux actuels et le taux crédibilisés chargés pour les 20 rubriques les plus importantes en poids sur les 11 années d'historique. Les TRE ci-dessous sont anonymisés.

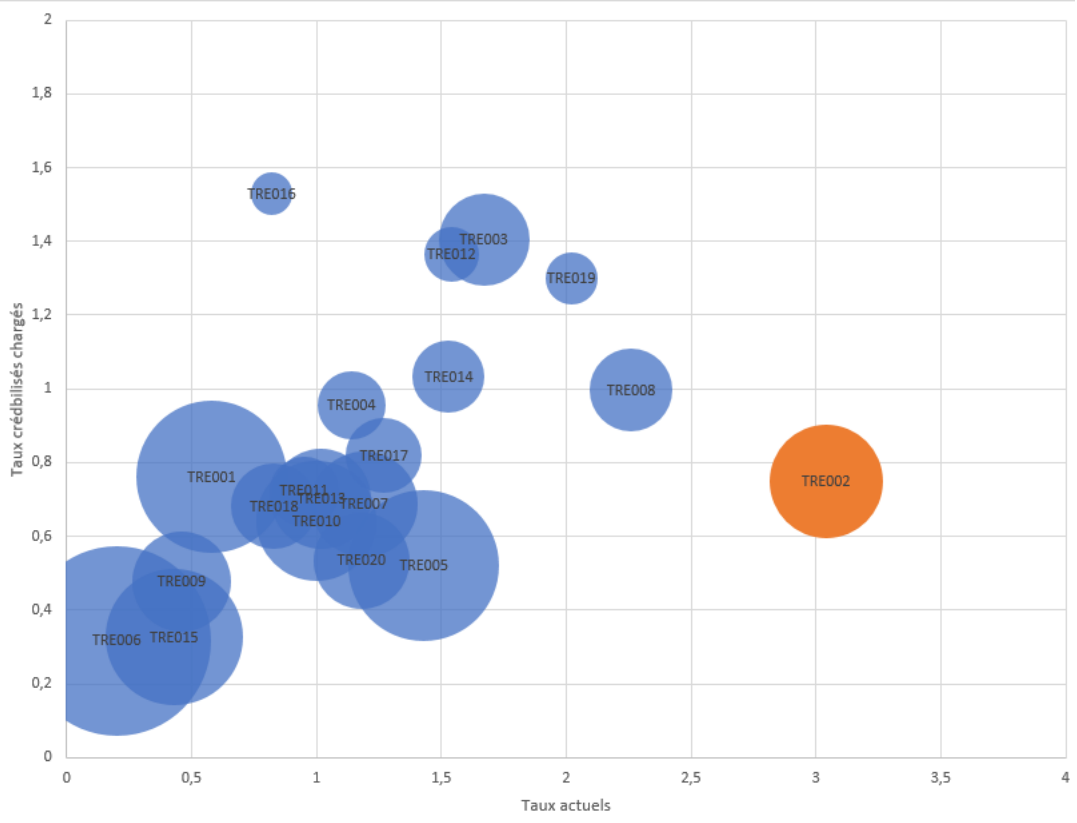


Figure 50 - Evolution tarifaires

La rubrique 2 (TRE002) passe d'un taux de 3 à 0,75. L'étude du S/P incendie par année donne le résultat suivant :

Année	S/P incendie
2012	0,09
2013	0,14
2014	0,10
2015	0,51
2016	0,04
2017	0,64
2018	0,48
2019	1,80
2020	0,57
2021	1,37

Cette rubrique est fortement impactée par deux années exceptionnelles, l'écrêtement a permis de limiter l'effet sur la tarification (hors écrêtement le taux est de 1,10). Cependant le passage de la prime pure actuelle à celle obtenue par ces travaux engendrerai un risque au vu de l'augmentation de l'intensité des sinistres sur les années récentes.

L'application de la théorie de la crédibilité hiérarchique après écrêtement et mutualisation permet bien à AXA France une remise à plat de son tarif incendie de la branche risque industriel.

La prime pure qui ressort du modèle étudié permet d'ajuster la tarification actuelle avec une meilleur qualité qu'une simple étude du SC, qui ne prend pas en compte l'information collective. La refonte voulu par AXA de la tarification du risque industriel ne peut se limiter à l'incendie, un travail équivalent sur les garanties annexes (pertes d'exploitation, bris de glace…) permettra une réelle remise à plat de la tarification.

6 CONCLUSION

L'étude des risques incendie est un exercice difficile. La forte spécificité de la sinistralité avec des fréquences de sinistres faibles mais à forte intensité ne permet pas l'application des modèles de tarifications contemporains comme les modèles paramétriques (GLM) ou non paramétrique (le Machine Learning) pour la tarification des rubriques d'activités.

L'utilisation de la crédibilité a largement contribué à une meilleure pénalisation des rubriques d'activités. Cependant le choix arbitraire des niveaux hiérarchiques et des seuils de mutualisations de sinistres biaise les études.

La théorie des valeurs extrêmes pour l'étude de la sinistralité et l'utilisation de modèles non paramétriques limité aux choix de niveaux hiérarchiques permettent une amélioration des résultats de la crédibilité.

L'application de ces 3 étapes de tarification permet à AXA une véritable maitrise et connaissance de son exposition aux risques incendies aux niveaux des rubriques d'activités.

La théorie des valeurs extrêmes permet une meilleure connaissance des intensités de risques, ce qui dans la continuité des travaux de cette étude permettra de revoir rapidement les classifications commerciales (ou segments « cible, standard… ») avec une remise à plat du seuil de mutualisation des sinistres extraordinaires actuellement utilisé chez AXA France.

Le seuil actuel de 1,2 M€ entrainant une trop forte mutualisation entre rubriques d'activités, alors que l'étude montre un seuil pertinent proche des 2,5 M€. La revue de ce seuil va permettre de mettre en exergue les rubriques fortement touchées.

L'apport de l'ANOVA et la confirmation par les modèles non paramétriques permet déterminer avec pertinence les niveaux hiérarchiques utilisées pour la théorie de crédibilité hiérarchique. Une ouverture possible de cette étape est l'utilisation de ces méthodes pour la création de regroupement de rubrique d'activité (TRE), cependant cela complexifierait les échanges entre les différents services des Risques Industriels, chaque service ayant à force une segmentation différentes (Segment, NARI, fascicule, segmentation tarifaire…).

L'étude se limite pour le moment à la remise à plat des taux de prime pure des rubriques d'activités (TRE), la continuité de ces travaux sera la remise à plat des coefficients de pénalisation des études de prévention et de protection sur l'incendie permettant la personnalisation du tarif de la rubrique à l'entreprise (Note de qualité incendie). Pour cela il est important d'avoir un historique conséquent pour appliquer des modèles paramétriques adaptés, ce qui fut l'élément bloquant pour rentrer dans l'objectif de l'étude.

Les évolutions rapides des algorithmes de machine Learning vont ouvrir des nouvelles opportunités de tarifications, du fait des améliorations de prédictions à partir de faibles échantillons de données et de l'augmentation des historiques de sinistres.

Les événements récents de crise énergétique font apparaitre de nouveaux risques sur les industries, la notion de blackout, c'est-à-dire, une coupure généralisée de l'approvisionnement en électricité sur toute ou partie d'un territoire, est de plus en plus étudiée par l'actuariat au sein d'AXA France.

7 BIBLIOGRAPHIE

Cary Chi-Liang Tsai, Adelaide Di Wu. *Application du modèle de crédibilité hiérarchique à la modélisation des taux de mortalité de plusieurs populations.*

Charpentier & Denuit. 2004. *Mathématiques de l'assurance non vie - Volume 1*.

Charpentier & Denuit. 2004. *Mathématiques de l'assurance non vie - Volume 2*.

Charpentier, Arthur. 2017. *ACTUARIAT DE L'ASSURANCE NON-VIE #6.* 2017.

Cohen Dupin et Levy. 1985. *TARIFICATION DE L'INCENDIE DES RISQUES INDUSTRIELS*.

Desaegher, Caroline. *L'histoire d'Axa de Caroline Desaegher.*

DOUVILLE, Cécile. 2004. *TARIFICATION DES RISQUES INDUSTRIELS*.

Hans Bühlmann, Alois Gisler. 2005. *A course in Credibility Theory and its Applications*.

Howard C. Malher, Curtis Gary Dean. 1971. *Credibility*.

Journal de la société statistique de Paris. 1995. *La théorie des valeurs extrêmes : présentation et premières applications en finance.* 1995.

Le « Tarif rouge » est dépoussiéré pour encourager la prévention. **2004**, Argus de l'assurance.

Légifrance. [En ligne] https://www.legifrance.gouv.fr/codes/id/LEGITEXT000006073984/.

Models for exceedances over high thresholds. **Davison A., Smith R. 1990**.

Novaro, Yoann. *TPE Statistique des Extrêmes.*

Regression model with scalar credibility weights. **F., DE VYLDER. 1981**.

Thérond, Pierre. Année universitaire 2017-2018. *Théorie de la crédibilité.* Année universitaire 2017-2018.

Tableau des figures