**Mémoire présenté le :**

**pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA**

**et l'admission à l'Institut des Actuaires**

Par : Pierre Mourier

Titre Ultimates as intervals in reinsurance

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

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

*Membres présents du jury de l'Institut des Actuaires* Signature

*Entreprise :*

Nom : QBE Re

*Signature :*

*Membres présents du jury de l'ISFA*

*Directeur de mémoire en entreprise :*

*Nom : Vincent Daxbek*

*Signature :* Vincent Daxbek Analytics Manager QBE — Digitally signed by Vincent Daxbek Date: 2021.03.30 16:39:47 +02'00'

*Invité :*

*Nom : Aurélien Dubois*

*Signature :* Aurélien Dubois Head of Technical & Analytics QBE — Digitally signed by Aurelien Dubois Date: 2021.03.30 09:50:37 +02'00'

***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

Vincent Daxbek Analytics Manager QBE — Digitally signed by Vincent Daxbek Date: 2021.03.30 17:11:35 +02'00'

Signature du candidat

Institut de Science Financière et d'Assurances

M2 Actuariat

# Ultimates as intervals in reinsurance

*Auteur :*
Pierre Mourier
*Tuteur ISFA :*
Stéphane Loisel

*Tuteurs entreprise :*
Vincent Daxbek
Aurélien Dubois

# Résumé

**Mots-clefs :** Réassurance, Tarification, Développement de sinistre, apprentissage non-supervisé.

Les réassureurs sont confrontés à un manque de données et de qualité de ces données. Ils doivent pourtant réaliser des études de tarification complexes en vue d'estimer au mieux le prix d'une couverture de réassurance. L'objectif de ce mémoire est de développer une nouvelle méthode d'estimation du coût ultime des sinistres de réassurance en vue d'augmenter la qualité de la tarification. Dans un premier temps, ce mémoire s'intéressera à l'impact de l'estimation de la valeur ultime du sinistre lors de la tarification, puis à la méthode classique utilisée en réassurance : Chain Ladder. Ensuite, en utilisant des méthodes d'apprentissage non-supervisé, une nouvelle méthode d'estimation du coût ultime des sinistres sera développée. Celle-ci aura pour but de donner plus d'information sur les possibilités de développement en ultimes des sinistres. L'étude de la pertinence de cette méthode permettra de conclure sur l'intérêt de cette nouvelle manière de développer des sinistres en leur situation ultime.

# Abstract

**Keywords:** Reinsurance, Underwriting, Claims development, clustering.

Reinsurers face with a lack of data and data quality. Nevertheless they must carry out complex pricing studies in order to best estimate the price of reinsurance coverage. The aim of this master thesis is to develop a new method for estimating the ultimate cost of reinsurance claims, with a view to improving the quality of pricing. In this paper, I will first focus on the impact of estimating the ultimate value of the claim at the time of pricing, and then on the method traditionally used in reinsurance: Chain Ladder. Then, using unsupervised learning methods, I will develop a new method for estimating the ultimate cost of claims. This method will aim to provide more information on the possibilities of ultimate claims development. The study of the relevance of this method will allow me to conclude on the interest of this new way of developing claims in their ultimate situation.

# Remerciements

Je tiens tout d'abord à remercier Vincent Daxbek, mon tuteur au sein de QBE Re, pour son immense investissement tout au long de mon stage et de mon alternance. Sa pédagogie, ses conseils et son expérience m'ont été d'une très grande aide pour comprendre le monde de la réassurance et les spécificités de mes sujets d'étude. Je tiens à le remercier également pour la relecture de ce mémoire.

Je remercie également Aurélien Dubois pour m'avoir offert l'opportunité d'effectuer mon alternance à QBE Re, mais aussi pour son encadrement lors de la réalisation de ce mémoire.

Je souhaite également remercier tous mes collègues à QBE Re pour leur accueil chaleureux, leur bienveillance et leur disponibilité tout au long de mon alternance.

Je remercie mon tuteur Stéphane Loisel pour les conseils qu'il m'a donnés ainsi que pour ses relectures.

Je remercie tout particulièrement Yannis BIOLAY, mon colocataire préféré qui a essayé des mois durant de m'apprendre à écrire avec style. Les nombreuses lettres d'excuse qu'il m'a rédigées ont été des exemples que, malheureusement, je suis incapable d'égaler.

# Table des matières

# Introduction

Lorsqu'une entreprise de réassurance souhaite tarifer un contrat de réassurance en excédent de sinistre, celle-ci reçoit de la part de la cédante un triangle de sinistres. Ce triangle est constitué de l'entièreté des sinistres dont le coût historique aurait, au moins une année, dépassé un seuil appelé la priorité des statistiques. La priorité des statistiques est définie contractuellement et est souvent exprimée en un pourcentage de la priorité du contrat (généralement 50% ou 75%).

Ces données nous montrent que le coût estimé des sinistres entre leur année de déclaration et celle de clôture est très variable. Ces variations peuvent être dues à une mauvaise estimation initiale du sinistre, à un changement de l'état de la victime, à un jugement d'un tribunal à propos de la partie responsable, à un changement de la législation, etc. Or, le réassureur a besoin d'estimer le coût ultime qu'un sinistre représentera une fois clôturé pour réaliser la tarification. Ce processus d'estimation s'appelle dans le jargon « le développement en ultime ». Cette estimation se doit donc d'être précise, car elle servira de base à la tarification du contrat de réassurance. Elle est l'élément central du modèle de risque collectif : d'un côté par un impact direct sur la sévérité, et d'un autre côté par un impact indirect sur la fréquence.

Une méthode classique de développement des sinistres en situation ultime utilisée en réassurance s'appelle « Chain Ladder ». Des coefficients de passage sont calculés entre chaque année de développement et offrent la possibilité d'être ajustés par un technicien selon son expertise. Ensuite, ces coefficients sont appliqués à chacun des sinistres en vue d'obtenir une estimation de leur coût ultime.

Cette méthode d'estimation de la valeur ultime des sinistres est problématique sur plusieurs points. Premièrement, chaque sinistre se voit attribué le même coefficient par année de développement sans prendre en compte ses caractéristiques particulières telles que son montant actuel estimé, sa nature (matériel ou corporel), la situation de la victime, etc. Deuxièmement, la méthode de Chain Ladder est une méthode de provisionnement qui s'applique à des données agrégées et qui permet seulement de déduire une évolution moyenne des sinistres. Troisièmement, le fait d'appliquer une évolution moyenne commune à chaque sinistre fait que nous obtiendrons une valeur discrète une fois le sinistre développé. Ce résultat déterministe peut être problématique dans l'estimation de la fréquence qui s'opère au-delà d'un seuil.

Dans ce mémoire une nouvelle méthode d'estimation de la valeur ultime des sinistres est proposée. L'objectif de cette nouvelle méthode est de pouvoir développer des sinistres en situation

ultime sous forme d'intervalle de probabilité en estimant la fonction de répartition des valeurs ultimes. Ce nouveau résultat permettra notamment de résoudre les problèmes rencontrés lors de l'estimation de la fréquence avec des valeurs ultimes déterministes proches du seuil d'analyse.

Pour ce faire, nous créerons de nouvelles variables qui seront appelées des « Métriques » et qui seront basées sur l'historique des montants estimés et des paiements du sinistre. Ces métriques permettront de décrire le comportement du sinistre, ce qui sera une réelle augmentation de la qualité des données utilisées lors de nos estimations.

Ensuite, des techniques de *clustering* seront appliquées aux données de manière à classifier les sinistres selon leur comportement. A la suite de cela, les différentes évolutions individuelles des sinistres parmi ces groupes seront combinées en vue d'estimer la fonction de répartition correspondante à chaque groupe à la place de coefficients de développement communs à tout l'échantillon.

Finalement, les résultats obtenus seront comparés aux résultats de la méthode actuelle : Chain Ladder. Nous mettrons aussi en place une comparaison des résultats avec ceux obtenus en utilisant le modèle de Mack, qui serait l'amélioration la plus directe de la méthode de Chain Ladder.

L'exploitation des résultats de cette nouvelle méthode de développement impose des changements sur la manière dont les paramètres des lois de probabilité nécessaires à la modélisation du risque tarifé sont estimés. C'est pour cela que ce mémoire a été réalisé en étroite collaboration avec BIBER L., qui a travaillé sur la manière d'estimer des paramètres à partir d'un ensemble d'intervalles de probabilité dans le mémoire « *Study of the Body and the Tail* » (Biber, 2021)[1]. Ces deux mémoires permettent d'aboutir à une nouvelle méthode de tarification plus précise des contrats de réassurance.

Après une conclusion, nous nous pencherons sur les futurs développements de cette approche proposée.

# Chapitre 1 : Introduction et motivations

L'objectif de ce chapitre introductif est de contextualiser l'utilisation de l'estimation du montant ultime des sinistres. Pour cela, j'introduirai les concepts clefs de la réassurance non-vie en commençant par une présentation générale de la réassurance, notamment des contrats non-proportionnels. Ensuite, je détaillerai les risques couverts par le réassureur qui seront étudiés dans ce mémoire. Enfin, je présenterai l'entreprise qui m'a accueilli pour réaliser ce mémoire : QBE Re.

Les données de réassurance seront aussi explicitées, ceci de manière à cadrer au mieux les possibilités et limitations de celles-ci. Nous en profiterons pour étudier la méthode de mise en « AS-IF » des données, qui est une étape extrêmement importante dans la tarification d'un contrat de réassurance.

Ensuite, nous présenterons la méthode de tarifications utilisée pour les contrats en excédent de sinistre : le modèle de risque collectif. Nous séparerons cette présentation entre l'estimation de la sévérité et l'estimation de la fréquence.

Enfin, grâce à cette mise en contexte nous pourrons pleinement exprimer la problématique du développement de sinistre. Nous pourrons alors étudier son impact et, par conséquent, l'importance d'avoir l'estimation la plus juste possible.

## 1.1 Introduction à la réassurance

### 1.1.1 Présentation générale

La réassurance est classiquement définie comme « l'assurance de l'assurance ». Il s'agit pour une société d'assurance (appelée cédante) de céder à une société de réassurance un risque présent dans son portefeuille de clients qu'elle ne souhaite plus assumer, et cela, en échange d'une prime. Le mode de transfert des risques est alors prévu par un contrat appelé traité de réassurance. Ces contrats ont la plupart du temps une période de couverture d'un an. Dans les faits, un assureur achète de la réassurance pour faire face aux sinistres les plus extrêmes qui pourraient remettre en cause sa solvabilité. Par exemple, un assureur peut se protéger des conséquences d'une catastrophe naturelle. La réassurance peut aussi avoir pour rôle d'aider une cédante à lancer un nouveau produit via des traités proportionnels. Dans ce cas, la prime et les coûts des sinistres sont partagés de manière proportionnelle entre la cédante et le réassureur.

On peut distinguer deux branches principales en réassurance : la Vie et la Non-Vie. Ces branches sont ensuite séparées en deux modes de réassurance : la réassurance obligatoire et la réassurance facultative. La réassurance obligatoire couvre le portefeuille entier de la cédante dans une sous-branche (habitation, auto, etc.). La réassurance facultative couvre certaines polices spécifiques de la cédante explicitement décrites, car ces risques sont atypiques (la couverture de voitures de luxe peut, par exemple, faire l'objet d'un traité facultatif).

Enfin il y a plusieurs formes de réassurance : la réassurance proportionnelle et la réassurance non-proportionnelle. La réassurance proportionnelle consiste pour le réassureur à recevoir une proportion de la prime en échange de la prise en charge de la même proportion des sinistres. La réassurance non proportionnelle consiste pour le réassureur à intervenir seulement lorsque les pertes ou les sinistres dépassent un seuil.

Ce mémoire portera exclusivement sur la partie Non-vie (plus particulièrement sur les branches à développement long) et pour des contrats de réassurance à forme non-proportionnelle.

### 1.1.2 Réassurance non-proportionnelle – Excédent de sinistre

La réassurance non-proportionnelle permet de protéger la solvabilité des assureurs qui pourrait être affectée par une trop grande sinistralité. Dans cette forme de réassurance, l'assureur supporte tous les sinistres jusqu'à un seuil appelé priorité (ou *deductible*). Lorsque la priorité est dépassée, le réassureur prend alors en charge les pertes au-dessus de ce seuil. Souvent, la prise en charge au-dessus du seuil est limitée à un certain montant, que l'on appelle la couverture (*cover*).

Il y a deux formes de réassurance non-proportionnelle. Premièrement, l'excédent de perte (ou *Stop-loss*) est une forme de réassurance non-proportionnelle qui intervient lorsque le ratio $\frac{sinistre}{prime}$ d'une année entière dépasse un seuil, et ce, jusqu'à la limite. Deuxièmement, l'excédent de sinistre - XS (*Excess of Loss – XL*) est la deuxième forme de réassurance non-proportionnelle. C'est la plus courante en réassurance. Le réassureur indemnise la cédante pour le montant du sinistre entre la priorité et la limite. On note D la priorité et C la couverture (*cover*), avec $limite = C + D$. Le contrat en excédent de sinistre est noté C xs D. Pour un sinistre de montant Xi, le réassureur devra indemniser la cédante d'un montant $R_i$, tel que :

$$R_i = \min(max(0, X_i - D), C)$$

C'est dans le cas de la réassurance par excédent de sinistre que l'estimation de la valeur ultime du sinistre est une étape particulièrement importante. En effet, dans ce cas le réassureur indemnise la cédante sinistre par sinistre. En vue de connaître le risque supporté, le réassureur doit avoir la meilleure estimation des anciens sinistres.

Prenons un exemple pour illustrer ce mécanisme : Supposons un contrat de réassurance en deux tranches :

- Tranche 1 : 2.000.000 xs 2.000.000
- Tranche 2 : 2.000.000 xs 4.000.000

Avec les sinistres suivants :

| Montant des sinistres | Rétention à la charge de la cédante (en deca de la priorité) | Coût pour le réassureur Tranche 1 | Coût pour le réassureur Tranche 2 | Reste à la charge de la cédante (au-dessus de la limite) |
|---|---|---|---|---|
| 1 000 000 | 1 000 000 | 0 | 0 | 0 |
| 3 000 000 | 2 000 000 | 1 000 000 | 0 | 0 |
| 5 000 000 | 2 000 000 | 2 000 000 | 1 000 000 | 0 |
| 7 000 000 | 2 000 000 | 2 000 000 | 2 000 000 | 1 000 000 |

Nous pouvons aussi représenter ces sinistres visuellement, ce qui rend l'exemple plus clair :

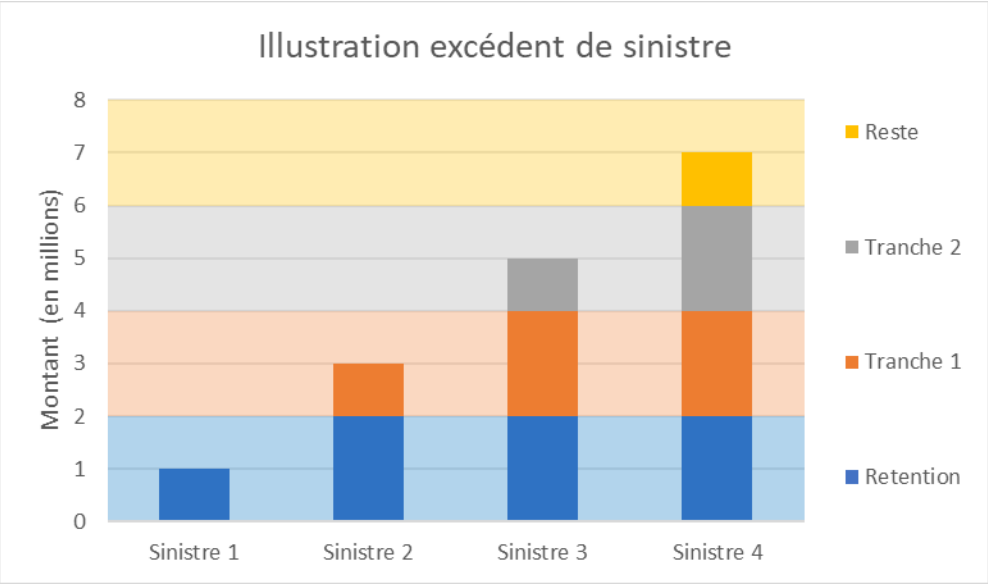


*Figure 1 - Illustration Excédent de sinistre*

On retrouve deux types de couvertures dans un traité en excédent de sinistre. Les couvertures par risque ou celles par événement.

- **Par risque** : le réassureur indemnise la cédante lorsqu'un risque se matérialise et vient à toucher le réassureur, car les dégâts occasionnés dépassent la priorité du contrat (e.g. un accident de voiture)
- **Par événement** : le réassureur intervient lorsque les dommages totaux causés par un même événement touchant plusieurs risques en portefeuille (e.g. une tempête qui endommage plusieurs maisons) dépassent la priorité. Cela oblige à définir la notion d'événement, ce qui est alors fait dans le traité de réassurance.

En réalité, la cédante crée selon ses besoins sa propre structure de réassurance. Ceci en cumulant plusieurs types de couverture et plusieurs tranches de réassurance et parfois de manière illimitée lorsque la loi l'impose. C'est par exemple le cas pour les assurances auto obligatoires en France et en Belgique qui couvrent les dégâts des tiers. On peut aussi noter qu'il est commun que les réassureurs ne participent à une structure de réassurance qu'à hauteur d'un pourcentage défini contractuellement de certains sous-contrats composant la structure. Ceci permet aux réassureurs de diversifier plus facilement leurs risques.

De plus, il est important de comprendre qu'il y a un effet de levier dans une tranche de réassurance en excédent de sinistre. En effet, la charge pour le réassureur (calculée en proportion de la prime de réassurance) augmente plus vite que l'augmentation du coût du sinistre.

Par exemple, pour un contrat $100\ xs\ 100$ :

| | Montant | Coût pour le réassureur |
|---|---|---|
| Sinistre 1 | 150 | 50 |
| Sinistre 2 | 175 | 75 |

Le montant du sinistre 2 est 16,66% plus élevé que celui du sinistre 1 mais le coût pour le réassureur est 50% plus élevé. C'est à cause de cet effet de levier qu'il est aussi extrêmement important d'avoir la meilleure estimation du montant ultime du sinistre en réassurance.

### 1.1.3 Présentation de QBE Re

QBE Re est une filiale du groupe australien d'assurance QBE. QBE Re est un réassureur vie et non-vie international dont les bureaux se situent à Bruxelles, Londres, Dublin, New York et Dubaï. Le bureau de Bruxelles où j'ai été accueilli pour y réaliser mon alternance a été racheté à KBC, grand groupe de banque et d'assurance belge, en 2010 par le groupe QBE. QBE Re génère environ 1 milliard d'euros de prime brute annuelle, dont 30% provient du bureau de Bruxelles. La prime acquise dans le cadre de la réassurance non-vie non-proportionnelle en représente une partie importante, d'où la nécessité de maintenir et d'améliorer les méthodes de tarification. Le bureau de Bruxelles est composé de 85 personnes, dont environ 25 actuaires. L'activité d'actuariat est organisée en 3 départements, que sont le département Analytics (Recherche et Développement), le département vie, et le département non-vie. Le département Analytics s'occupe du développement des modèles et des outils, qui vont être utilisés par les départements actuariels pour calculer le prix technique des traités de réassurance. C'est au sein de ce département que ce mémoire a été écrit.

## 1.2 Données de réassurance

### 1.2.1 Données de renouvellement

Lorsqu'un réassureur souhaite se positionner sur un contrat, la cédante (ou le courtier) lui envoie ce que l'on appelle les données de renouvellement (*renewal package*). Ces données sont sous la forme de deux triangles. Le premier contient l'historique des valeurs estimées de chaque sinistre ayant au moins une fois dépassé la priorité des statistiques. Le second contient l'historique de paiement des sinistres présents dans le premier triangle.

| | Année de développement des sinistres | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Année d'accident | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 2009 | $C_{0,0}$ | $C_{0,1}$ | $C_{0,2}$ | $C_{0,3}$ | $C_{0,4}$ | $C_{0,5}$ | $C_{0,6}$ | $C_{0,7}$ | $C_{0,8}$ | $C_{0,9}$ | $C_{0,10}$ |
| 2010 | $C_{1,0}$ | $C_{1,1}$ | $C_{1,2}$ | $C_{1,3}$ | $C_{1,4}$ | $C_{1,5}$ | $C_{1,6}$ | $C_{1,7}$ | $C_{1,8}$ | $C_{1,9}$ | |
| 2011 | $C_{2,0}$ | $C_{2,1}$ | $C_{2,2}$ | $C_{2,3}$ | $C_{2,4}$ | $C_{2,5}$ | $C_{2,6}$ | $C_{2,7}$ | $C_{2,8}$ | | |
| 2012 | $C_{3,0}$ | $C_{3,1}$ | $C_{3,2}$ | $C_{3,3}$ | $C_{3,4}$ | $C_{3,5}$ | $C_{3,6}$ | $C_{3,7}$ | | | |
| 2013 | $C_{4,0}$ | $C_{4,1}$ | $C_{4,2}$ | $C_{4,3}$ | $C_{4,4}$ | $C_{4,5}$ | $C_{4,6}$ | | | | |
| 2014 | $C_{5,0}$ | $C_{5,1}$ | $C_{5,2}$ | $C_{5,3}$ | $C_{5,4}$ | $C_{5,5}$ | | | | | |
| 2015 | $C_{6,0}$ | $C_{6,1}$ | $C_{6,2}$ | $C_{6,3}$ | $C_{6,4}$ | | | | | | |
| 2016 | $C_{7,0}$ | $C_{7,1}$ | $C_{7,2}$ | $C_{7,3}$ | | | | | | | |
| 2017 | $C_{8,0}$ | $C_{8,1}$ | $C_{8,2}$ | | | | | | | | |
| 2018 | $C_{9,0}$ | $C_{9,1}$ | | | | | | | | | |
| 2019 | $C_{10,0}$ | | | | | | | | | | |

*Figure 2 - Représentation visuelle des données de renouvellement sur 11 ans*

Lorsque que l'on travaille sur des branches à développement long, dites « *long tail* », on reçoit en général un historique des 5 ou 10 dernières années. Les branches long tail correspondent aux branches où les sinistres peuvent se développer longtemps. Ce sont les branches qui peuvent être touchées par des dommages corporels : l'état de la victime évolue au cours du temps (et donc le coût du sinistre aussi). Mais cela peut aussi être d'autres risques qui durent ou évoluent dans le temps (assurance dommage ouvrage par exemple).

On peut ainsi déjà comprendre une des sources de difficulté du travail du réassureur : le manque de données. Dans un premier temps, le nombre de sinistres au-dessus d'un seuil peut être très faible, même quand les données des 10 dernières années sont disponibles. Ce manque de données se fait ressentir plus la priorité du contrat est élevée et il complexifie grandement le travail de tarification. De plus ces données sont très souvent sommaires, il n'y a seulement que deux variables par sinistre : le montant estimé et le paiement déjà effectué. Ce qui ne permet pas de développer une analyse précise et pertinente sur le développement des sinistres, tel que pourrait le faire un expert des règlements de sinistres.

Enfin, se pose la question de la pertinence de mélanger entre eux des sinistres étant survenus à plusieurs années d'intervalle.

### 1.2.2 Mise en AS-IF

Pour permettre de prendre en compte des sinistres d'années de survenance différentes, les réassureurs réalisent une mise en « *AS-IF* ». Cela permet de remettre les sinistres sur une base actualisée commune : la réactualisation des montant sde sinistre vis-à-vis de l'inflation via une méthode d'indexation des sinistres. L'indexation peut être plus large qu'uniquement le sens monétaire et peut intégrer un changement de tarif ou des changements légaux qui ont un impact sur les montants réservés (augmentation du salaire minimum des soins à domicile par exemple). De cette manière il devient possible d'utiliser des sinistres d'années de survenance différentes pour la tarification.

Cependant, ces modifications réévaluent généralement les montants de sinistres à la hausse (sous l'hypothèse que l'indice utilisé pour la mise en as-if est continuellement croissant – autrement la priorité pourrait baisser). Ce qui entraîne une réévaluation de fait de la priorité des statistiques. Cela fait que pour chaque année de sinistralité de nos données, il y a une troncature à gauche à un seuil de plus en plus élevé que les années de survenance des sinistres sont anciennes. Il devient alors évident que l'on ne peut donc pas garder les sinistres dont le montant est inférieur à la plus grande priorité des statistiques réévaluée, appelée l'*Amin*.

Pour rendre cela plus explicite, prenons l'exemple suivant :

- Sinistres de 2000 à 2020
- Priorité des statistiques : 1 000 000
- Inflation : 2% par an
- Changements législatifs : +20% (2006) et +10% (2015)

L'Amin est donc :

$$1000000 \times 1{,}02^{20} \times 1{,}2 \times 1{,}1 \ = 1961451$$

Nous pouvons d'autant mieux observer ce phénomène à travers une représentation graphique

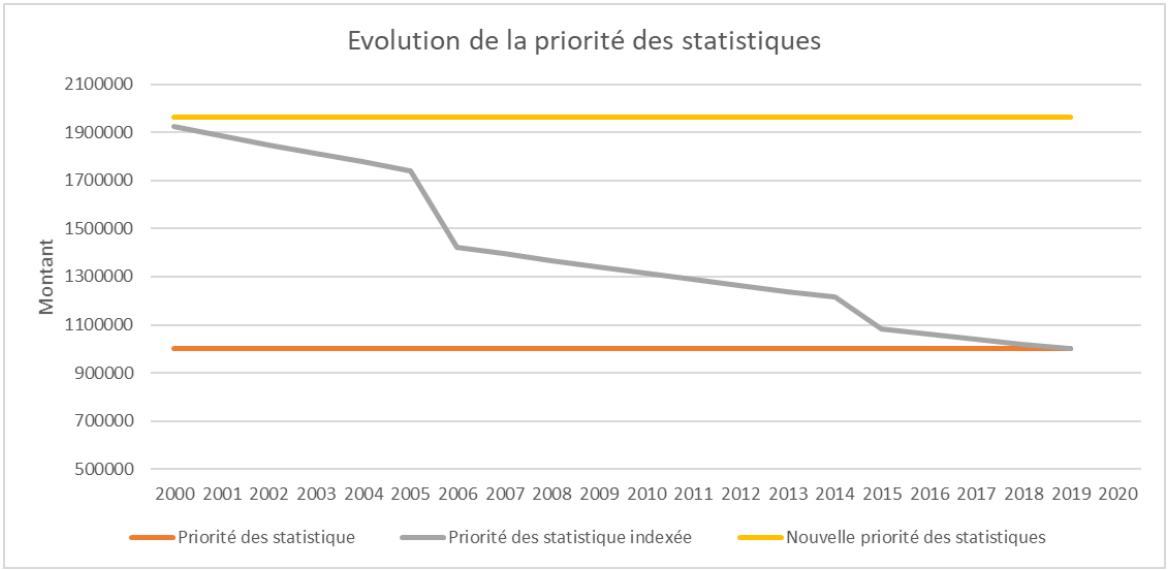


*Figure 3 - Priorité des statistiques indexée par année*

La mise en ASIF des sinistres réduit donc encore le nombre de sinistres utilisables pour la tarification, car ne sont plus utilisables que les sinistres dépassant l'Amin.

## 1.3 Méthode de tarification

La méthode de tarification la plus utilisée par les réassureurs est le modèle du risque collectif : coût moyen × fréquence. Ce modèle repose sur deux hypothèses :

- Les montants des sinistres sont indépendants et identiquement distribués
- Les montants des sinistres sont indépendants de la fréquence de sinistres

Cependant pour modéliser la sévérité et la fréquence du modèle de risque collectif, il faut que les sinistres soient développés en leur valeur ultime. En effet, les paramètres de la sévérité et de la fréquence doivent être estimés à partir de sinistres développés, pour que le coût pris en compte soit

le plus proche de leur coût réel final. C'est en cela que la méthode présentée dans ce mémoire de développement des ultimes a un impact sur la tarification. Ainsi, nous comprenons que l'estimation du montant ultime des sinistres est d'une très grande importance.

Actuellement les sinistres sont développés par la méthode de Chain ladder. Cette méthode est déterministe et permet d'attribuer un montant ultime à chaque sinistre. Une fois les sinistres développés, le réassureur peut procéder à l'estimation des différents paramètres de la sévérité et de la fréquence.

### 1.3.1 Modélisation de la sévérité

L'estimation de la sévérité est découpée en deux parties. La première partie consiste à estimer la queue (*Tail*) de la fonction de répartition. Une fois que les paramètres de la queue de la fonction de répartition sont estimés, le corps (*Body*) de la fonction de répartition peut être estimé. Il est défini comme la partie de la sévérité située avant la queue.

#### *1.3.1.1 Tail*

Classiquement, la queue de la fonction de répartition est modélisée par une loi de Pareto (ou une distribution de Pareto généralisée). Pour ce faire il faut estimer le seuil (noté A) à partir duquel la sévérité des sinistres suit une loi de Pareto et estimer le paramètre de forme (noté α).

Pour rappel la fonction de répartition d'une variable aléatoire X suivant une loi de Pareto est :

$$\mathrm{P}(\mathrm{X} \leq x) = \begin{cases} 1 - \left(\frac{x}{A}\right)^{-\alpha} & si\ x > A \\ 0 & si\ x \leq A \end{cases}$$

**Estimation du seuil**

Pour estimer le paramètre de position A d'une loi de Pareto, la méthode graphique du *Mean Excess Plot* est souvent utilisée.

Définissons la *Mean Excess Function* :

$$e(t) = E[X - t | X > t]$$

Supposons que nous ayons un échantillon : $\{x_1,\dots,x_n\}$ i.i.d provenant d'une même loi, alors la *Mean Excess Function* peut être estimée empiriquement par :

$$\hat{e}_n(t) = \frac{\sum_{j=1}^{n} x_j \, \mathbb{1}_{\{x_j > x\}}}{\sum_{j=1}^{n} \mathbb{1}_{\{x_j > x\}}} - t$$

Ensuite, il faut tracer cette estimation empirique de la Mean Excess Function et choisir le seuil A tel que $\hat{e}_n(t)$ soit approximativement linéaire pour les valeurs x > A.

Voici un exemple avec des données de réassurance, on remarque la linéarité (en rouge) à partir de 8500000. C'est donc notre paramètre de position (ce point est aussi appelé point d'attachement).

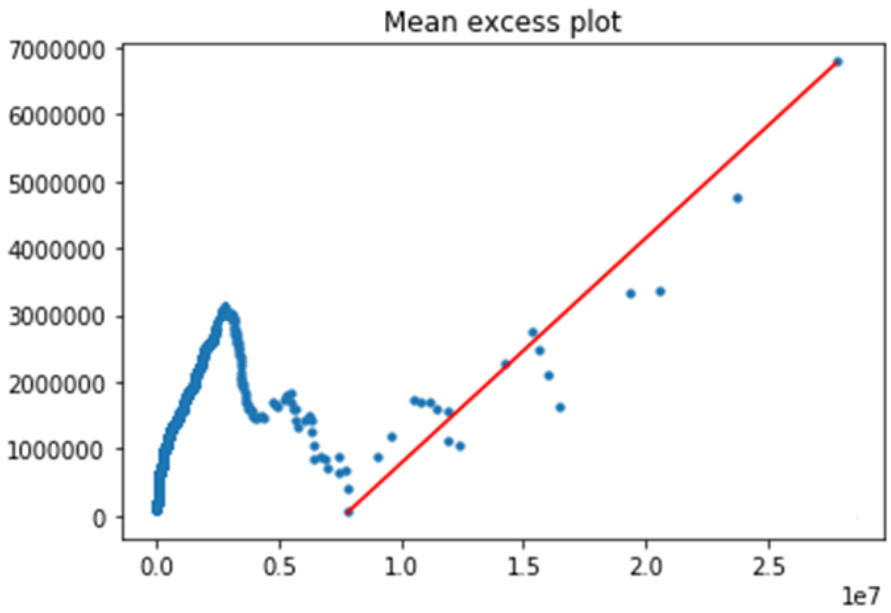


*Figure 4- Exemple Mean Excess Plot*

**Estimation de la forme**

Pour estimer le paramètre de forme α d'une loi de Pareto, le plus souvent la méthode graphique de *l'Alpha Plot* est utilisée.

L'estimateur de maximum de vraisemblance de α au-dessus du seuil A (dont le point d'attachement est $X_{m\text{-}k,m}$) est défini comme :

$$\hat{\alpha}_{k,m} = \frac{k}{\sum_{j=1}^{k} ln\left(\frac{X_{m-j+1,m}}{X_{m-k,m}}\right)}$$

Puisqu'en pratique le point d'attachement n'est pas connu, il faut tracer un α-plot : ($X_{m\text{-}k,m}$ , $\hat{\alpha}_{k,m}$). Le paramètre estimé est le alpha autour duquel l'ensemble des derniers points sont stables.

Voici un exemple avec des données de réassurance, on remarque une stabilité autour de 2. C'est donc notre paramètre de forme.

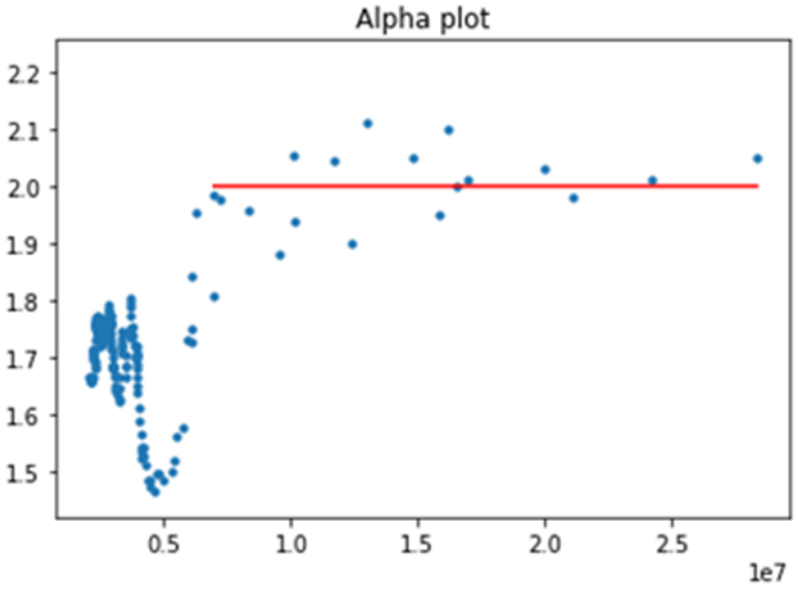


*Figure 5- Exemple Alpha Plot*

#### *1.3.1.2 Body*

Ce qui est appelé le *Body* est défini comme ce qui n'est pas la *Tail*. Ainsi en définissant le seuil de début de Pareto, le seuil de la fin du Body est défini. Cette partie de la sévérité est souvent modélisée par une modélisation empirique.

Il pourrait être intéressant de modéliser le *Body* avec une estimation paramétrique.

### 1.3.2 Modélisation de la fréquence

La fréquence du nombre de sinistres au-dessus d'un seuil est modélisée par une loi de Poisson de paramètre λ. Pour cela, on utilise l'estimateur du maximum de vraisemblance du paramètre λ.

Supposons que nous ayons un échantillon $\{x_1,\dots,x_n\}$ du nombre annuel de sinistres dépassant la priorité pour les n dernières années. Alors l'estimateur du maximum de vraisemblance du paramètre λ est :

$$\hat{\lambda} = \frac{1}{\mathrm{n}}\sum_{i=1}^{n} x_i$$

C'est la moyenne empirique et c'est un estimateur convergent, sans biais, efficace, complet, exhaustif.

## 1.4 Problématique de développement des sinistres pour la tarification

Cette mise en contexte explique bien pourquoi le réassureur doit estimer la valeur ultime des sinistres, dans quel cadre cette estimation doit se faire, et les conséquences de cette estimation pour la tarification. L'estimation de la valeur ultime se doit d'être précise, car elle peut affecter la tarification aussi bien d'un point de vue fréquence que d'un point de vue sévérité.

### 1.4.1 Fréquence

L'impact de l'estimation du coût ultime des sinistres sur la fréquence est évident. C'est d'ailleurs la première raison qui a poussé QBE Re à travailler sur le sujet.

En effet, les réassureurs sont régulièrement confrontés à des effets de seuils, qui sont très prononcés dans les cas où l'estimation du sinistre final est au voisinage de la priorité. La méthode utilisée actuellement n'est pas très satisfaisante dans ce cas-là. On peut très bien imaginer qu'une méthode permettant au réassureur d'estimer la probabilité de dépasser le seuil serait une grande avancée sur cette question et permettrait de réaliser une tarification plus juste par rapport au risque.

### 1.4.2 Sévérité

L'impact sur la sévérité de l'estimation du coût ultime des sinistres n'est pas à négliger. Il est légitime de se questionner sur le manque de flexibilité des résultats actuels.

En effet, calculer un simple coefficient multiplicateur pour chaque année de développement ne permet pas d'envisager d'autres évolutions. Un cas particulier mais assez courant est celui des sinistres pour lesquels le montant final a chuté par rapport au montant actuel. C'est le cas de certains sinistres où un jugement changeant l'attribution de la responsabilité du sinistre a été rendu. Il n'est pas rare pour un réassureur d'avoir des sinistres dont le montant final est ridicule par rapport au montant estimé dans un premier temps.

En particulier, dans le cas où le sinistre était initialement estimé à un montant supérieur au seuil de la loi de Pareto, le fait qu'il ne soit plus dans l'échantillon change fortement la valeur estimée des paramètres. Dans un cas comme celui-ci, le prix technique du contrat s'en retrouve grandement modifié. C'est pour cela qu'il serait intéressant pour le réassureur de pouvoir estimer le montant ultime du sinistre comme un intervalle, avec lequel le réassureur aurait une estimation de la fonction de répartition de la valeur ultime du sinistre.

# Chapitre 2 : Méthode classique de développement de sinistres

L'objectif de ce chapitre est de faire un état des lieux de la méthode de Chain Ladder. C'est la méthode classique de développement de sinistres en réassurance. Cette méthode sera présentée, avec ses hypothèses, son application et ses limites dans le cadre des développements de sinistres.

Ensuite, le modèle de Mack sera présentée. Ce modèle peut être défini comme une version stochastique du modèle de Chain Ladder. L'utilisation de ce modèle peut être vue comme une amélioration directe par rapport à la méthode actuelle. Les hypothèses et les résultats de son application seront étudiés.

Enfin, après avoir étudié le modèle de Chain Ladder et rejeté l'utilisation du modèle de Mack pour obtenir une meilleure estimation du montant ultime des sinistres, les limites structurelles de ces modèles seront mises en exergues. Grâce à ce travail préliminaire, la compréhension des enjeux de la nouvelle méthode de développement des sinistres sera bien plus évidente.

## 2.1 La méthode de Chain Ladder

### 2.1.1 Présentation

Chain Ladder est la méthode classique de développement des sinistres en réassurance. Cependant cette méthode a pour principale limite d'être déterministe. C'est cette limite qui pousse aujourd'hui QBE RE à chercher à changer la méthode de développement.

Chain Ladder est une méthode de référence en provisionnement. Elle est utilisée en assurance non-vie et s'applique à des triangles de sinistres cumulés.

Voici un exemple de triangle de sinistres cumulés où les sinistres se développent jusqu'à 10 ans après leur survenance :

| | Année de développement des sinistres | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Année d'accident | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 2009 | $C_{0,0}$ | $C_{0,1}$ | $C_{0,2}$ | $C_{0,3}$ | $C_{0,4}$ | $C_{0,5}$ | $C_{0,6}$ | $C_{0,7}$ | $C_{0,8}$ | $C_{0,9}$ | $C_{0,10}$ |
| 2010 | $C_{1,0}$ | $C_{1,1}$ | $C_{1,2}$ | $C_{1,3}$ | $C_{1,4}$ | $C_{1,5}$ | $C_{1,6}$ | $C_{1,7}$ | $C_{1,8}$ | $C_{1,9}$ | |
| 2011 | $C_{2,0}$ | $C_{2,1}$ | $C_{2,2}$ | $C_{2,3}$ | $C_{2,4}$ | $C_{2,5}$ | $C_{2,6}$ | $C_{2,7}$ | $C_{2,8}$ | | |
| 2012 | $C_{3,0}$ | $C_{3,1}$ | $C_{3,2}$ | $C_{3,3}$ | $C_{3,4}$ | $C_{3,5}$ | $C_{3,6}$ | $C_{3,7}$ | | | |
| 2013 | $C_{4,0}$ | $C_{4,1}$ | $C_{4,2}$ | $C_{4,3}$ | $C_{4,4}$ | $C_{4,5}$ | $C_{4,6}$ | | | | |
| 2014 | $C_{5,0}$ | $C_{5,1}$ | $C_{5,2}$ | $C_{5,3}$ | $C_{5,4}$ | $C_{5,5}$ | | | | | |
| 2015 | $C_{6,0}$ | $C_{6,1}$ | $C_{6,2}$ | $C_{6,3}$ | $C_{6,4}$ | | | | | | |
| 2016 | $C_{7,0}$ | $C_{7,1}$ | $C_{7,2}$ | $C_{7,3}$ | | | | | | | |
| 2017 | $C_{8,0}$ | $C_{8,1}$ | $C_{8,2}$ | | | | | | | | |
| 2018 | $C_{9,0}$ | $C_{9,1}$ | | | | | | | | | |
| 2019 | $C_{10,0}$ | | | | | | | | | | |

*Figure 6 – Exemple de triangle de sinistres cumulés*

Chain Ladder permet de développer les montants cumulés des sinistres vers leur situation ultime. Ceci, en estimant les coefficients $f_{i,j} = \frac{C_{i,j+1}}{C_{i,j}}$ pour $\mathrm{i} = 1, \dots, \mathrm{n}$ et pour $\mathrm{j} = 1, \dots, \mathrm{n}$

### 2.1.2 Hypothèses et conséquences

La méthode de Chain Ladder repose sur deux hypothèses fortes :

- $\{C_{i,1}, \dots, C_{i,n}\}$ sont indépendants de $\{C_{j,1}, \dots, C_{j,n}\}, si\ i \neq j$

- $\forall j \in 1, \dots, n$ les facteurs de développement $f_{i,j}$ sont indépendants de i

Cette hypothèse entraîne qu'il est possible de considérer des coefficients de passage d'une année à l'autre communs pour chaque année de survenance. Ces coefficients peuvent être estimés par :

$$\widehat{f_j} = \frac{\sum_{i=0}^{n-j+1} C_{i,j+1}}{\sum_{i=0}^{n-j+1} C_{i,j}}, \mathrm{j} = 0, \dots, \mathrm{n}$$

Il devient possible, en appliquant les différents facteurs $f_j$, d'estimer le montant total des sinistres pour chaque année de survenance. Cette estimation se fait de manière très simple et intuitive, tout en donnant un résultat facilement intelligible. De plus ces coefficients peuvent être ajustés facilement par un expert en vue de les rendre meilleurs.

### 2.1.3 Application au développement de sinistres et exemple

Bien que la méthode de Chain Ladder soit pensée pour faire du provisionnement, elle est utilisée pour développer les sinistres individuellement en réassurance. Pour ce faire, les coefficients

$\widehat{f_j}$ , $\forall j = 0, \dots, n$ sont estimés, puis ils sont ensuite appliqués individuellement à chaque sinistre, donnant ainsi la valeur ultime estimée de chaque sinistre.

Voici un exemple sur le triangle de sinistres suivant :

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 2000 | 87852289 | 98056635 | 1.02E+08 | 1.14E+08 | 1.19E+08 | 1.29E+08 | 1.6E+08 | 1.62E+08 | 1.61E+08 | 1.57E+08 |
| 2001 | 1E+08 | 1.17E+08 | 1.4E+08 | 1.39E+08 | 1.48E+08 | 1.81E+08 | 1.88E+08 | 1.88E+08 | 1.87E+08 | 0 |
| 2002 | 94940679 | 1.17E+08 | 1.25E+08 | 1.35E+08 | 1.73E+08 | 1.81E+08 | 1.91E+08 | 1.84E+08 | 0 | 0 |
| 2003 | 1.16E+08 | 1.35E+08 | 1.52E+08 | 2.02E+08 | 2.17E+08 | 2.3E+08 | 2.32E+08 | 0 | 0 | 0 |
| 2004 | 1.34E+08 | 1.61E+08 | 2.1E+08 | 2.19E+08 | 2.14E+08 | 2.06E+08 | 0 | 0 | 0 | 0 |
| 2005 | 1.64E+08 | 2.19E+08 | 2.32E+08 | 2.45E+08 | 2.42E+08 | 0 | 0 | 0 | 0 | 0 |
| 2006 | 3.53E+08 | 3.6E+08 | 3.82E+08 | 3.98E+08 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2007 | 3.45E+08 | 3.49E+08 | 3.28E+08 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2008 | 3.59E+08 | 3.63E+08 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2009 | 3.36E+08 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

*Figure 7 - Triangle de sinistre 1*

On en déduit les facteurs de développement suivants :

| Lambda | $\hat{f}_0$ | $\hat{f}_1$ | $\hat{f}_2$ | $\hat{f}_3$ | $\hat{f}_4$ | $\hat{f}_5$ | $\hat{f}_6$ | $\hat{f}_7$ | $\hat{f}_8$ |
|---|---|---|---|---|---|---|---|---|---|
| | 1.093602 | 1.074062 | 1.081464 | 1.05611 | 1.064725 | 1.069175 | 0.990522 | 0.993226 | 0.972971 |

Et les facteurs de développement cumulés suivants :

| Cumulé | $\hat{f}_0$ | $\hat{f}_1$ | $\hat{f}_2$ | $\hat{f}_3$ | $\hat{f}_4$ | $\hat{f}_5$ | $\hat{f}_6$ | $\hat{f}_7$ | $\hat{f}_8$ |
|---|---|---|---|---|---|---|---|---|---|
| | 1.512725 | 1.38325 | 1.287868 | 1.190856 | 1.127587 | 1.059041 | 0.990522 | 0.983812 | 0.957221 |

On peut alors procéder à l'estimation des valeurs finales des sinistres :

Pour la visualisation, j'ai choisi de sélectionner seulement les sinistres en année de développement 1 pour les besoins des futurs exemples. De cette manière, les exemples suivants seront facilement comparables.

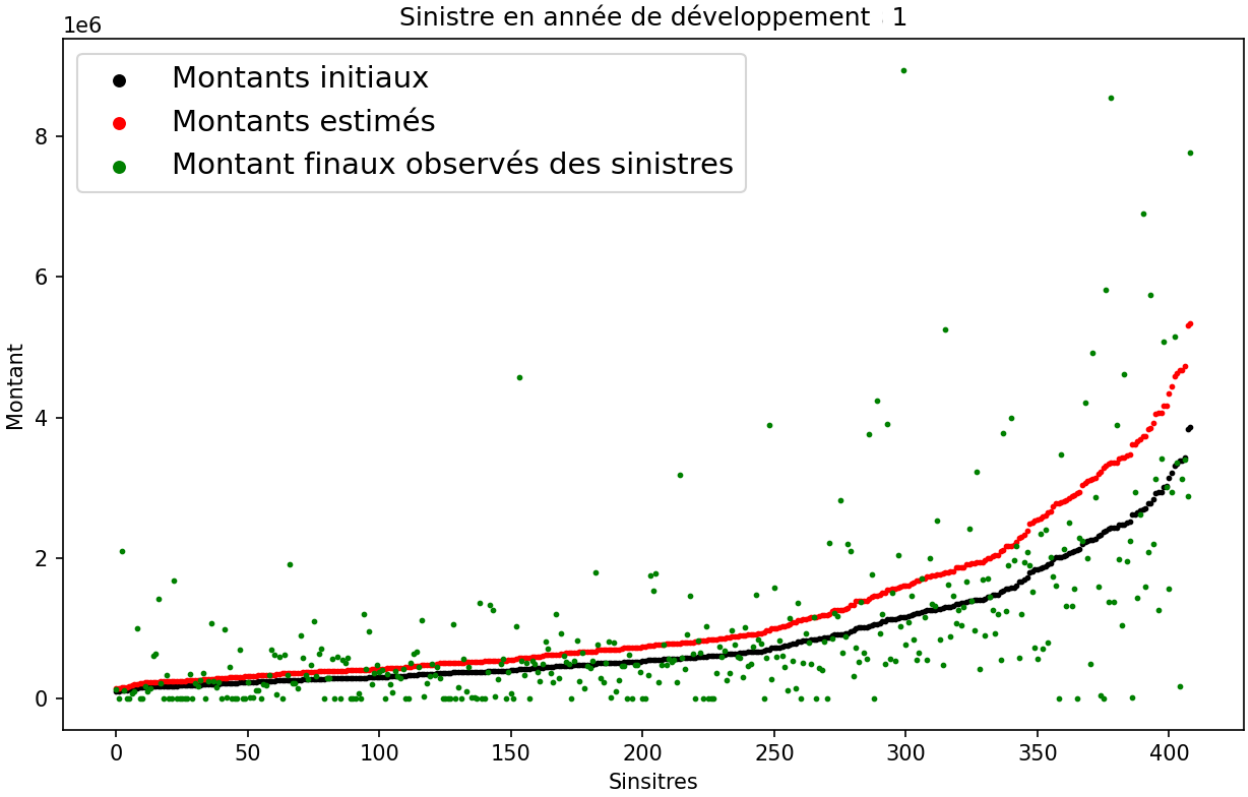


*Figure 8 - Exemple d'estimation du montant ultime des sinistres avec Chain Ladder*

Les résultats obtenus ne sont pas convaincants. L'estimation réalisée ne donne qu'une tendance de ce qu'il se passe réellement. Il semble évident que cette application est un détournement de l'objectif initial de la méthode de Chain Ladder. Les coefficients calculés par la méthode de Chain Ladder ne sont que des moyennes empiriques de l'évolution. Ainsi, les appliquer à chaque sinistre individuellement rend l'estimation obtenue grossière et déterministe, ce qui ne permet pas une bonne compréhension de l'incertitude présente dans cette estimation.

Une méthode stochastique pourrait permettre de résoudre ce problème de développement déterministe.

## 2.2 Modèle de Mack

### 2.2.1 Présentation

Le modèle de Mack (Mack, 1993) [3] peut être présenté comme étant la version stochastique du modèle de Chain Ladder. Il a été développé en 1993 et s'utilise lorsque l'on a disposition un triangle de sinistres cumulés. Il permet de donner une estimation de deux paramètres : $\lambda_j$, et $\sigma_j^2$, tels que :

$$\mathrm{E}[\mathrm{C_{i,j}}|\mathrm{C_{i,0,}} \dots, \mathrm{C_{i,j-1}}] = \lambda_{i,j} C_{i,j-1}\,, \forall i = 1, \dots, n,, \forall j = 1, \dots, n$$

$$\mathrm{V}[\mathrm{C_{i,j}}|\mathrm{C_{i,0,}} \dots, \mathrm{C_{i,j-1}}] = \sigma_{i,j}^2 C_{i,j-1}\,, \forall i = 1, \dots, n, \forall j = 2, \dots, n$$

### 2.2.2 Hypothèses et conséquences

Le modèle de Mack repose sur plusieurs hypothèses fortes, ressemblantes à celles de la méthode de Chain ladder :

- $\{C_{i,1}, \dots, C_{i,n}\}$ sont indépendants de $\{C_{j,1}, \dots, C_{j,n}\}, si\ i \neq j$
- $\forall j \in 1, \dots, n$ Les facteurs de développement $\lambda_{i,j}$ sont indépendants de i
- $\forall j \in 1, \dots, n$ Les facteurs de variance $\sigma_{i,j}^2$ sont indépendants de i

En s'appuyant sur ces hypothèses, Mack propose des estimateurs non biaisés :

$$\widehat{\lambda_j} = \frac{\sum_{i=1}^{n-j+1} C_{i,j}}{\sum_{i=1}^{n-j+1} C_{i,j-1}}, 1 \leq j \leq n$$

$$\hat{\sigma}_j^2 = \frac{1}{n-j} \sum_{i=1}^{n-j+1} C_{i,j-1} \left(\frac{C_{i,j}}{C_{i,j-1}} - \widehat{\lambda_j}\right)^2, 1 \leq j \leq n-1$$

L'estimateur $\widehat{\lambda_j}$ de $\lambda_{\mathrm{i},j}$ est semblable à celui de $\widehat{f_j}$. C'est une moyenne pondérée des facteurs de développement individuels constatés. Mack a montré dans son article de 1993 que cet estimateur est non biaisé et que l'estimation proposée ici est celle à variance minimale dans la classe des estimateurs non biaisés.

L'estimateur $\hat{\sigma}_j^2$ est une moyenne pondérée des résidus où le dénominateur représente le nombre de résidus sur lequel on calcule la moyenne moins un, ce qui est courant lorsqu'on cherche à faire de l'inférence sur une variance. Cela permet d'obtenir un estimateur non biaisé.

L'estimation de $\hat{\sigma}_n^2$ n'est pas prévue par Mack, à cause du fait que seulement une observation est disponible en n. Plusieurs solutions peuvent être trouvées pour résoudre ce problème. Par exemple, il est possible de garder l'estimateur $\hat{\sigma}_{n-1}^2$ pour $\hat{\sigma}_n^2$. Il est aussi possible de fixer le dernier estimateur en prenant la moyenne des k derniers estimateurs.

Le calcul de la réserve estimée peut-être réalisé pour chaque année de survenance i, notée $\widehat{R_i}$, et l'erreur de l'estimation de la réserve peut être mesurée par la moyenne des écarts de prédiction au carré (MSPE – Mean squared prediction error) :

$$MSPE[\widehat{R_i}] = E[(\widehat{R_i} - R_i)^2 | I]$$

Sous les hypothèses énoncées et si $\hat{C}_{i,n-i} = C_{i,n-i}$ l'erreur peut être estimée par :

$$MSPE[\widehat{R_i}] = \hat{C}_{i,n}{}^2 \sum_{j=n-i}^{n-1} \frac{\widehat{\sigma}_j^2}{\widehat{f}_J} \left(\frac{1}{\widehat{C_{i,n}}} + \frac{1}{\sum_{k=1}^{n-j} C_{i,k}}\right)$$

L'erreur standard relative au montant de sinistre en i est donnée par : $\frac{\sqrt{\{MSPE[\widehat{R_i}]\}}}{R_i}$

Grâce à cette erreur relative, il est possible de calculer l'erreur pour chaque sinistre selon son montant. De plus, en ajoutant une hypothèse paramétrique sur la forme de la fonction de répartition des provisions, calculer des quantiles et donner des intervalles de confiance sur nos prédictions.

En me basant sur le mémoire « *Adaptation de la méthode de Merz & Wutrich à des cas non standards* » (Busson, 2012) [2], j'ai choisi de retenir une loi de probabilité classique en assurance : la loi log-normale. Les deux paramètres de cette loi seront estimés en utilisant les estimation $\widehat{R_i}$ et $MSPE[\widehat{R_i}]$.

Ainsi on a $ln(R_i) \hookrightarrow N(\mu, \sigma^2)$, avec $(\mu, \sigma^2)$ solution de :

$$\begin{cases} e^{\mu + \frac{\sigma^2}{2}} = \widehat{R_i} \\ (e^{\sigma^2 - 1}) e^{2\mu + \sigma^2} = (MSPE[\widehat{R_i}])^2 \end{cases}$$

On obtient :

$$\begin{cases} \sigma^2 = ln(1 + \dfrac{MSPE\left[\widehat{R_i}\right]\}}{\widehat{R_i}^{\,2}}) \\ \mu = ln(\widehat{R_i}) - \dfrac{\sigma^2}{2} \end{cases}$$

Sous cette hypothèse de loi, l'intervalle de confiance à 95% de $R_i$ est $[e^{\mu-1,96\sigma}; e^{\mu+1,96\sigma}]$

### 2.2.3 Application au développement de sinistres et exemple

Etant donné que l'objectif de ce mémoire est d'améliorer la méthode de développement des sinistres actuelle, passer de la méthode de Chain Ladder au modèle de Mack est une vraie amélioration. De plus cette amélioration est la plus évidente, car le modèle de Mack peut être considéré comme une amélioration directe de Chain Ladder. Cela aurait pour avantage de rester dans les notions connues en actuariat et donc de rendre le modèle plus intelligible.

En réalisant l'estimation des paramètres $\lambda_j$, et $\sigma_j^2$, $2 \leq j \leq n$ , il devient possible d'estimer l'espérance de l'évolution du coût des sinistres agrégés et la variance du coût des sinistres agrégés. On pourrait imaginer que la fonction de répartition du coût des sinistres agrégés soit log-normale et que de cette manière nous connaîtrions la fonction de répartition des coûts totaux.

Voici un exemple en réutilisant le triangle de sinistres précédent :

Les facteurs de développement cumulés sont identiques :

| Cumulé | $\hat{\lambda}_1$ | $\hat{\lambda}_2$ | $\hat{\lambda}_3$ | $\hat{\lambda}_4$ | $\hat{\lambda}_5$ | $\hat{\lambda}_6$ | $\hat{\lambda}_7$ | $\hat{\lambda}_8$ | $\hat{\lambda}_9$ |
|---|---|---|---|---|---|---|---|---|---|
| | 1.512725 | 1.38325 | 1.287868 | 1.190856 | 1.127587 | 1.059041 | 0.990522 | 0.983812 | 0.957221 |

Les estimations des $\sigma_i$ sont :

| Sigma² | $\hat{\sigma}_1^{\,2}$ | $\hat{\sigma}_2^{\,2}$ | $\hat{\sigma}_3^{\,2}$ | $\hat{\sigma}_4^{\,2}$ | $\hat{\sigma}_5^{\,2}$ | $\hat{\sigma}_6^{\,2}$ | $\hat{\sigma}_7^{\,2}$ | $\hat{\sigma}_8^{\,2}$ | $\hat{\sigma}_9^{\,2}$ |
|---|---|---|---|---|---|---|---|---|---|
| | 2632854 | 2431635 | 1916743 | 1890888 | 1477930 | 1608099.66 | 151071.5 | 62.44149 | 0 |

Puisque les sinistres choisis pour l'exemple sont en année de développement 1, cela signifie qu'il faut calculer la réserve de la neuvième année de développement :

$$\widehat{R_9} = 484726103$$

On peut alors estimer $\sqrt{\{MSPE[\widehat{R_9}]\}}$ =76 598 865

Et donc l'erreur relative au montant de sinistre est de $\frac{\sqrt{\{MSPE[\widehat{R_i}]\}}}{R_i}$ = 15,8%

Ainsi l'estimation des paramètres μ et σ est :

$$\begin{cases} \sigma^2 = ln(1 + 0{,}1582^2) = 0.02465749 \\ \mu = ln(S_i) - \frac{\sigma^2}{2}, \text{ avec } S_i \text{ le montant du i}^{\text{ème}} \text{ sinistre à estimer} \end{cases}$$

On peut alors procéder à l'estimation de la valeur finale des sinistres ainsi que de l'intervalle de confiance à 95% :

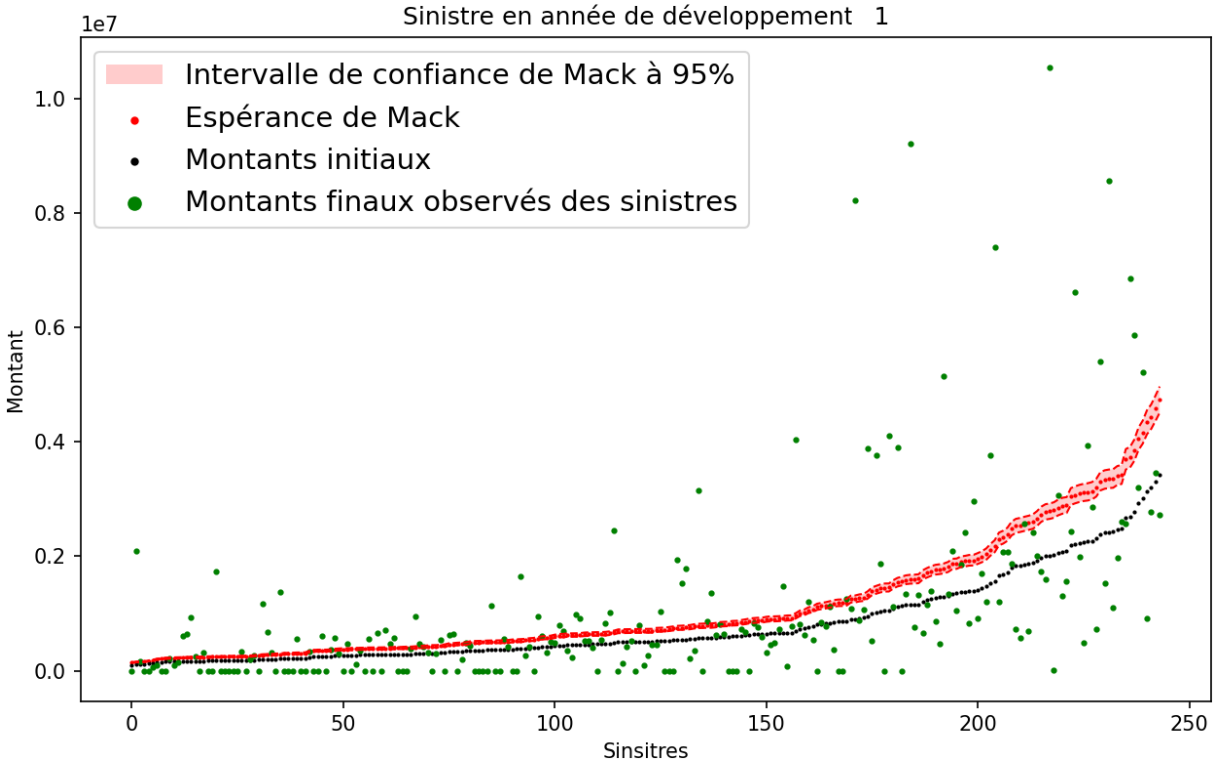


*Figure 9 - Exemple d'estimation du montant ultime des sinistres avec le modèle de Mack*

Comme avec la méthode de Chain Ladder, il semble évident que cette application est un détournement de l'objectif initial du modèle de Mack. C'est une méthode de provisionnement et elle est pensée pour être mise en place sur des sinistres agrégés. C'est pour cela que les intervalles obtenus ne sont pas satisfaisants car trop petits. 95% des sinistres finaux sont censés être compris dans cette intervalle, alors que ce n'est pas du tout le cas. Le modèle de Mack n'est pas une solution satisfaisante pour mieux estimer le montant final des sinistres.

## 2.3 Objectif d'une nouvelle méthode de développement de sinistres

Les méthodes actuarielles classiques utilisées à QBE Re ne sont pas adaptées à l'estimation du montant ultime des sinistres. Plusieurs pistes d'amélioration peuvent être imaginées pour rendre les estimations meilleures.

L'objectif étant d'obtenir une fonction de répartition des montant finaux possibles, il est évident que la première étape est d'enrichir l'information possédée sur chaque sinistre en vue d'améliorer la robustesse de l'estimation du montant final. Pour la suite, des nouvelles variables qui permettront de décrire le comportement du sinistre seront créées. Par exemple, une variable correspondra à la somme des variations du montant du sinistre des trois dernières années.

Ensuite, j'utiliserai ces différentes variables permettant de mesurer des aspects du comportement du sinistre pour créer des groupes de comportement de sinistres.

# Chapitre 3 : Introduction d'une notion de comportement de sinistres

En vue d'obtenir une nouvelle manière plus précise de développer les sinistres, des nouvelles variables ont été créées. Ces nouvelles variables sont basées sur l'historique des estimations des montants des sinistres et sur l'historique des paiements effectués par le réassureur à la cédante.

Les données utilisées dans cette étude sont l'entièreté des sinistres auto et responsabilité civile générale en France et en Belgique (les sinistres accident du travail seront aussi étudiés pour la Belgique) de QBE Re entre 2000 et 2020.

Ces variables seront le fondement de l'idée de comportement de sinistre. La notion de comportement d'un sinistre peut être définie comme : la manière dont le montant estimé d'un sinistre évolue dans le temps. Grace à cette très grande base de données, des groupes de comportement de sinistres seront définis. Ils permettront de classer les futurs sinistres d'une manière plus pertinente qu'un classement basé uniquement sur la cédante déclarant le sinistre.

## 3.1 Comportement de sinistres

### 3.1.1 Partitionnement des données

Dans un premier temps, les données seront séparées selon leur origine géographique : les sinistres France et Belgique seront traités séparément. Ceci car ils ne se comportent pas de la même manière pour des raisons législatives et culturelles. Un sinistre auto français peut avoir un coût extrêmement élevé, là où les sinistres auto belges ont des coût plus modérés.

Dans un second temps, les sinistres auto, responsabilité civile générale, et accident du travail seront traités séparément. Ces sinistres sont de natures différentes, par conséquent ils ne se comportent pas de la même manière.

Dans un troisième temps, les observations seront regroupées selon leur années de développement. En particulier, si un sinistre en $k^{ième}$ année de développement est dans la base de données, alors nous disposons des informations pour les années de développement de $0$ à $k$. Donc ce sinistre sera utilisé comme observation pour les bases de données des $k$ premières années de développement. En revanche ce sinistre aura dans la $i^{ième}$ base de données la valeur qui lui était attribuée lors de sa $i^{ième}$ année de développement.

Dans la suite du mémoire, les exemples et les visualisations se feront sur les sinistre auto français. Ce sont les sinistres les plus importants pour QBE Re et ceux dont nous disposons de la plus grande base de données.

### 3.1.2 Définition des nouvelles variables - Métriques

En vue de pouvoir créer des catégories de sinistres, de nouvelles variables ont été créées. L'objectif de ces variables est de réussir à catégoriser les comportements des sinistres. Ces variables seront nommées « Métriques », car elles ont pour objectif de mesurer certains aspects du comportement des sinistres. Je vais détailler chacune de ces métriques.

Pour certaines variables, la notion de « montant significatif » est utilisée. Pour mettre en place la méthode, le montant de 100 000€ a été défini comme significatif. Cela veut dire que le sinistre sera pris en compte seulement lorsqu'il aura au moins une fois dépassé la valeur de 100 000€, de manière à éviter les problèmes posés par les sinistres déclarés tardivement et pour lesquels le coût estimé du sinistre pour les premières années de développement est très faible.

| Nom de la variable | Calcul | Description |
|---|---|---|
| $Incurred_n$ | Variable présente dans les données de renouvellement | Dernier montant total du sinistre estimé par la cédante pour l'année n actuelle. |
| $Incurred_0$ | Variable présente dans les données de renouvellement | Montant de la première évaluation significative du sinistre par la cédante. |
| Incurred evolution | $\frac{Incurred_n}{Incurred_{n-1}}$ | Evolution du montant estimé du sinistre par rapport à l'estimation de l'année précédente. |
| % Min | $\frac{\min\limits_{k=0\dots n}(Incurred_k)}{Incurred_n}$ | Cette métrique permet de comparer le montant minimum significatif estimé depuis l'année de survenance du sinistre par rapport à l'estimation actuelle du montant du sinistre. |
| % Max | $\frac{Incurred_n}{\max\limits_{k=0\dots n}(Incurred_k)}$ | Cette métrique permet de comparer le montant actuel estimé par rapport à l'estimation maximum du montant du sinistre depuis son année de survenance. |
| Incurred moyen | $\frac{\sum_{k=0}^{n} \mathrm{Incurred_k}}{n}$ | Cette métrique représentante moyenne des estimations du montant du sinistre. |
| Moyenne des trois derniers Incurred | $\frac{\sum_{k=n-2}^{n} \mathrm{Incurred_k}}{3}$ | Moyenne des trois dernières estimations du montant du sinistre. |

| | | |
|---|---|---|
| Somme des incurred relatif | $\frac{\sum_{k=0}^{n} \text{Incurred}_k}{Incurred_n}$ | Somme des situations passées rapport à la situation actuelle |
| Grande variation | $\begin{cases} 1 \; si \; Incurred \; evolution > 1{,}5 \\ 1 \; si \; Incurred \; evolution < 0{,}5 \\ 0 \; sinon \end{cases}$ | Métriques valant 1 si le sinistre a beaucoup évolué depuis la dernière situation (±50%) |
| Nombre de très grandes variations | $\sum_{k=0}^{n} \text{Grande variation}_k$ | Nombre total d'années où le sinistre a connu une grande variation |
| Nombre de changement de sens de la variation | | Nombre total d'année ou le sinistre a changé de sens de variation : passage de la hausse à la baisse (ou inversement) entre deux années de développement |
| Nombre d'années ou le sinistre était évalué à un montant insignifiant | $\sum_{k=0}^{n} \mathbb{1}_{\text{Incurred}_k < 100000}$ | Cette métrique donne une indication sur le nombre d'année où le sinistre était évalué à un montant non significatif. |
| Paid | Variable présente dans les données de renouvellement | Montant total du sinistre payé en année n. |
| Proportion Paid | $\frac{Paid_n}{Incurred_n}$ | Proportion du sinistre payée par rapport à l'estimation du montant |
| Paid evolution | $\frac{Paid_n}{Paid_{n-1}}$ | Evolution du montant de Paid entre l'année précédente et l'année actuelle |

Ces métriques sont toutes dérivées des triangles de sinistres et de paiements fournis par la cédante lors des renouvellements. Elles permettent de capter les différents aspects du comportement des sinistres. Toutes ces variables ont ensuite été centrées et réduites pour faciliter l'analyse de données.

Un certain nombre de métriques avait été imaginées, mais seulement les métriques les plus pertinentes au sens d'une Analyse en composantes principales (ACP) ont été conservées. Parmi les métriques non conservées, il y avait la moyenne des 5 derniers incurred et des 10 derniers incurred qui étaient très redondante avec celle des 3 derniers incurred. On peut aussi noter que des métriques par rapport au nombre de variation à la hausse et à la baisse n'ont pas été conservées. Elles n'apportaient pas plus d'information que le Nombre de très grandes variations.

Cependant, créer des groupes de comportements de sinistre à partir de ces métriques est un processus complexe. C'est pour cela que des techniques de classification non supervisée, aussi appelées *Clustering*, ont été utilisées.

## 3.2 Classification non supervisée

En vue de créer des groupes de comportement de sinistre, plusieurs algorithmes de classification non supervisée seront décrits et testés.

### 3.2.1 Introduction à la classification non supervisée

La classification supervisée, aussi appelée Clustering, a pour objectif d'attribuer à chaque observation $x_i$ une étiquette $y_i \in \mathrm{N}$. Cette étiquette matérialise l'appartenance à une classe. La classification est dite non supervisée, car l'algorithme de classification n'utilise pas un jeu de données déjà classées au préalable pour apprendre à prédire. Dans le cas de la classification non supervisée, l'objectif de l'algorithme est de créer des classes qui à la fois maximisent l'homogénéité intra-classes et l'hétérogénéité inter-classes.

La problématique de créer des groupes de sinistres dont le comportement serait similaire peut donc trouver sa solution par un algorithme de clustering.

### 3.2.2 Algorithmes de clustering

Il existe un grand nombre d'algorithme de clustering, chacun avec ses avantages et ces inconvénients. Les différentes techniques de clustering utilisées lors de la réalisation de ce mémoire seront détaillées.

#### *3.2.2.1 Algorithmes du k-Means*

L'algorithme du K-Means regroupe les données en essayant de séparer les échantillons en $k$ groupes (ou clusters), de manière à minimiser l'inertie intra-groupe. Cet algorithme exige que le nombre de groupes soit préalablement spécifié. Il s'adapte bien aux grands échantillons et a été utilisé dans de nombreux domaines différents. L'algorithme du K-Means est un algorithme basé sur des centroïdes. Cela signifie que les clusters sont entièrement définis par leur centre et les observations sont classées dans le cluster dont le centre est le plus proche.

Il divise un ensemble de $N$ échantillons $x_i$ en $k$ groupes disjoints $C$, chacun étant décrit par la moyenne $\mu_j$ des échantillons du cluster. Les $\mu_j$ sont les « centroïdes » des clusters. K-Means vise à choisir des centroïdes qui minimisent l'inertie, ou le critère de la somme des carrés dans un cluster :

$$\sum_{i=0}^{n} \min_{\mu_j \in C} \left( ||x_i - \mu_j||^2 \right)$$

L'inertie peut être vue comme une mesure de l'homogénéité des clusters. Elle souffre cependant d'un grand inconvénient : l'inertie suppose que les clusters sont convexes, ce qui n'est pas toujours le cas. Il réagit mal aux clusters allongés, aux formes irrégulières et aux groupes de variance différente.

Or, certaines des métriques (Nombre de très grandes variations, grande variation, Nombre de changement de signe de la variation) prennent leur valeur dans l'ensemble des entiers naturels. D'autres ont des distributions très différentes. Cela pose problème pour l'utilisation d'un algorithme de K-Means sur les métriques créées.

De plus cet algorithme demande à l'utilisateur de choisir le nombre de clusters, ce qui n'est pas aisé lorsque la dimension des observations est grande.

Voici des exemples (scikit-learn.org, 2020) [7] des problèmes posés par cet algorithme :

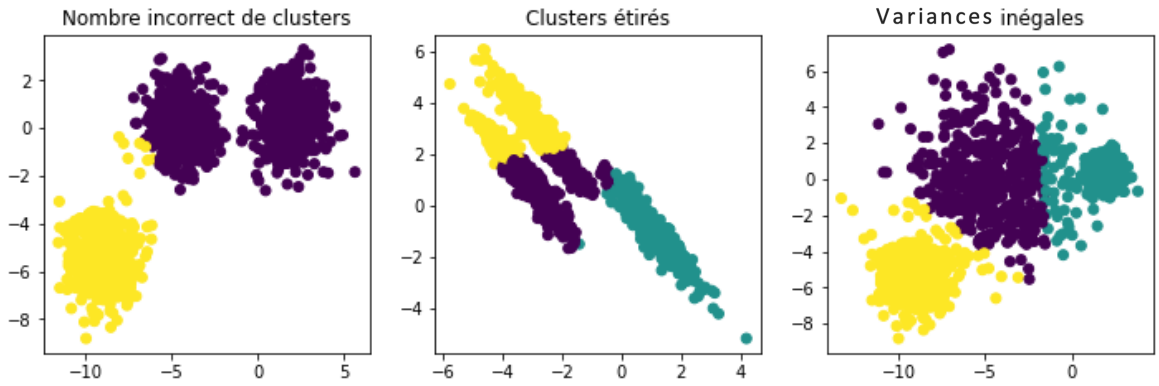


*Figure 10 - Limites de l'algorithme K-Means*

C'est pour cela que la méthode ne donnait pas de bons résultats lorsque l'algorithme K-means était utilisé. D'autres algorithmes ont été testés par la suite.

### *3.2.2.2 Algorithmes de densité (DBSCAN)*

Mon attention s'est ensuite portée sur les algorithmes basés sur la densité, notamment sur l'algorithme DBSCAN (Martin Ester, 1996) [5]. Ils fonctionnent en détectant les zones où les observations sont concentrées et sont séparées par des zones de vides (ou quasi-vide). Les points qui ne font pas partie d'un groupe sont qualifiés de bruit.

Ces algorithmes ont tous un paramètre commun : le nombre $k$ minimum d'observations par clusters. Cela signifie qu'il faut minimum $k$ observations assez denses pour former un cluster. Dans le cas où le nombre d'observations assez denses entre elles est inférieur à $k$ ces observations seront considérées comme du bruit.

Pour l'algorithme DBSCAN, un autre paramètre doit être fourni par l'utilisateur : la distance de recherche. Ce paramètre désigne la distance minimale qui doit séparer deux observations pour les considérer dans le même cluster.

Les clusters trouvés par DBSCAN peuvent avoir n'importe quelle forme, par opposition à ceux trouvés par K-Means qui suppose que les clusters sont de forme convexe.

Cependant après expérimentations de l'algorithme, je me suis rendu compte qu'il n'était pas adapté aux observations. Le premier problème vient du fait que les clusters ont des densités différentes. Certains clusters devraient être très dense, alors que d'autre beaucoup moins. De plus, le fait de traiter certain cas comme du bruit empêche de classifier toutes les observations, ce qui n'est pas du tout satisfaisant en vue de la suite des travaux à réaliser.

#### *3.2.2.3 Classification ascendante hiérarchique*

La classification ascendante hiérarchique est une famille d'algorithmes de classification non supervisée qui construit des clusters en regroupant successivement des observations ou des clusters d'observations. Les résultats de cette méthode de clustering sont représentés sous forme d'un arbre (appelé dendrogramme). La racine de l'arbre est l'unique grappe qui rassemble tous les échantillons (souvent représentée en haut de l'arbre), les feuilles étant les clusters composés d'une seule observation. La classification ascendante hiérarchique est composée de deux paramètres :

**Le paramètre de distance :** Une distance d sur l'espace de dimension $n$ permet de mesurer la proximité entre deux observations. Voici quelques exemples de distance classique :

| **Distance** | $\boldsymbol{d(x, y)}$ |
|---|---|
| Distance euclidienne | $\left\|\|x - y\|\right\| = \sqrt{\sum_i (x_i - y_i)^2}$ |
| Distance de Manhattan | $\|\|x - y\|\| = \sum_i \|x_i - y_i\|$ |
| Norme de de la convergence uniforme | $\|\|x - y\|\| = \max_i \|x_i - y_i\|$ |

**Le critère de lien :** C'est le critère sur lequel se basera le regroupement des observations et des clusters. Voici quelques exemples de critère de lien simple et classique :

| Critère de lien | $\Delta(C_1, C_2)$ |
|---|---|
| Lien simple | $min\{d(x,y) \mid x \in C_1, y \in C_2\}$ |
| Lien complet | $max\{d(x,y) \mid x \in C_1, y \in C_2\}$ |
| Lien Moyen | $\frac{1}{\#C_1 \#C_2} \sum_{x \in C_1} \sum_{X \in C_2} d(x,y)$ |

Il existe aussi d'autre critère de lien souvent utilisé. Notamment le critère de Ward (Ward, 1963) [8] qui s'appuie sur la forte connexion entre les notions de distances et de variance. L'objectif de l'algorithme est d'agréger les individus ou les classes afin de faire varier le moins possible l'inertie intra-classe à chaque étape. Ceci revient à rendre minimale la perte d'inertie inter-classe.

Il est plus facile de visualiser le fonctionnement de l'algorithme par un exemple :

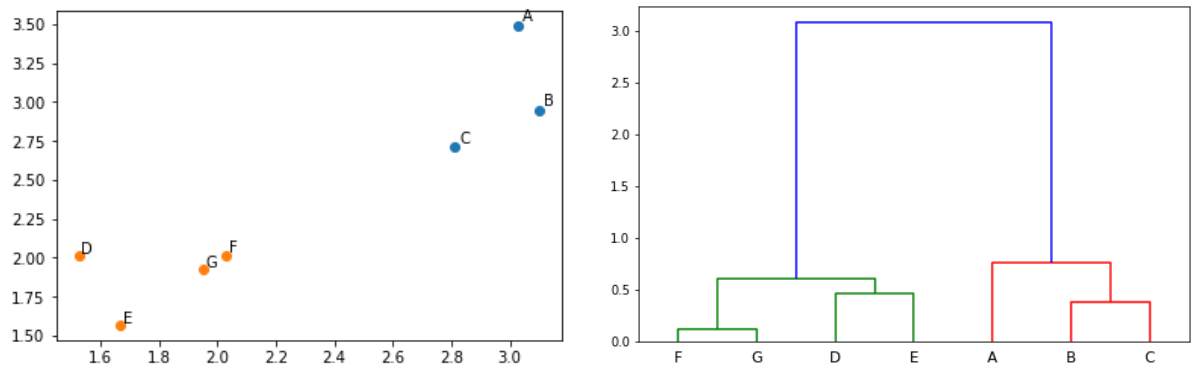


*Figure 11 - Fonctionnement de la classification ascendante hiérarchique*

Voici des données en deux dimensions nommées de A à G. A côté le dendrogramme associé à ces données. Plus la hauteur des nœuds reliant deux points (ou groupes de points) est élevée, plus ceux-ci étaient éloignés. On remarque facilement que l'on peut séparer nos données en deux groupes.

## 3.3 Clustering des comportements similaires

Après avoir testé les algorithmes de clustering présentés précédemment, celui qui est le plus adapté aux données est l'algorithme de classification ascendante hiérarchique. En effet cet algorithme peut créer des groupes non convexes, tout en n'excluant aucune observation.

Cependant, la classification ascendante hiérarchique dépend de deux paramètres qu'il conviendra de choisir intelligemment, car ceux-ci définissent entièrement les clusters créés. Pour être certain de faire le bon choix de paramètre, une analyse des résultats obtenus est indispensable.

Enfin, la classification ascendante hiérarchique n'est pas un algorithme qui permet de traiter de nouvelles observations (contrairement au K-Means). Pour ce faire, une méthode permettant de rattacher de nouvelles observations à des clusters existants sera présentée.

### 3.3.1 Choix des paramètres

Le choix des paramètres est extrêmement important dans une classification ascendante hiérarchique. Chacun des paramètres possibles a été étudié avec soin. Mais une méthode de regroupement m'a semblée plus pertinente que les autres : la méthode de Ward. Cette méthode vise à maximiser l'inertie inter-classe et donc d'avoir les classes les plus différentes les unes des autres.

Cependant, pour créer les clusters, un algorithme de classification ascendante hiérarchique a besoin de connaître la distance maximale autorisée entre deux observations d'un même cluster, ce que l'on appelle communément le seuil auquel on « coupe le dendrogramme ». Pour ce faire, il existe plusieurs méthodes permettant d'automatiser le choix. Cependant ces méthodes sont critiquables, car elles automatisent une décision qui est avant tout une question d'interprétation. C'est pour cette raison que j'ai choisi de ne pas faire appel à ces méthodes et de choisir la distance maximale entre les observations de manière manuelle.

### 3.3.2 Présentation des résultats et analyse

Pour chaque catégorie, les dendrogrammes ont été créés et je les ai analysés pour déterminer le seuil de coupe du dendrogramme. Les seuils choisis ont toujours été autour de 25% de la hauteur maximale, mais adaptés pour chaque cluster.

Voici plusieurs exemples dendrogrammes avec, en noir, la coupe qui a été réalisée.

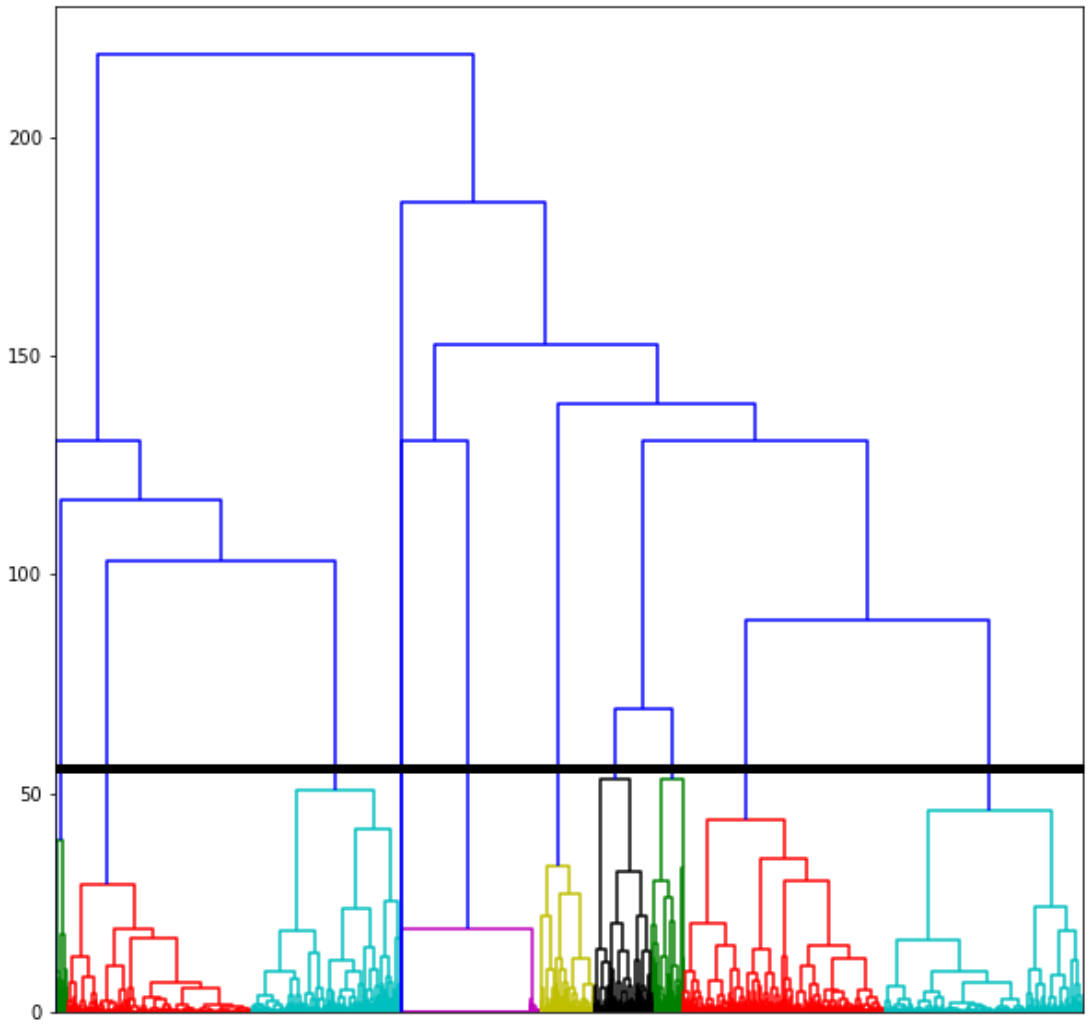


*Figure 12 - Dendrogramme des sinistres auto français en année de développement 0*

Dans le cas des sinistres auto français en année de développement 0, j'ai choisi de couper le dendrogramme à 12 clusters (environ 40 de hauteur). Cela ne fait pas beaucoup de groupe par rapport aux 8620 observations parce que lorsque que les sinistres sont en année de développement 0, la plupart des métriques prennent la même valeur car il n'y a pas d'historique pour ces sinistres.

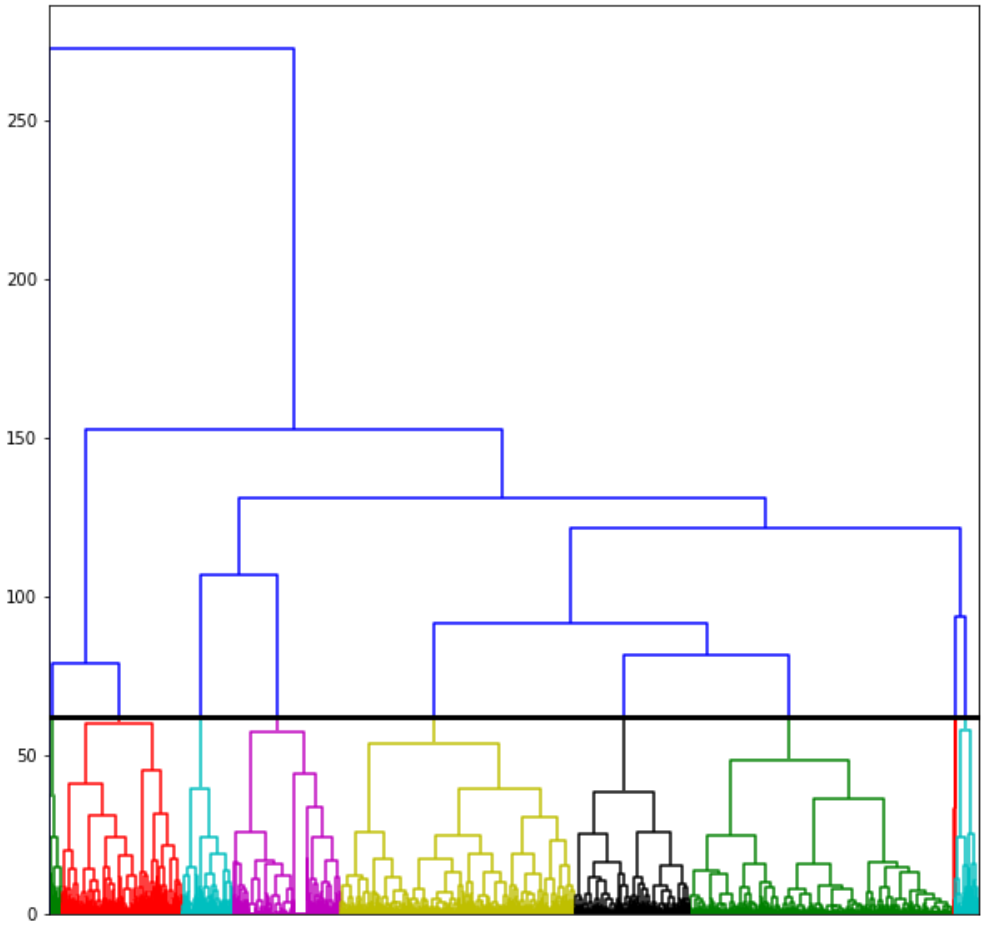


*Figure 13 - Dendrogramme des sinistres auto français en année de développement 6*

Dans le cas des sinistres auto français en année de développement 6, j'ai choisi de couper le dendrogramme à 10 clusters (environ 67 de hauteur). Ce dendrogramme montre bien l'intérêt des métriques et du clustering. En choisissant la bonne méthode de clustering on peut voir clairement que des groupes se dégagent.

Ces groupes peuvent ensuite être analysés à travers un *pairplot* qui est juste l'ensemble des nuages de points pour chaque duo de variable possible. Certain duo de variables sont très intéressants et montre que le clustering est efficace.

Voici un exemple de pairplot (sur une sélection limitée de variables pour plus de clarté) sur les clusters des sinistres auto français en année de développement 6 :

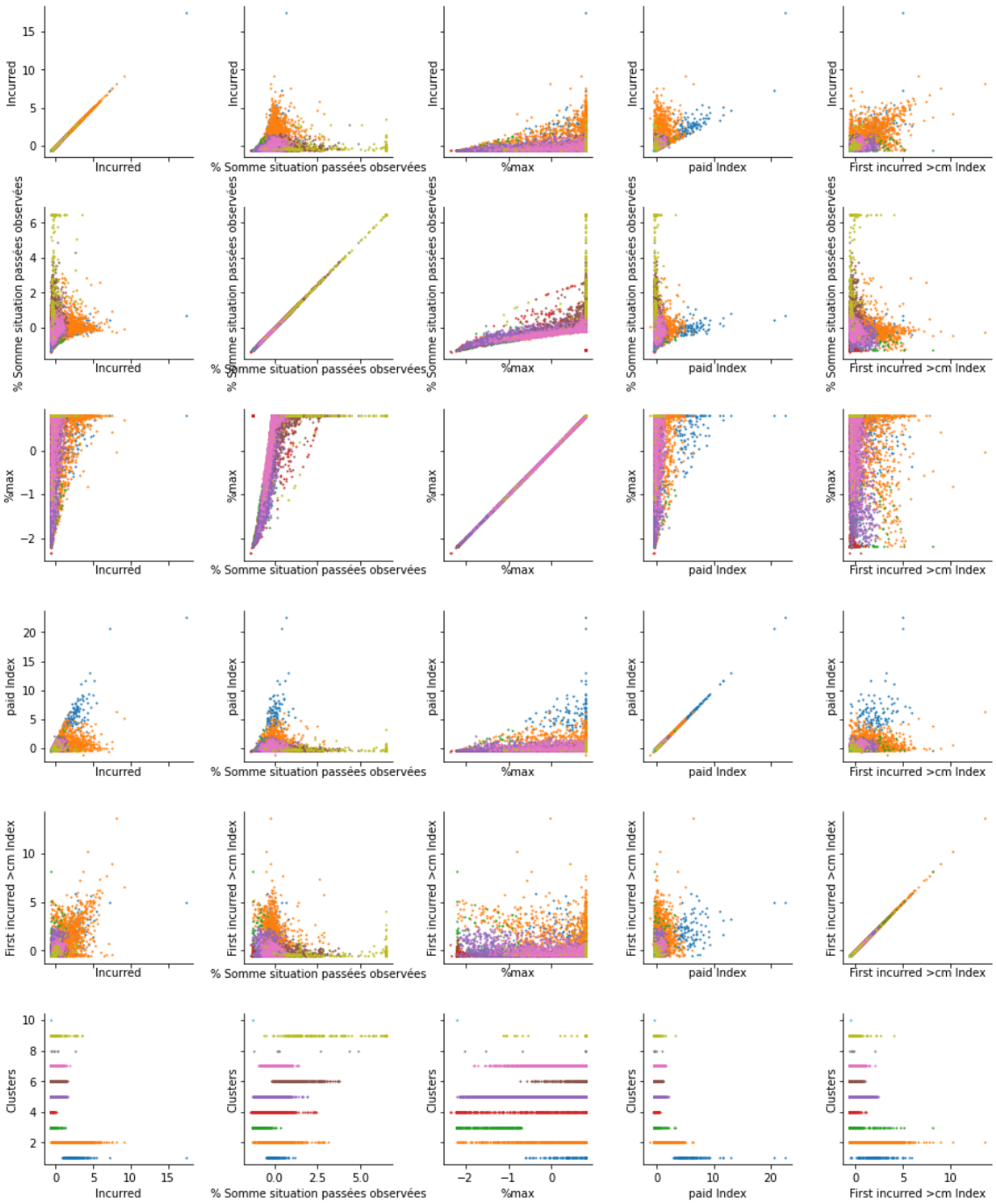


*Figure 14 - Selection de Pairplot pour les sinistre auto français en année de développement 6*

Tout d'abord il est intéressant de noter la repartions des observations au sein des clusters : 8 des 10 clusters ont des effectifs supérieur 100 mais les clusters numéro 8 et 10 comportent respectivement 7 et 1 observations. Cela pourrait poser des problèmes dans la suite et c'est pour cela que lorsque les résultats du clustering seront utilisés je veillerai à bien à ce que les petits clusters (en quantité d'observations) ne perturbent pas les estimations du montant ultime des sinistres.

Ensuite, on remarque que les clusters ont une cohérence pour chaque duo de de variables. Bien sûr les limites de chacun des clusters ne sont pas parfaitement visibles et certaines semble même

se chevauchées, mais cela est dû au fait que nos observations sont en dimensions 15 et que seulement 5 variables sont représentées ici.

Il est aussi intéressant de regarder la dernière ligne du Pairplot qui montre le rapport entre chaque cluster et chaque variable. Ce qui nous permet de comprendre les caractéristiques des sinistres présent dans chaque cluster.

### 3.3.3 Gestion des nouveaux sinistres

Un des problèmes de la classification ascendante hiérarchique est qu'il est impossible d'attribuer à une nouvelle observation une étiquette. En effet les classes créer par la classification ascendante hiérarchique sont figées.

Or, l'objectif du clustering était d'utiliser les données des sinistres possédées par QBE Re pour créer des classes de comportement de sinistre en vue de classer de nouveaux sinistres en fonction leur propre comportement.

Une manière d'attribuer une classe aux nouveaux sinistres a été trouvée : Utiliser un algorithme de prédiction. L'algorithme de prédiction le plus adapté à ce besoin est celui des $k$ plus proches voisins ($k$-*nearest neighbors*). C'est un algorithme de classification supervisée dont le principe est très simple : on associe à une nouvelle observation $x_i$ la classe la plus présente parmi les k plus proches observations (au sens d'une distance euclidienne). Cette méthode semble être tout à fait adaptée pour prolonger le clustering.

Après avoir créé un modèle des $k$ plus proches voisins pour $k \in \{1, \dots, 20\}$, le meilleur $k$ au sens de la précisions est $k = 1$.

Voici des exemples (pour les sinistres auto en année de développement 0, 4, et 8) de la précision du modèle selon la valeur de $k$ sur un set de test composé de 10% de l'échantillon :

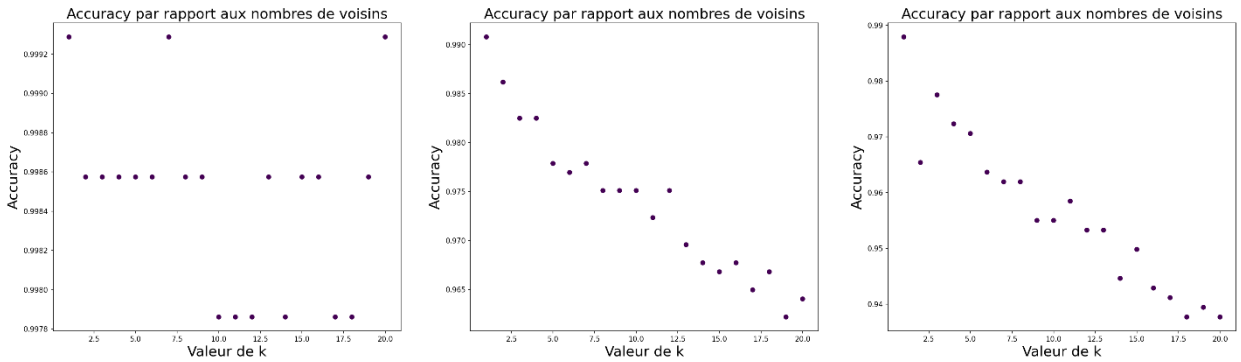


*Figure 15 - Précisions du classifieur selon la valeur de k*

Malheureusement, cette méthode n'est pas parfaite. Tout d'abord, on remarque que la précision n'est pas de 100%, ce n'est donc pas un prolongement parfait pour la classification. Il serait intéressant de pouvoir avoir la probabilité d'être dans chacun de des clusters, pour pouvoir comprendre de manière plus fine la prédiction. Or, ce n'est pas possible avec l'algorithme des $k$ plus proches voisins.

Une deuxième limite d'utiliser l'algorithme des $k$ plus proches voisins est que chacune des nouvelles observations sera classée individuellement. Or, cela peut ne pas être rigoureux, par exemple si un nouveau cluster se développe, il ne pourra pas être compris par l'algorithme comme un nouveau cluster. Enfin, si des données « prolongent » un cluster, celle-ci pourrait ne pas être traitées comme telles. Voici un exemple ou les nouvelles données prolongent un cluster actuel mais si la classification de ces données se fait une à une, alors elles ne seront pas toutes rattachées au bon cluster.

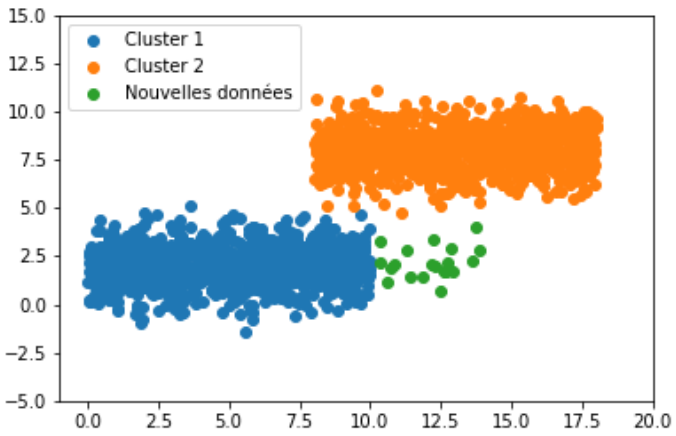


*Figure 16 - Limite du traitement des nouvelles données*

## 3.4 Conclusion

La création des métriques pour décrire le comportement des sinistres remplie ses objectifs. Des clusters cohérents de comportement de sinistres ont pu être créés. Ceci en utilisant la classification ascendante hiérarchique et la distance de Ward.

Cependant, cette manière de créer les clusters ne permet pas, en théorie, de classer de nouvelles observations. Pour cela un algorithme des $k$ plus proches voisins est utilisé et permet, à quelques limites près, de rattacher à chaque nouvelle observations un cluster.

Maintenant que chaque observation a permis d'enrichir notre classification et que chaque nouveau sinistre peut être rattaché à un cluster, nous pouvons utiliser ces résultats pour définir un modèle de prédiction de la valeur ultime des sinistres.

# Chapitre 4 : Application de la nouvelle méthode

L'objectif de ce chapitre 4 est d'utiliser les clusters de comportement des sinistres historiques en vue de construire un modèle de prédiction de la fonction de répartition des valeurs ultimes des sinistres.

Pour ce faire, une description générale de la nouvelle méthode présentée dans ce mémoire sera proposée. La réflexion générale que j'ai menée et les choix que j'ai fait y seront décrit. Cela permettra de comprendre la philosophie générale du modèle, tout en donnant des clefs pour y apporter des améliorations dans le futur.

Ensuite, je décrierai les quatre grandes étapes de la création du modèle. Les deux premières étapes seront la base de la construction des résultats des étapes 3 et 4. La première étape consistera à calculer et explorer les transitions entre les différents clusters à chaque année de développement. La seconde est de calculer les probabilités de clôture dans chaque cluster et dans chaque année de développement.

Les deux dernières étapes sont celles de la construction du modèle final. Premièrement, j'expliquerai comment construire un modèle permettant de donner à chaque sinistre ses valeurs possibles en une année de développement future. Puis j'utiliserai ces résultats pour créer un modèle général, donnant pour chaque sinistre la fonction de répartition de son montant ultime.

## 4.1 Description de la nouvelle méthode

La méthode proposée dans ce mémoire s'appuie sur les groupes de comportement similaire créés précédemment. Ainsi, pour chaque pays, ligne de business, et chaque année de développement entre 10 et 20 groupes ont été créés à partir des sinistres disponibles dans la base de données de QBE Re depuis 2000.

Grâce à ces clusters et à la méthode choisie de gestion des nouveaux sinistres, chaque nouveau sinistre peut se voir attribuer une étiquette correspondant au groupe auquel il appartient. Ensuite, ce sinistre sera entièrement assimilé comme appartenant à ce groupe.

Dans un premier temps, les probabilités de passages d'un cluster d'une année de développement vers un cluster à une année de développement future seront calculées. En faisant cette opération pour chaque cluster, il sera possible de calculer la probabilité de chaque chemin de développement. Ainsi, pour chaque nouveau sinistre, les chemins de développement possibles seront connus et leur probabilité d'occurrence calculée.

Ensuite, les probabilités de chacun des clusters d'observer un sinistre qui se clôture en son sein seront calculées. Ainsi pour chaque cluster de chaque année de développement, la probabilité qu'un sinistre se clôture sera calculée. Un sinistre clôturé correspond à un sinistre dont le développement est considéré comme stabilisé définitivement. Ces probabilités de clôture serviront pour améliorer les probabilités de passages entre chaque cluster en prenons en compte le fait qu'un sinistre peut arrêter d'évoluer.

Enfin, grâce aux deux premiers chapitres, les mouvements entre cluster et les probabilités associés à ces mouvements seront connues. Il faudra alors calculer les évolutions des montants de sinistres liés à ces mouvements. Ainsi, les évolutions historiques serviront à créer une fonction de répartition historique des évolutions possibles entre deux clusters.

Le premier modèle créé permettra de donner une fonction de répartition des montants possibles du sinistre pour une future année de développement donnée. Cela permettra de créer les briques en vue de réaliser le modèle final. Ainsi, seront présentés la manière dont ce modèle a été réalisé, les difficultés rencontrées et les solutions mises en place.

Le second modèle généralisera le premier en donnant une estimation de la fonction de répartition du montant ultime du sinistre. Certaines simplifications ont été adoptées en vue de rendre réalisables certains calculs. Ces simplifications seront présentées en détail, j'expliquerai pourquoi ces choix ont été fait, et quels sont les conséquences de ces choix. Cette partie présentera aussi des pistes de réflexion pour améliorer la méthode proposée.

## 4.2 Transition entre clusters

### 4.2.1 Inspirations

L'idée de passer de cluster en cluster à chaque année de développement est inspirée de l'article « *Transition Matrix Theory And Individual Claim Loss Development* » (Mahon, 2005) [4]. L'auteur explique que les techniques de développement individuelle de sinistre ne sont pas adaptées à la réalité du processus d'évolution des sinistres et que cela n'est pas optimal pour la tarification en réassurance.

MAHON.J propose alors une méthode de développement basée sur une matrice de transitions entre des tranches de montant de sinistre. Pour chaque année de développement d'un sinistre, Mahon propose d'estimer, pour chaque transition entre année de développement, une matrice de transition

spécifique (une pour passer de l'année 2 à l'année 3, puis une nouvelle pour l'année 3 à l'année 4, etc. jusqu'à l'année de développement finale).

Je me suis inspiré de cette idée de transition entre des états pour chaque année de développement. La principale hypothèse faite, est que le processus de transition est un processus de Markov, c'est-à-dire que la prédiction du futur à partir du présent n'est pas rendue plus précise par des éléments d'information concernant le passé.

Pour cela il faut créer des matrices de transitions de cluster pour chaque année de développement.

### 4.2.2 Calcul des matrices de transition

Une matrice de transition est une matrice carrée contenant les probabilités de passer d'un état à une autre état (Moffat, 2018) [6]. Dans ce mémoire, les états initiaux seront les clusters de l'année de développements $k$ et les états finaux ceux de l'année $k+1$. La matrice sera de taille 26, car c'est le nombre de classe maximum pour une année de développement (année de développement 4). Dans la suite de cette partie, j'utiliserais comme exemple le passage entre l'année de développement 5 et l'année de développement 6 des sinistres auto français.

Matrice de comptage des clusters des sinistres auto français entre l'année 5 et l'année 6 de développement

| Classes | Classes initiales | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| finales | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | ... | 26 |
| 1 | 0 | 0 | 196 | 196 | 0 | 0 | 29 | 0 | 0 | 32 | 0 | 0 | 18 | 0 | 0 | 0 | 2 | 1 | 0 | 0 | 0 | 0 | ... | 0 |
| 2 | 0 | 0 | 22 | 22 | 53 | 0 | 21 | 0 | 0 | 9 | 0 | 1 | 0 | 0 | 1 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| 3 | 0 | 5 | 14 | 14 | 0 | 22 | 15 | 56 | 61 | 8 | 52 | 36 | 10 | 40 | 0 | 7 | 4 | 50 | 12 | 0 | 55 | 0 | ... | 0 |
| 4 | 0 | 12 | 0 | 0 | 0 | 39 | 0 | 13 | 85 | 1 | 5 | 26 | 0 | 17 | 0 | 15 | 2 | 149 | 215 | 1 | 332 | 0 | ... | 0 |
| 5 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | ... | 0 |
| 6 | 0 | 0 | 0 | 0 | 0 | 18 | 3 | 22 | 23 | 2 | 13 | 11 | 0 | 7 | 0 | 12 | 1 | 3 | 85 | 0 | 1 | 1 | ... | 0 |
| 7 | 2 | 7 | 1 | 1 | 0 | 544 | 54 | 107 | 245 | 1 | 1 | 530 | 150 | 386 | 2 | 14 | 36 | 174 | 0 | 0 | 6 | 0 | ... | 0 |
| 8 | 0 | 0 | 0 | 0 | 0 | 5 | 44 | 742 | 602 | 0 | 0 | 8 | 0 | 3 | 1 | 103 | 49 | 0 | 6 | 0 | 0 | 0 | ... | 0 |
| 9 | 0 | 0 | 74 | 74 | 0 | 4 | 193 | 21 | 53 | 305 | 907 | 482 | 47 | 85 | 0 | 10 | 10 | 1 | 1 | 0 | 0 | 0 | ... | 0 |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | ... | 0 |
| 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 26 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | 2 | 27 | 307 | 307 | 53 | 632 | 359 | 961 | 1071 | 358 | 978 | 1094 | 225 | 538 | 4 | 161 | 108 | 379 | 319 | 1 | 396 | 1 | 0 | 0 |

*Figure 17 - Exemple de matrice de comptage des clusters entre l'année de développement 5 et 6*

La première étape est de compter les changements de cluster entre les années, ces résultats peuvent être résumés dans une matrice de comptage :

En divisant chaque colonne par son total, nous obtenons la matrice de transition de Markov associé au processus de passage de l'année $k$ à l'année $k+1$. Chaque valeur représente la probabilité que la classe initiale associée fasse la transition vers la classe finale associée. On notera cette matrice de transition $C_k$

La matrice de transition de Markov $C_5$, associée à la Figure 17 est la suivante :

| Matrice de transition des clusters des sinistres auto français entre l'année 5 et l'année 6 de développement | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Classes | Classes initiales | | | | | | | | | | | | | | | | | | | | | | | |
| finales | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | ... | 26 |
| 1 | 0 | 0 | 0.64 | 0.64 | 0 | 0 | 0.08 | 0 | 0 | 0.09 | 0 | 0 | 0.08 | 0 | 0 | 0 | 0.02 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| 2 | 0 | 0 | 0.07 | 0.07 | 1 | 0 | 0.06 | 0 | 0 | 0.03 | 0 | 0 | 0 | 0 | 0.25 | 0 | 0.04 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| 3 | 0 | 0.19 | 0.05 | 0.05 | 0 | 0.03 | 0.04 | 0.06 | 0.06 | 0.02 | 0.05 | 0.03 | 0.04 | 0.07 | 0 | 0.04 | 0.04 | 0.13 | 0.04 | 0 | 0.14 | 0 | ... | 0 |
| 4 | 0 | 0.44 | 0 | 0 | 0 | 0.06 | 0 | 0.01 | 0.08 | 0 | 0.01 | 0.02 | 0 | 0.03 | 0 | 0.09 | 0.02 | 0.39 | 0.67 | 1 | 0.84 | 0 | ... | 0 |
| 5 | 0 | 0.11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| 6 | 0 | 0 | 0 | 0 | 0 | 0.03 | 0.01 | 0.02 | 0.02 | 0.01 | 0.01 | 0.01 | 0 | 0.01 | 0 | 0.07 | 0.01 | 0.01 | 0.27 | 0 | 0 | 1 | ... | 0 |
| 7 | 1 | 0.26 | 0 | 0 | 0 | 0.86 | 0.15 | 0.11 | 0.23 | 0 | 0 | 0.48 | 0.67 | 0.72 | 0.5 | 0.09 | 0.33 | 0.46 | 0 | 0 | 0.02 | 0 | ... | 0 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0.01 | 0.12 | 0.77 | 0.56 | 0 | 0 | 0.01 | 0 | 0.01 | 0.25 | 0.64 | 0.45 | 0 | 0.02 | 0 | 0 | 0 | ... | 0 |
| 9 | 0 | 0 | 0.24 | 0.24 | 0 | 0.01 | 0.54 | 0.02 | 0.05 | 0.85 | 0.93 | 0.44 | 0.21 | 0.16 | 0 | 0.06 | 0.09 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 26 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |

*Figure 18- Exemple de matrice de transition des clusters entre l'année de développement 5 et 6*

Cette matrice présente pour chaque colonne les probabilités de passer d'un état initial en année de développement 5 à un état final en année de développement 6.

Dans le cas où les matrices de transitions entre chaque année sont connues, il est alors possible de connaître les probabilités pour qu'un sinistre dans le cluster $i$ en année $k$, soit dans le cluster $j$ en année $n$.

Prenons un exemple avec les sinistres auto français en année de développement 5 classés dans le cluster 6. Tracer le Diagramme des catégories parallèles permet de comprendre les mouvements entre l'année de développement 5 et des années ultérieures :

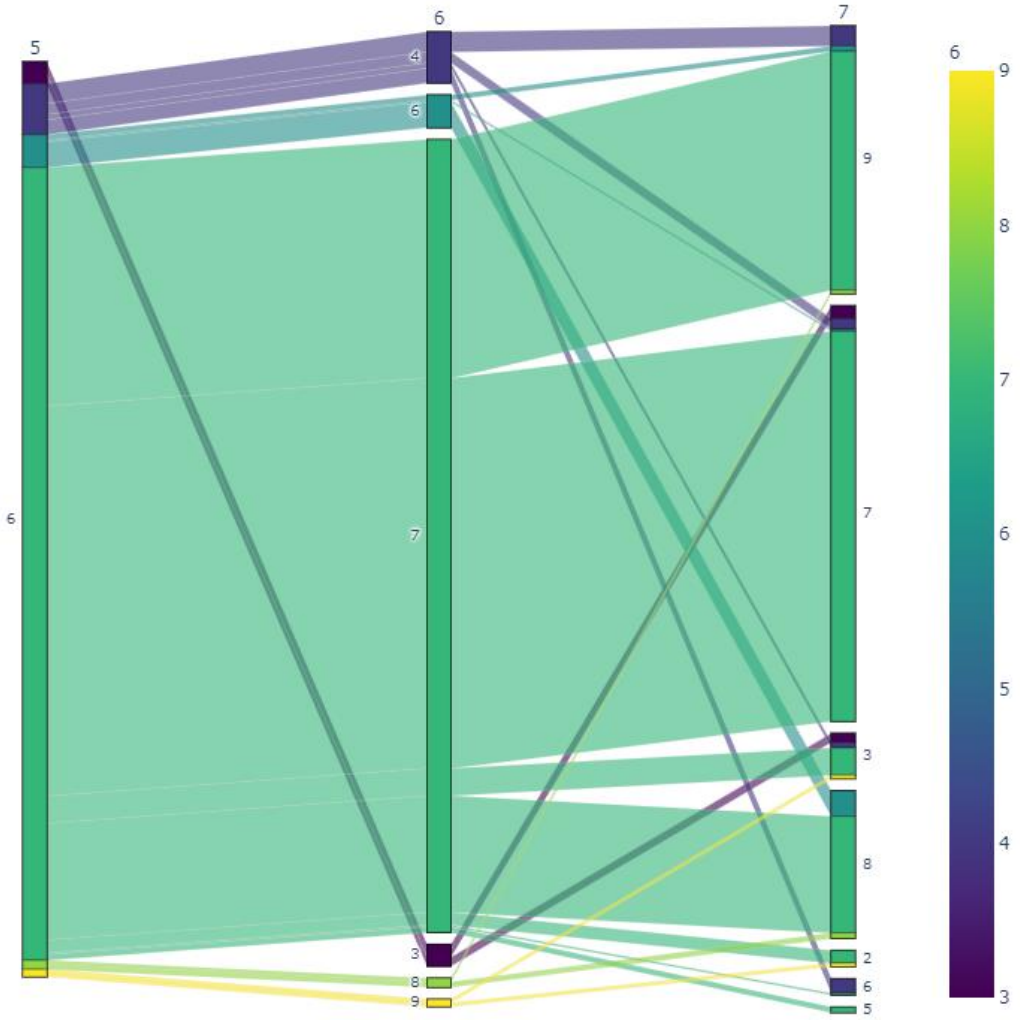


*Figure 19 - Evolution à partir du cluster 6 de l'année de développement 5 vers l'année de développement 7*

Grace à cette représentation graphique, il est possible de visualiser les évolutions entre chaque année de développement.

Cependant cette visualisation est limitée par le fait que les classes ne soient pas les même chaque année. En effet, les classes sont créées automatiquement par l'algorithme de classification ascendante hiérarchique et donc le cluster 6 de l'année 5 n'a, à priori, aucun rapport avec les clusters 6 des autres années de développement.

Cependant la transition entre les clusters n'est pas le seul phénomène de transition qui doit être étudié. L'étude du passage de l'état ouvert vers l'état clôturé doit aussi être effectuée.

## 4.3 Calcul des probabilités de clôture

On appelle ouvert un sinistre dont l'estimation du coût final n'est pas encore considérée comme définitif. A l'inverse, on dit qu'un sinistre est clôturé lorsque le coût final de ce sinistre est connu et définitif.

En réassurance, les sinistres traités sont le plus souvent les grands sinistres. Ces grands sinistres sont très souvent des accidents corporels graves. Or, le coût de ces sinistres dépend de l'état de santé de(s) la victime(s). Etat de santé qui est souvent instable. De plus la responsabilité des sinistres peut faire l'objet d'une bataille juridique entre les assureurs. Pour ces raisons, les sinistres peuvent mettre plusieurs années avant de se clôturer (parfois plus de 15 ans).

### 4.3.1 Calcul des probabilités

En vue de prendre en compte les différentes possibilités de clôture dans mon modèle, la proportion de sinistres clôturés a été estimée, pour chaque cluster de chaque année de développement. Cette proportion nous servira d'estimation de la probabilité qu'à un sinistre de se clôturer à chaque année de développement en fonction du cluster qu'il rejoindra. On notera la matrice de transition entre l'état ouvert et clôturé de l'année $k, T_k$

| Matrice de comptage des sinistres auto français en année 6 de développemen clôturés par clusters | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Classes finales | Classes initiales | | | | | | | | | |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Ouvert | 297 | 145 | 324 | 833 | 194 | 2007 | 1532 | 2140 | 7 | 1 |
| Clôt | 1 | 16 | 123 | 79 | 8 | 259 | 31 | 132 | 0 | 0 |
| Total | 298 | 161 | 447 | 912 | 202 | 2266 | 1563 | 2272 | 7 | 1 |

*Figure 20 - Exemple de matrice de comptage des sinistres clôturés (auto français année 6)*

Pour calculer ces probabilités j'utilise, comme au précédent chapitre, une matrice de comptage :

De la même manière que précédemment, la matrice de transition $T_6$ est la suivante:

| Matrice de markov des sinistres auto français en année 6 de développemen clôturés par clusters | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Classes finales | Classes initiales | | | | | | | | | |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Ouvert | 0.9966 | 0.9006 | 0.7248 | 0.9134 | 0.9604 | 0.8857 | 0.9802 | 0.9419 | 1 | 1 |
| Clôt | 0.0034 | 0.0994 | 0.2752 | 0.0866 | 0.0396 | 0.1143 | 0.0198 | 0.0581 | 0 | 0 |
| Total | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

*Figure 21 - Exemple de matrice de transition des sinistres clôturés (auto français année 6)*

Il est alors possible de combiner la matrice de transition entre les clusters et celle de transition vers l'état de clôture. Cette nouvelle matrice représente entièrement les états que peuvent prendre les sinistres. Elle sera notée $M_{i,j}$ (Matrice finale de transition entre l'année i et l'année j).

Voici la matrice $M_{5,6}$:

Matrice de transition des clusters des sinistres auto français entre l'année 5 et l'année 6 de développement

| Classes finales | Classe initiales 1 Ouv | 2 Ouv | 3 Ouv | 4 Ouv | 5 Ouv | 6 Ouv | 7 Ouv | 8 Ouv | 9 Ouv | 10 Ouv | 11 Ouv | 12 Ouv | 13 Ouv | 14 Ouv | 15 Ouv | 16 Ouv | 17 Ouv | 18 Ouv | 19 Ouv | 20 Ouv | 21 Ouv | 22 Ouv | ... | 26 Ouv | 1 Clôt | 2 Clôt | 3 Clôt | 4 Clôt | 5 Clôt | 6 Clôt | 7 Clôt | 8 Clôt | 9 Clôt | 10 Clôt | 11 Clôt | 12 Clôt | 13 Clôt | 14 Clôt | 15 Clôt | 16 Clôt | 17 Clôt | 18 Clôt | 19 Clôt | 20 Clôt | 21 Clôt | 22 Clôt | ... | 26 Clôt |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 Ouvert | 0 | 0 | 0.6 | 0.6 | 0 | 0 | 0.1 | 0 | 0 | 0.1 | 0 | 0 | 0.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| 2 Ouvert | 0 | 0 | 0.1 | 0.1 | 0.9 | 0 | 0.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| 3 Ouvert | 0 | 0.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.1 | 0 | 0 | 0 | 0.1 | 0 | 0 | 0.1 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| 4 Ouvert | 0 | 0.4 | 0 | 0 | 0 | 0.1 | 0 | 0 | 0.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0.1 | 0 | 0.4 | 0.6 | 0.9 | 0.8 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| 5 Ouvert | 0 | 0.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| 6 Ouvert | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.1 | 0 | 0 | 0.2 | 0 | 0 | 0.9 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | ... | 0 |
| 7 Ouvert | 1 | 0.3 | 0 | 0 | 0 | 0.8 | 0.1 | 0.1 | 0.2 | 0 | 0 | 0.5 | 0.7 | 0.7 | 0.5 | 0.1 | 0.3 | 0.4 | 0 | 0 | 0 | 0 | ... | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| 8 Ouvert | 0 | 0 | 0 | 0 | 0 | 0 | 0.1 | 0.7 | 0.5 | 0 | 0 | 0 | 0 | 0 | 0.2 | 0.6 | 0.4 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| 9 Ouvert | 0 | 0 | 0.2 | 0.2 | 0 | 0 | 0.5 | 0 | 0 | 0.9 | 0.9 | 0.4 | 0.2 | 0.2 | 0 | 0.1 | 0.1 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| 10 Ouvert | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| 11 Ouvert | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 26 Ouvert | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 Clôt | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| 2 Clôt | 0 | 0 | 0 | 0 | 0.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| 3 Clôt | 0 | 0.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| 4 Clôt | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.1 | 0.1 | 0.1 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| 5 Clôt | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| 6 Clôt | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.1 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| 7 Clôt | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| 8 Clôt | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| 9 Clôt | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| 10 Clôt | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| 11 Clôt | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 26 Clôt | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| Total | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |

*Figure 22 - Matrice de transition finale pour les sinistres auto français en année de développement 5*

Cette matrice est composée de quatre parties. En notant $t_{i,j}$ les éléments de la matrice de transition entre cluster, $c_{i,j}$ les éléments de la matrice de transition de l'état de clôture et $m_{i,j}$ les éléments de la matrice de transition finale, il devient facile de décrire les 4 parties de la matrice :

Le coin supérieur gauche correspond à la transition entre un sinistre ouvert en année de développement $k$ et un sinistre ouvert en année de développement $k+1$. Dans ce cas, on a :

$$m_{i,j} = t_{i,j} \times c_{2,i}$$

Voici le résultat associé à l'exemple :

| Classes finales | 1 Ouvert | 2 Ouvert | 3 Ouvert | 4 Ouvert | 5 Ouvert | 6 Ouvert | 7 Ouvert | 8 Ouvert | 9 Ouvert | 10 Ouvert | 11 Ouvert | 12 Ouvert | 13 Ouvert | 14 Ouvert | 15 Ouvert | 16 Ouvert | 17 Ouvert | 18 Ouvert | 19 Ouvert | 20 Ouvert | 21 Ouvert | 22 Ouvert | ... | Classe i 26 Ouvert |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 Ouver | 0 | 0 | 0.636 | 0.636 | 0 | 0 | 0.081 | 0 | 0 | 0.089 | 0 | 0 | 0.08 | 0 | 0 | 0 | 0.018 | 0.003 | 0 | 0 | 0 | 0 | ... | 0 |
| 2 Ouver | 0 | 0 | 0.065 | 0.065 | 0.901 | 0 | 0.053 | 0 | 0 | 0.023 | 0 | 8E-04 | 0 | 0 | 0.225 | 0 | 0.033 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| 3 Ouver | 0 | 0.134 | 0.033 | 0.033 | 0 | 0.025 | 0.03 | 0.042 | 0.041 | 0.016 | 0.039 | 0.024 | 0.032 | 0.054 | 0 | 0.032 | 0.027 | 0.096 | 0.027 | 0 | 0.101 | 0 | ... | 0 |
| 4 Ouver | 0 | 0.406 | 0 | 0 | 0 | 0.056 | 0 | 0.012 | 0.072 | 0.003 | 0.005 | 0.022 | 0 | 0.029 | 0 | 0.085 | 0.017 | 0.359 | 0.616 | 0.913 | 0.766 | 0 | ... | 0 |
| 5 Ouver | 0 | 0.107 | 0 | 0 | 0 | 0 | 0 | 0 | 0.002 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.003 | 0 | 0 | 0.002 | 0 | ... | 0 |
| 6 Ouver | 0 | 0 | 0 | 0 | 0 | 0.025 | 0.007 | 0.02 | 0.019 | 0.005 | 0.012 | 0.009 | 0 | 0.012 | 0 | 0.066 | 0.008 | 0.007 | 0.236 | 0 | 0.002 | 0.886 | ... | 0 |
| 7 Ouver | 0.98 | 0.254 | 0.003 | 0.003 | 0 | 0.844 | 0.147 | 0.109 | 0.224 | 0.003 | 0.001 | 0.475 | 0.653 | 0.703 | 0.49 | 0.085 | 0.327 | 0.45 | 0 | 0 | 0.015 | 0 | ... | 0 |
| 8 Ouver | 0 | 0 | 0 | 0 | 0 | 0.007 | 0.115 | 0.727 | 0.529 | 0 | 0 | 0.007 | 0 | 0.005 | 0.235 | 0.603 | 0.427 | 0 | 0.018 | 0 | 0 | 0 | ... | 0 |
| 9 Ouver | 0 | 0 | 0.241 | 0.241 | 0 | 0.006 | 0.538 | 0.022 | 0.049 | 0.852 | 0.927 | 0.441 | 0.209 | 0.158 | 0 | 0.062 | 0.093 | 0.003 | 0.003 | 0 | 0 | 0 | ... | 0 |
| 10 Ouve | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.003 | 0 | ... | 0 |
| 11 Ouve | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 26 Ouve | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

*Figure 23 - Coin supérieur gauche de la matrice de transition finale*

Le coin inférieur gauche correspond à la transition entre un sinistre ouvert en année de développement $k$ et un sinistre clôturé en année de développement $k+1$. Dans ce cas, on a :

$$m_{i,j} = t_{i,j} \times c_{1,i}$$

Voici le résultat associé à l'exemple :

| | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 Clôt | 0 | 0 | 0.002 | 0.002 | 0 | 0 | 3E-04 | 0 | 0 | 3E-04 | 0 | 0 | 3E-04 | 0 | 0 | 0 | 6E-05 | 9E-06 | 0 | 0 | 0 | 0 | ... | | 0 |
| 2 Clôt | 0 | 0 | 0.007 | 0.007 | 0.099 | 0 | 0.006 | 0 | 0 | 0.002 | 0 | 9E-05 | 0 | 0 | 0.025 | 0 | 0.004 | 0 | 0 | 0 | 0 | 0 | ... | | 0 |
| 3 Clôt | 0 | 0.051 | 0.013 | 0.013 | 0 | 0.01 | 0.011 | 0.016 | 0.016 | 0.006 | 0.015 | 0.009 | 0.012 | 0.02 | 0 | 0.012 | 0.01 | 0.036 | 0.01 | 0 | 0.038 | 0 | ... | | 0 |
| 4 Clôt | 0 | 0.038 | 0 | 0 | 0 | 0.005 | 0 | 0.001 | 0.007 | 2E-04 | 4E-04 | 0.002 | 0 | 0.003 | 0 | 0.008 | 0.002 | 0.034 | 0.058 | 0.087 | 0.073 | 0 | ... | | 0 |
| 5 Clôt | 0 | 0.004 | 0 | 0 | 0 | 0 | 0 | 0 | 7E-05 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1E-04 | 0 | 0 | 1E-04 | 0 | ... | | 0 |
| 6 Clôt | 0 | 0 | 0 | 0 | 0 | 0.003 | 1E-03 | 0.003 | 0.002 | 6E-04 | 0.002 | 0.001 | 0 | 0.001 | 0 | 0.009 | 0.001 | 9E-04 | 0.03 | 0 | 3E-04 | 0.114 | ... | | 0 |
| 7 Clôt | 0.02 | 0.005 | 6E-05 | 6E-05 | 0 | 0.017 | 0.003 | 0.002 | 0.005 | 6E-05 | 2E-05 | 0.01 | 0.013 | 0.014 | 0.01 | 0.002 | 0.007 | 0.009 | 0 | 0 | 3E-04 | 0 | ... | | 0 |
| 8 Clôt | 0 | 0 | 0 | 0 | 0 | 5E-04 | 0.007 | 0.045 | 0.033 | 0 | 0 | 4E-04 | 0 | 3E-04 | 0.015 | 0.037 | 0.026 | 0 | 0.001 | 0 | 0 | 0 | ... | | 0 |
| 9 Clôt | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | | 0 |
| 10 Clôt | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | | 0 |
| 11 Clôt | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 26 Clôt | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | |

*Figure 24 - Coin inférieur gauche de la matrice de transition finale*

Le coin supérieur droit n'est composé que de 0, car il représente la probabilité pour un sinistre clôturé de se rouvrir. Ce qui est par définition impossible.

Le coin inférieur droit est lui composé de la matrice identité $I_{26}$, car un sinistre clôturé ne peut que rester clôturé.

### 4.3.2 Exemple d'analyse des taux de clôture

Il est aussi possible d'analyser la répartition des clôtures via certaines visualisations. Voici une visualisation des clôtures des sinistres auto français en année de développement 6:

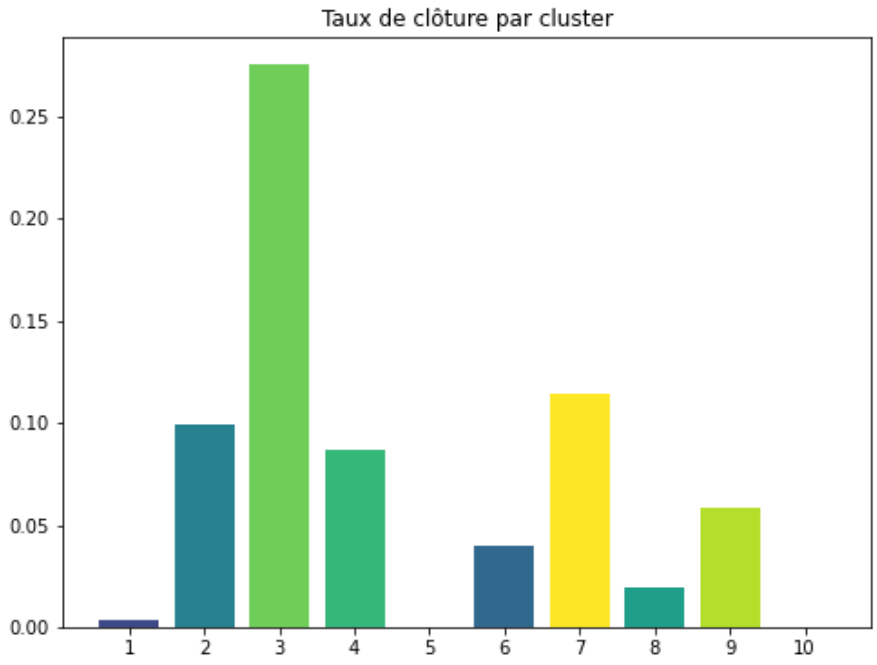


*Figure 25 - Exemple de taux de clôture pour les sinistres auto français en année de développement 6*

On remarque une grande hétérogénéité du taux de clôture selon les clusters. Ainsi, on ne peut que confirmer importance de bien prendre en compte la clôture dans les passages entre années de développement.

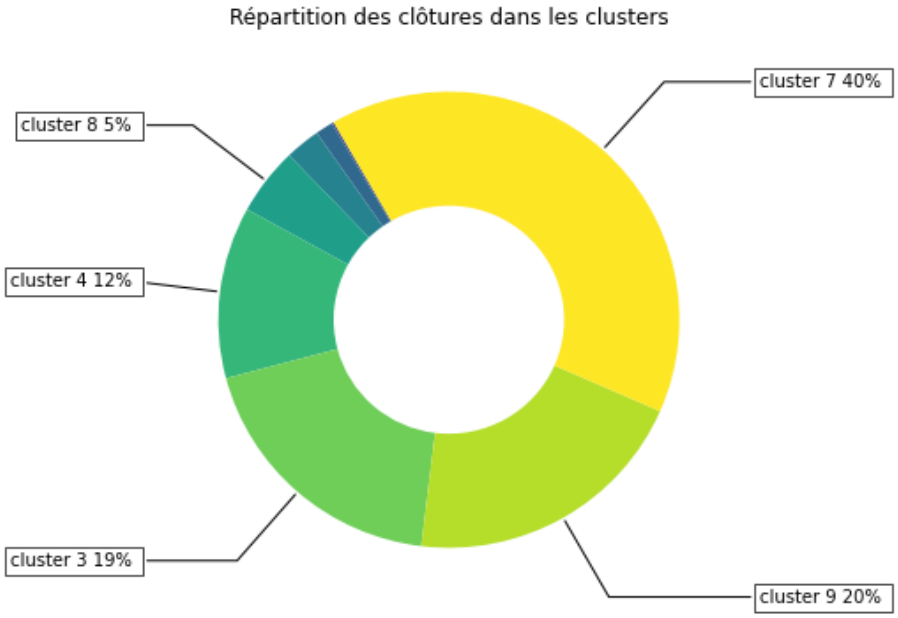


*Figure 26 - Répartition des clôtures selon les clusters pour les sinistres auto français en année de développement 6*

Ce que l'on peut voir ici, c'est que 5 des 10 clusters représentent 96% des clôtures. Ce sont sans surprise les clusters avec le plus d'observations qui représentent la plus grande part des sinistres clôturés.

## 4.4 Calcul de la fonction de répartition des évolutions

Maintenant que les probabilités des différentes transitions entre les clusters et des états de clôture des sinistres ont été calculées, il faut quantifier l'évolution du montant de sinistre à chaque transition.

Dans un premier temps, l'objectif sera de définir une fonction de répartition des montants finaux de chaque cluster de chaque année de développement, pour chaque année de développement future. Pour ce faire, les fonctions de répartitions des évolutions du montant des sinistres entre deux années de développement et deux clusters seront calculées. Puis elles seront combinées pour obtenir des fonctions de répartitions des évolutions du montant de sinistre pour toutes les années de sinistralité futures.

Dans un second temps, les résultats pratiques engendrés par la méthode présentée dans ce mémoire seront préentés.

### 4.4.1 Calcul de la fonction de répartition

La première étape pour calculer les fonctions de répartition des évolutions des montants de sinistre est de lister pour toutes les années de développement $i, i \in \{0, \dots, n-1\}$ les évolutions historiques du montant de sinistres pour chaque cluster de l'année de développement $i$ et vers l'année de développement $i+1$. Pour plus de clarté, les notations suivantes seront utilisées :

- $E_{i,c,j,k}$ : l'évolution des sinistres du cluster $c$ en année $i$ vers le cluster k dans l'année j.
- $E_{i,c}$ : l'évolution sur un an des sinistres du cluster $c$ en année $i$.

Prenons par exemple le passage entre le cluster 6 de l'année de développement 5 et le cluster 7 de l'année de développement 6. On peut calculer la fonction de répartition empirique $F_e$ des évolutions du montant estimé du sinistre $E_{5,6,6,7}$ :

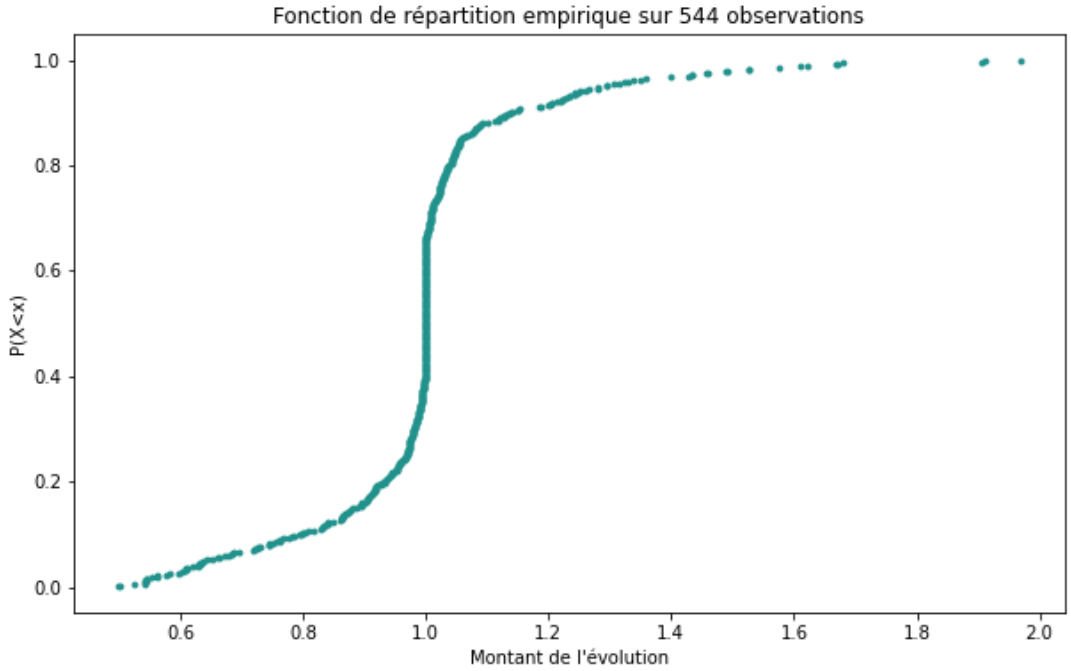


*Figure 27 - Exemple de fonction de répartition empirique*

L'objectif est de calculer toutes les fonctions de répartition des $E_{i,c,j,k}$ pour pouvoir les combiner et trouver la fonction de répartition des évolutions de chaque chemin de transition pour chaque cluster de départ de chaque année développement vers un année de développement future.

Par exemple, dans le cas d'un sinistre en année de développement 5, dans le cluster 6, il peut prendre un des 42 chemins suivant pour se développer jusqu'en année 8 :

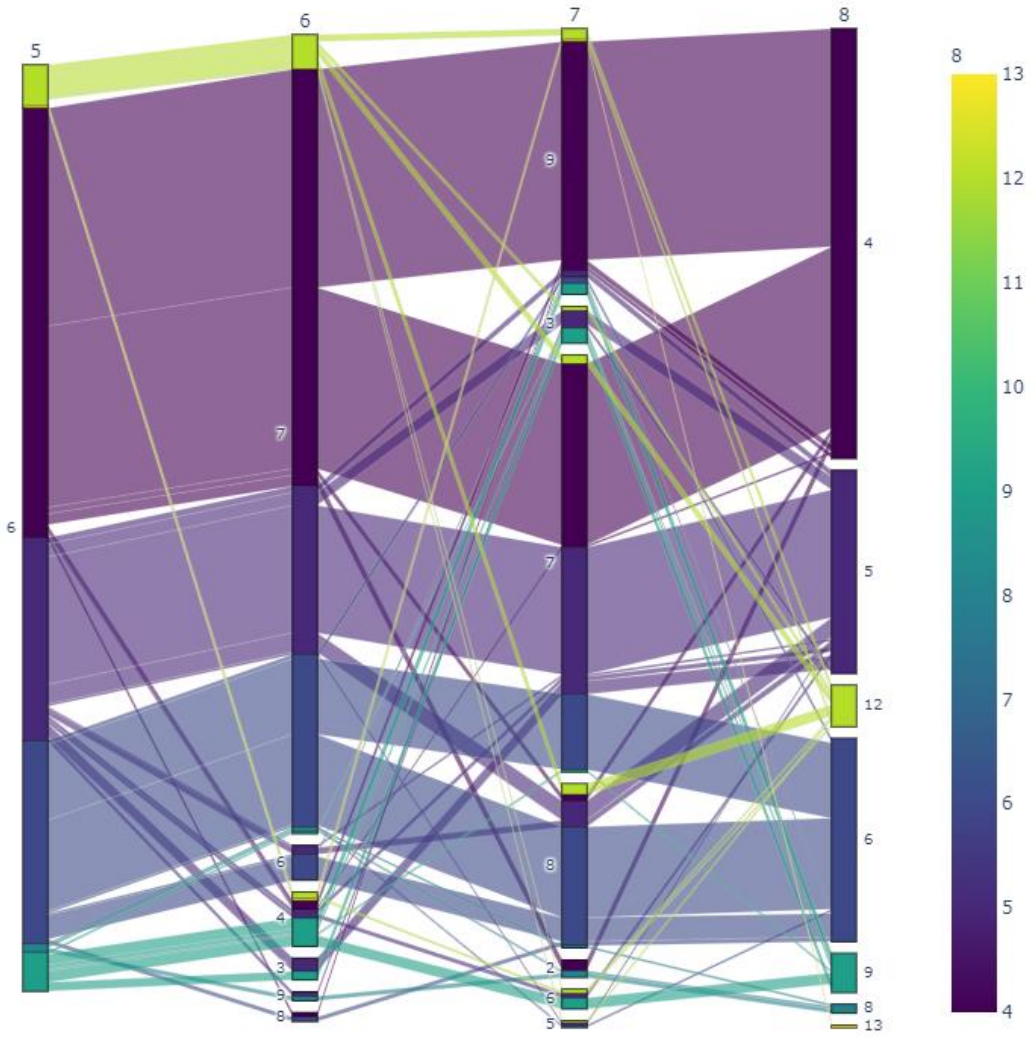


*Figure 28 - Chemins de développement possibles pour un sinistre en année 5 (cluster 6) vers l'année 8*

Pour chaque chemin, il faut combiner les différents $E_{i,c,j,k}$ correspondant. Par Exemple, pour le chemin 7-9-4, il faut trouver les trois fonctions de répartition de $E_{5,6,6,7}$, $E_{6,7,7,9}$, et $E_{7,9,8,4}$ :

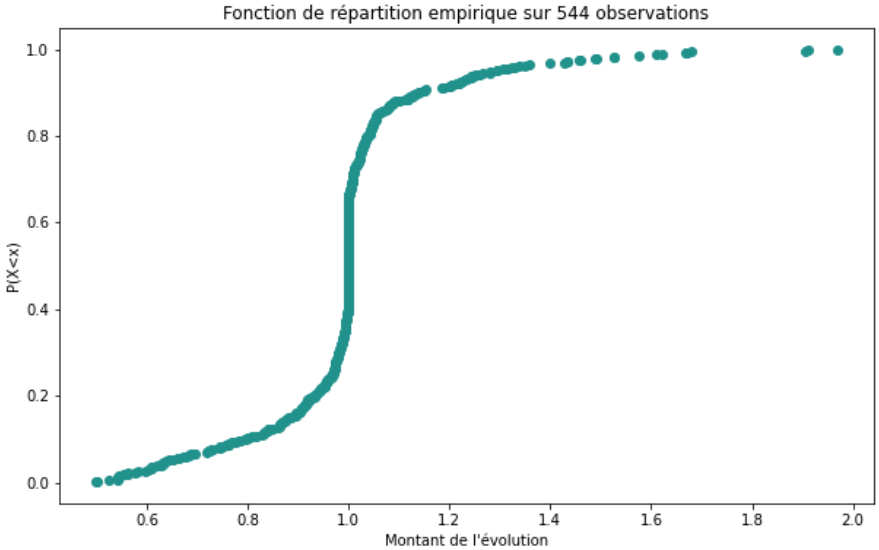


*Figure 29 - Fonction de répartition entre le cluster 6 (année 5) et le cluster 7 (année 6)*

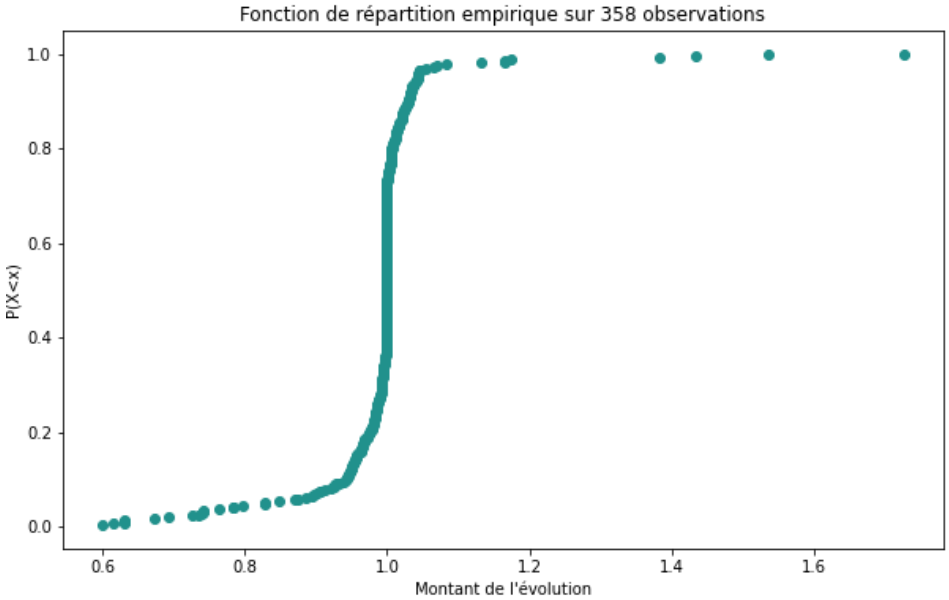


*Figure 30 - Fonction de répartition entre le cluster 7 (année 6) et le cluster 9 (année 7)*

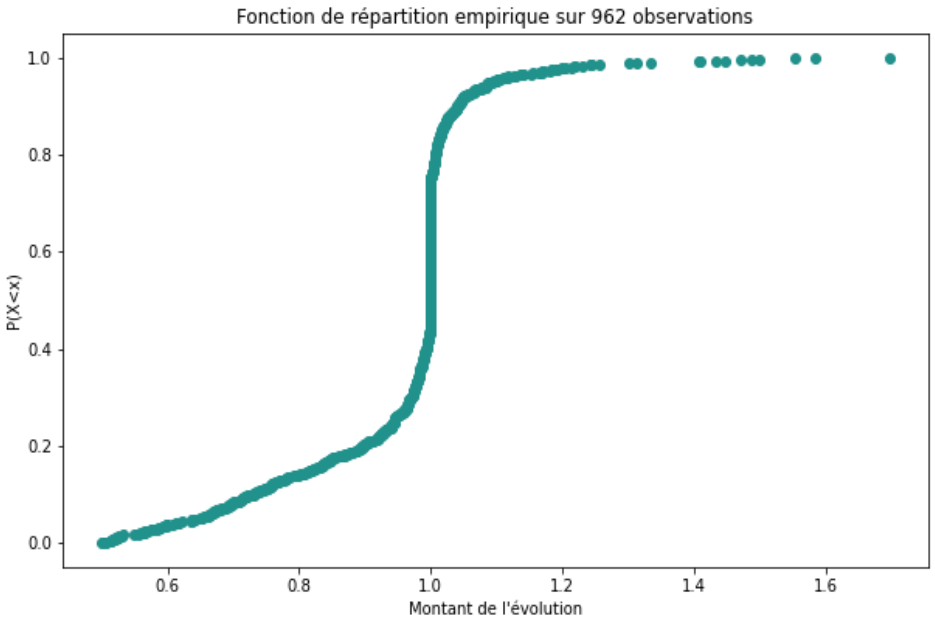


*Figure 31 - Fonction de répartition entre le cluster 9 (année 7) et le cluster 4 (année 8)*

Ensuite, je pose la notation suivante : $d_{i,c,j,k,l}$ correspondant à la l[ème] évolution du montant de sinistre entre le cluster $c$ de l'année $i$ et le cluster $k$ de l'année $j$. Ainsi, dans l'exemple ci-dessus, la fonction de répartition estimée des évolutions des sinistres empruntant ce chemin est l'ensemble des $\{d_{5,6,6,7,l6} * d_{6,7,7,9,l7} * d_{7,9,8,4,l8}, \forall l6 \in \{1, \dots, 544\}, \forall l7 \in \{1, \dots, 358\}, \forall l8 \in \{1, \dots, 962\}\}$

Chacun de ces évolutions à pour probabilité $P(d_{5,6,6,7,l6}) * P(d_{6,7,7,9,l7}) * P(d_{7,9,8,4,l8}), \forall l6 \in \{1, \dots, 544\}, \forall l7 \in \{1, \dots, 358\}, \forall l8 \in \{1, \dots, 962\}\}$

La première étape de la construction du modèle est alors de calculer tous les chemins possibles pour toutes les combinaisons d'année de développement d'origine, de cluster d'origine, d'année de développement finale et de cluster finale.

Ensuite, il faut agréger ces chemins en pondérant la probabilité de chacun des événements discrets de chaque chemin par la probabilité que ce chemin de développement soit emprunté par le sinistre. Le chemin 7-9-4 a, par exemple, une probabilité de $0{,}844 \times 0{,}172 \times 0{,}441 = 0{,}0632$ d'être emprunté par un sinistre du cluster 6 en année de développement 5. Et donc l'ensemble des probabilités d'évolutions associées à ce chemin doivent être pondérées par $0{,}0632$.

### 4.4.2 Complexité des calculs et proposition de simplification

Il est alors facile de se rendre compte que le calcul de la fonction de répartition pour des chemins long (tels que ce supérieur à trois années de développement) est extrêmement couteux en temps de calcul. Dans l'exemple précédent, déjà $187000000$ coefficients d'évolution étaient à calculer. Le nombre de coefficients à calculer est exponentiel en fonction du nombre d'années de développement. De plus, dans le cadre d'un sinistre appartenant au cluster 6 en année de développement 5, il y a 42 chemins de développement à 3 an. Comme pour les coefficients, le nombre de chemin augmente de manière exponentielle avec le nombre d'année de développement.

Pour rendre possible la réalisation des calculs, il faut adopter une simplification permettant de concentrer l'information disponible sur les évolutions des montants estimés de sinistres lors d'une transition entre clusters.

Pour faire cela, j'ai choisi de résumé des morceaux de la fonction de répartition des évolutions par leur moyenne. La Fonction de répartition simplifiée des évolutions du montant de sinistre, notée $F_s$, comportera 4 points :

- 0, tel que$F_s(0) = F_e(0.05)$
- $B = E\left[d_{i,c,j,k,l} | 0.05 < d_{i,c,j,k,l} < 0.98\right]$, tel que $F_s(B) = F_e(0.98)$
- 1, tel que$F_s(1) = F_e(1.02)$
- $H = E\left[d_{i,c,j,k,l} | 1.02 < d_{i,c,j,k,l}\right]$, tel que $F_s(H) = 1$

Voici un exemple pour rendre cela plus clair :

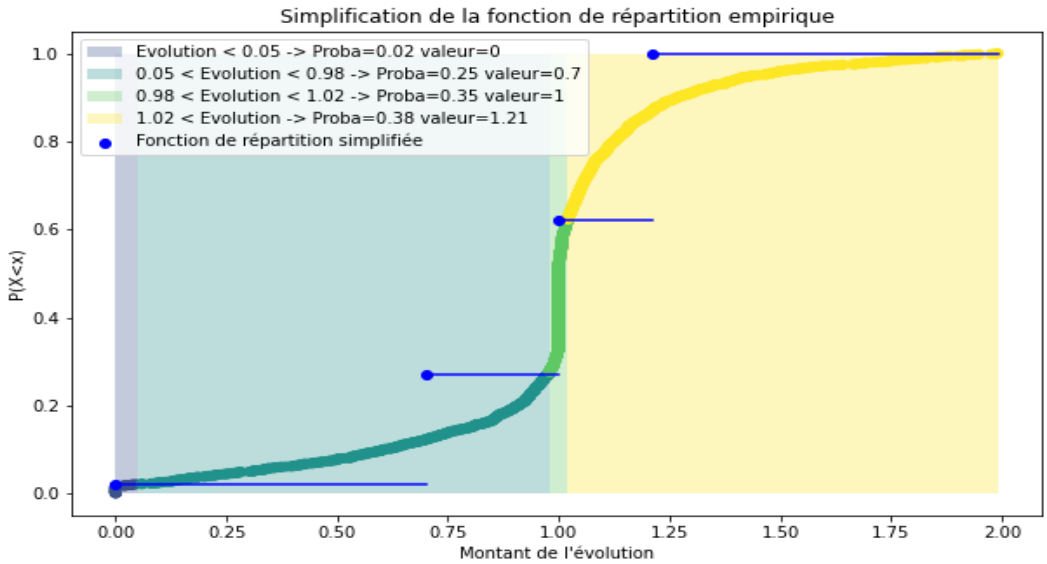


*Figure 32 - Exemple de simplification de fonction de répartition*

Cette simplification peut paraitre brutale, notamment pour les valeurs d'évolution élevées. C'est pour cela que dans la présentation des résultats, les grandes évolutions seront étudiées avec une grande attention.

Grâce à cette simplification, il devient possible de calculer les différentes fonctions de répartition simplifiées de chaque chemin. En effet, ces fonctions de répartitions simplifiées contiennent au maximum 4 observations. Dans ce cas, l'exemple précédent du chemin 7-9-4 ne comporterai pas $187000000$ coefficients mais $4 \times 4 \times 4 = 64$ coefficients.

Il suffit maintenant de calculer les 64 coefficients des 42 chemins allant du cluster 6 de l'année de développement 5 vers l'année de développement 8 et de les combiner pour pouvoir obtenir la fonction de répartition des évolutions des sinistres entre le cluster 6 (année 5) et l'année 8.

En réalisait cette tâche pour tous les clusters de toutes les années de développement, et pour chaque année de développement finale, j'obtiens un modèle permettant de prédire les évolutions possibles vers une année de développement future donnée.

Ainsi, la fonction de répartition des évolutions pour les sinistres du cluster 6 en année de développement 5 vers l'année de développement 8 est la suivante :

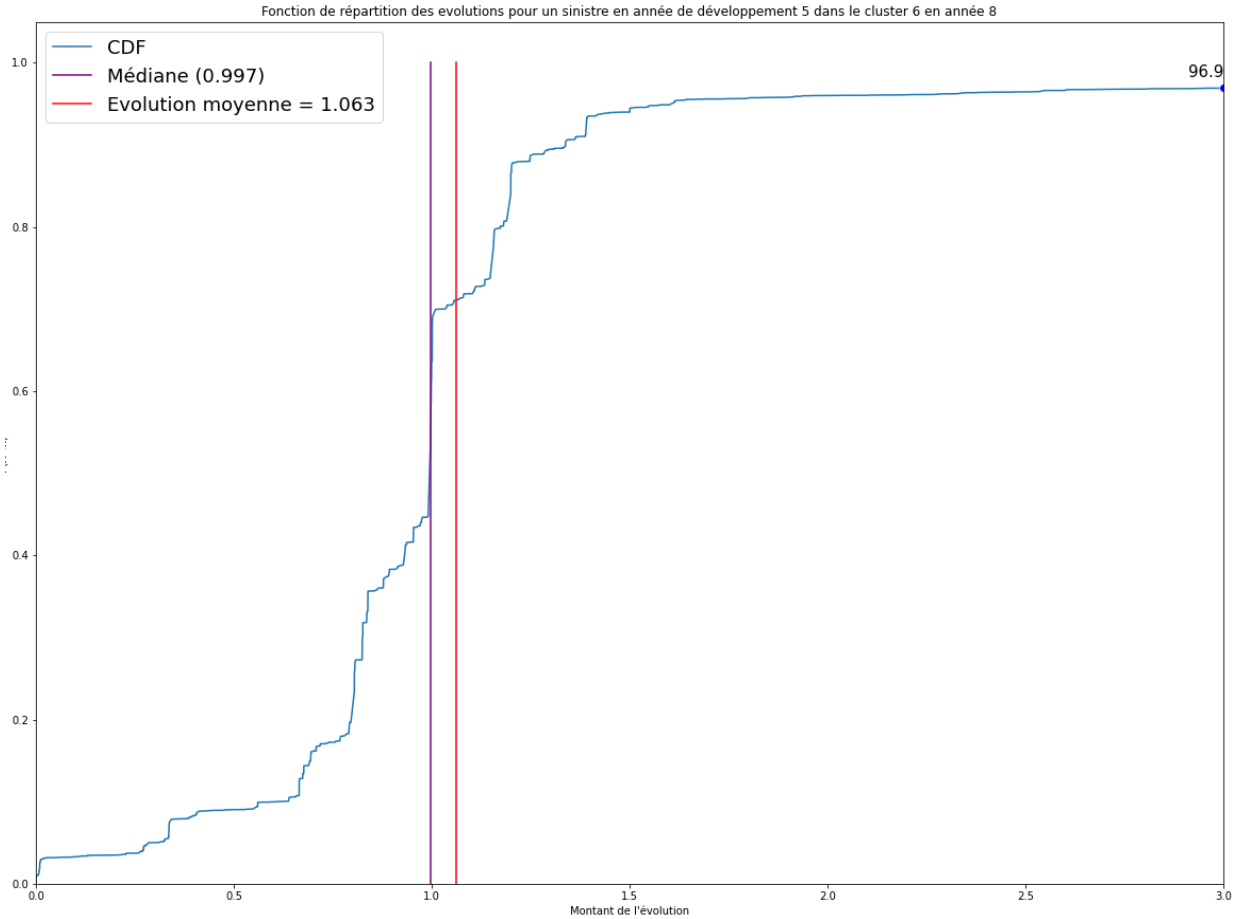


*Figure 33 – Fonction de répartition des évolutions du montant de sinistres entre le cluster 6 (année 5) et l'année 8*

Avec cette fonction de répartition, il est aisé de se rendre compte du progrès réalisé dans la connaissance du risque d'évolution du montant du sinistre. En effet, l'utilisation de la méthode de Chain Ladder ne permettra seulement de connaitre le coefficient moyen d'évolution.

Ce progrès est d'autant plus visible que l'année de développement finale est lointaine. Par exemple, voici la fonction de répartition des développements entre les sinistres du cluster 4 de l'année de développement 2 et l'année de développement 13 :

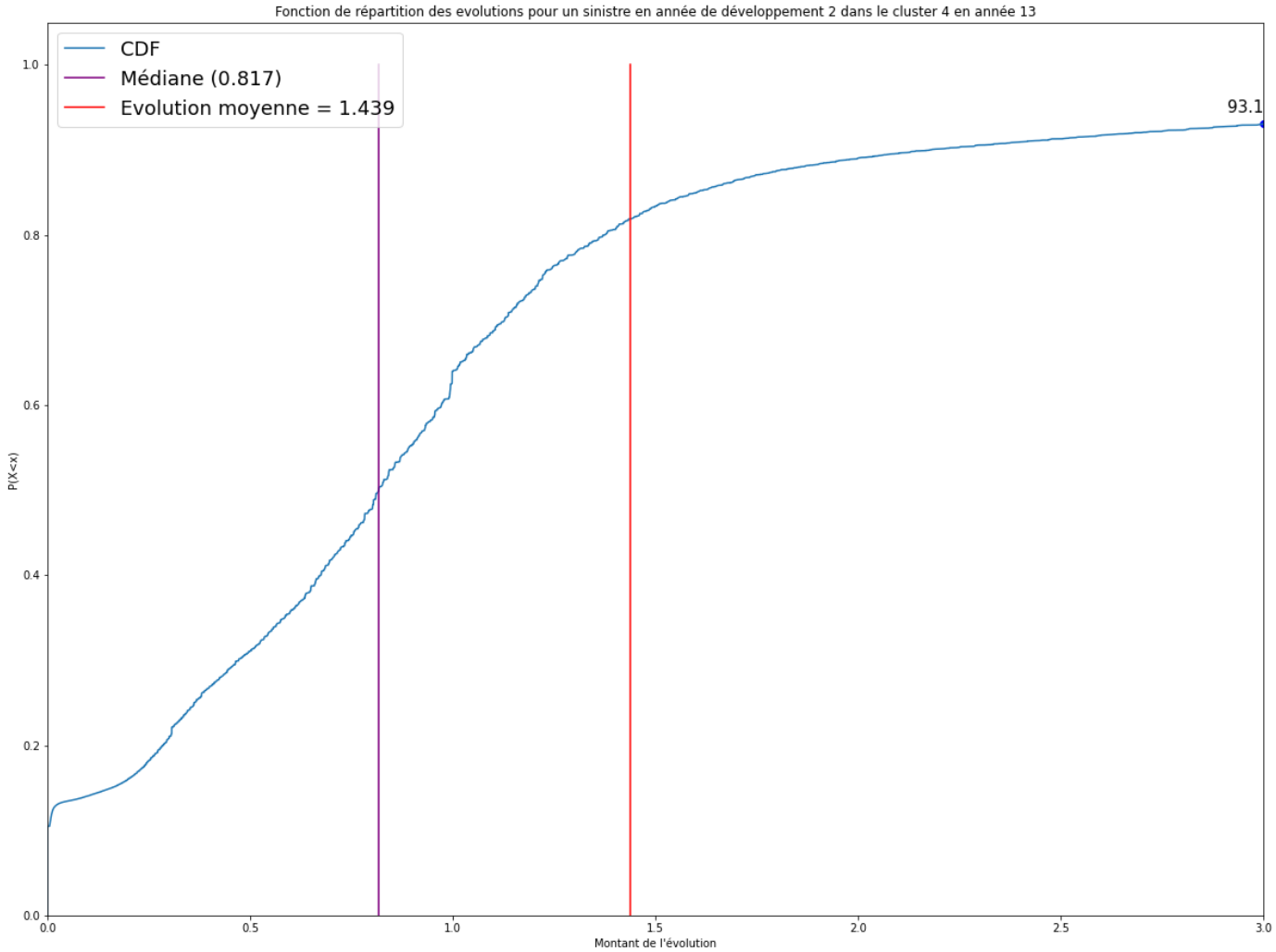


*Figure 34 - Fonction de répartition des évolutions de sinistres du cluster 4 de l'année de développement 2 et l'année de développement 13*

Cet exemple montre vraiment que l'estimation proposée est très intéressante pour comprendre les évolutions possibles des montants de sinistres, encore plus lorsque la méthode de Chain ladder ne permet d'obtenir qu'un coefficient moyen.

Il est aussi intéressant de noter qu'il y a une masse de proba en 0, cela aura de l'importance pour la suite de la réalisation du modèle.

### 4.4.3 Présentation des résultats

Pour étudier la qualité des résultats obtenus, j'ai choisi de créer des intervalles autour de la médiane des évolutions. Ensuite, la proportion de sinistres réels présents dans l'intervalle avec la taille de l'intervalle seront comparées. J'ai choisi de représenter trois intervalles : 25%, 50% et 75%.

Voici des exemples graphiques sur les sinistres du cluster 3 en année de développement 2, développés jusqu'en année 13 :

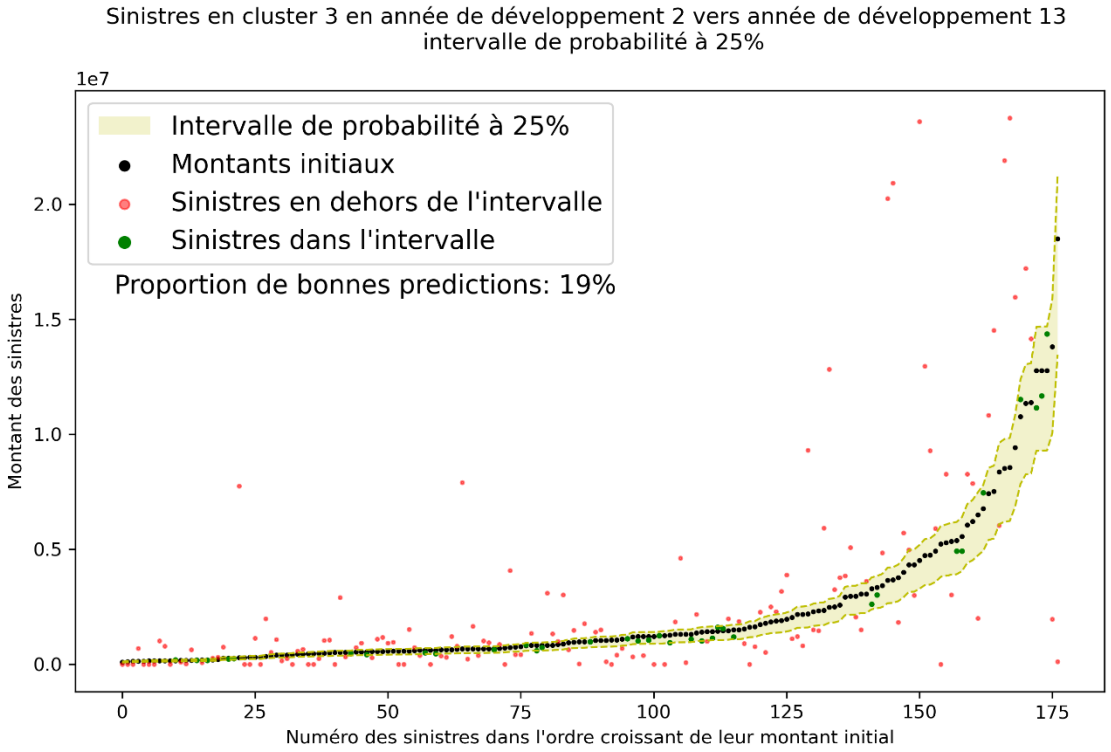


*Figure 35 - Exemple de cohérence (intervalle de probabilité à 25%)*

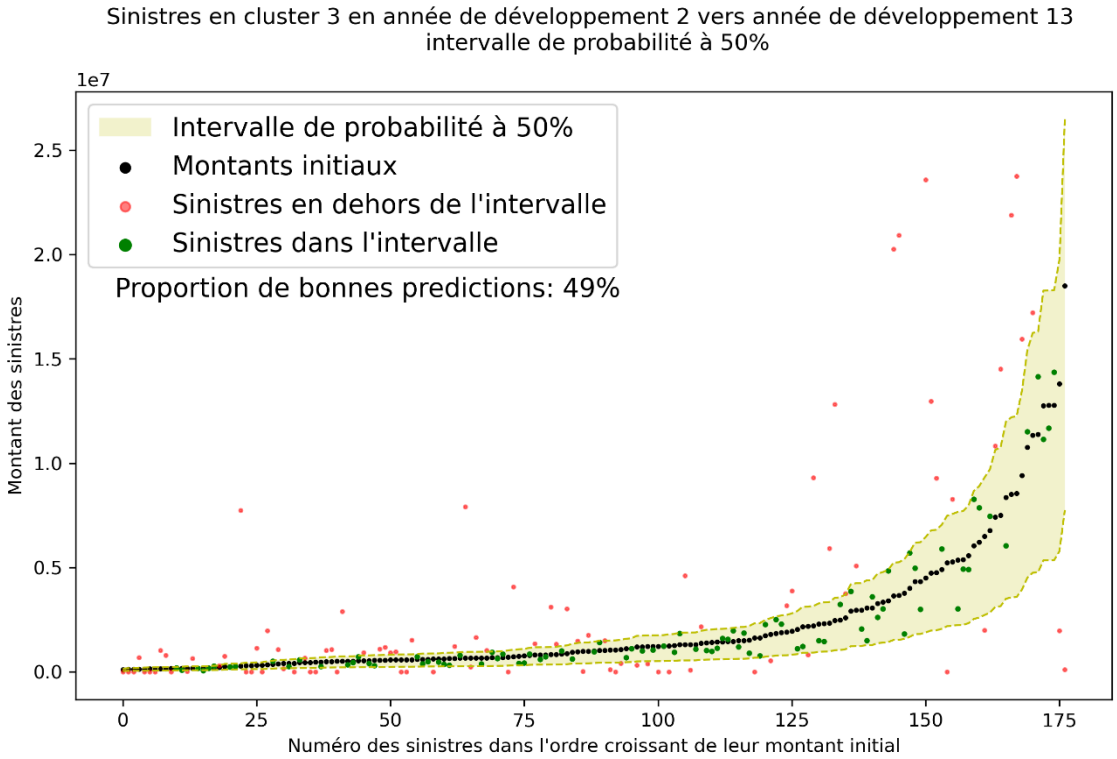


*Figure 36 - Exemple de cohérence (intervalle de probabilité à 50%)*

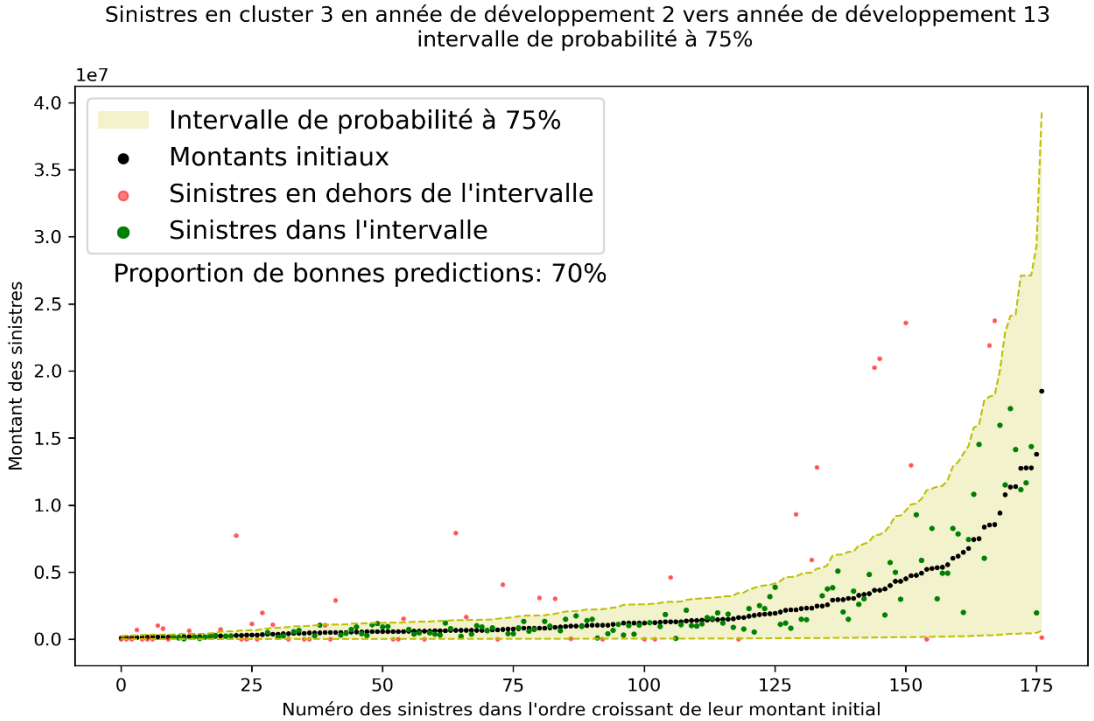


*Figure 37 - Exemple de cohérence (intervalle de probabilité à 75%)*

A travers ces trois exemples, on peut voir que la méthode donne de très bons résultats sur cet exemple. Bien que les chiffres des bonnes prédictions soient légèrement inférieurs à la valeur théorique, le modèle semble être satisfaisant dans ce cas.

Pour analyser de manière plus globale les résultats des modèles, les calculs de la proportion d'observations dans les intervalles pour toutes les combinaisons de clusters, d'années de développement d'origine et d'années de développement finales seront effectués.

Voici la boite à moustache (ou *Boxplot*) représentant les résultats :

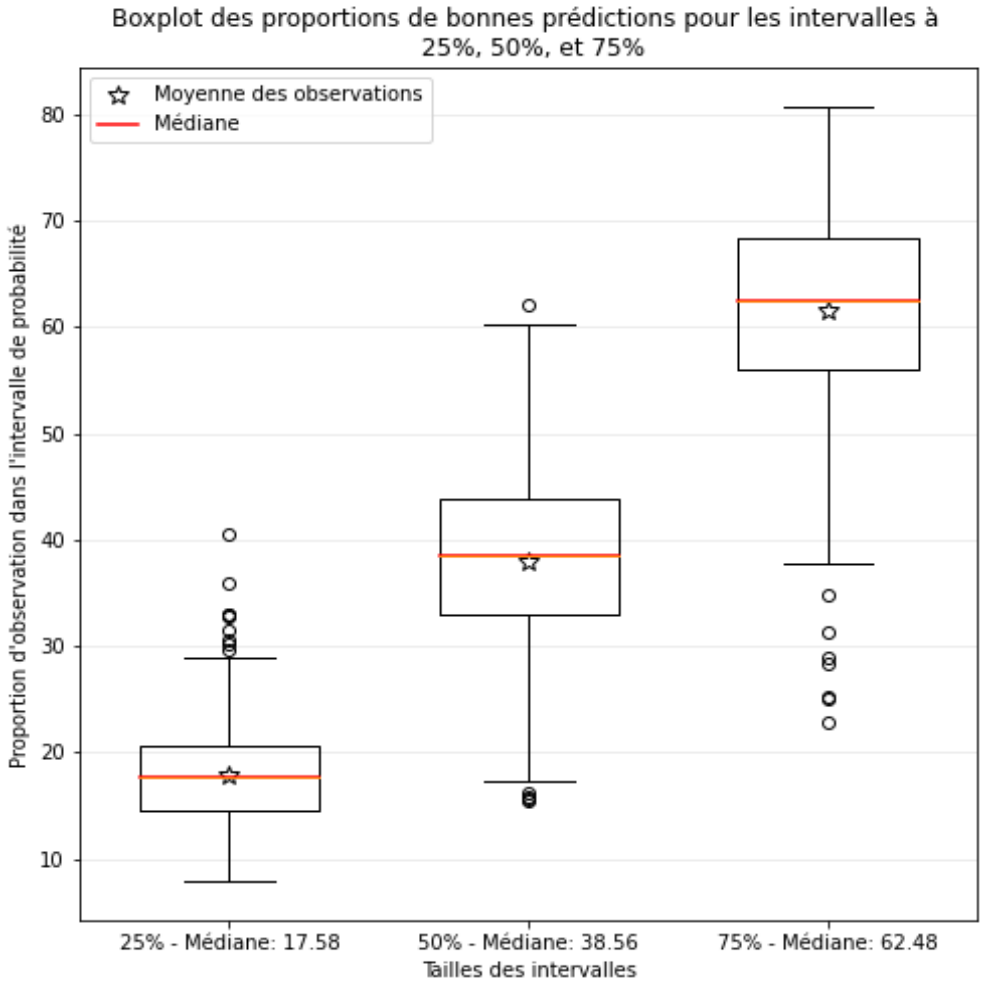


*Figure 38 – Boxplot des proportions de bonne prédiction pour les intervalles à 25%, 50% et 75%*

De manière général, les résultats ne peuvent pas être considérés comme mauvais, mais sont quand même mitigés. On note un écart médian de 30% (respectivement 23% et 17%) pour l'intervalle de probabilité à 25% (respectivement 50% et 75%).

Il y a effectivement un écart substantiel entre l'objectif théorique et les résultats pratique. Cependant ces écarts peuvent être expliqués.

En effet, tout d'abord, la simplification adoptée pour pouvoir calculer les fonctions de répartition des évolutions nous fait forcement perdre de la précision. Cela surtout pour la partie haute de l'intervalle, car les évolutions considérées comme haussières appartiennent à l'intervalle : $]1.02, +\infty[$. Or, la simplification adoptée résume toutes ces observations en un seul point, ce qui n'est

sans doute pas assez précis. Une manière de pallier ce problème serait de définir plus de points de simplification.

Enfin, une deuxième raison à ces résultats mitigés peut être la masse de probabilité en 0 des fonctions de répartition. On remarque qu'un grand nombre de sinistre finissent par évoluer vers un montant proche de 0. Cela créer donc une masse de probabilité en 0 qu'il faudrait traiter pour obtenir de meilleurs résultats.

Le prochain objectif est d'utiliser ces différentes fonctions de répartition et les matrices de transitions entre états calculées précédemment pour calculer une fonction de répartition unique pour chaque cluster de chaque année de développement.

## 4.5 Fonction de répartition des évolutions finales

Maintenant que les fonctions de répartition des évolutions pour une année future donnée sont connues, il faut les rassembler pour obtenir la fonction de répartition finale des évolutions d'un sinistre.

### 4.5.1 Calcul de la fonction de répartition des évolutions finales

Le calcul de la fonction de répartition finale consiste à ce qu'à chaque fois qu'une transition d'une année de développement à une autre est calculée, il y a une chance que le sinistre se clôture et donc il faut recalculer tous les chemins pour ne pas prendre en compte seulement les évolutions en étant d'ouverture mais pour prendre en compte aussi celle en état de clôture.

L'idée est de se dire : il y a une chance pour qu'à chaque année de développement le sinistre se clôture et que donc l'évolution du sinistre soit terminée. Ainsi en utilisant les fonctions de répartition précédemment calculées et en ajoutant ces possibilités de clôture, la fonction de répartition finale est obtenue.

### 4.5.2 Présentation des résultats

Pour étudier la qualité des résultats obtenus, la même méthode précédente sera réutilisée. J'ai choisi de créer des intervalles autour de la moyenne des évolutions. Ensuite, je compare la proportion de sinistres réels présents dans l'intervalle avec la taille de l'intervalle. J'ai choisi de représenter trois intervalles : 25%, 50% et 75%. Cependant un deuxième intervalle sera créé pour les

baisse comprises entre -90% et -100%. Ceci a pour objectif d'améliorer la précisons de l'intervalle, tout en prenant en compte la masse de probabilité en 0 de la fonction de réparation.

Voici des exemples graphiques sur les sinistres du cluster 3 en année de développement 2 :

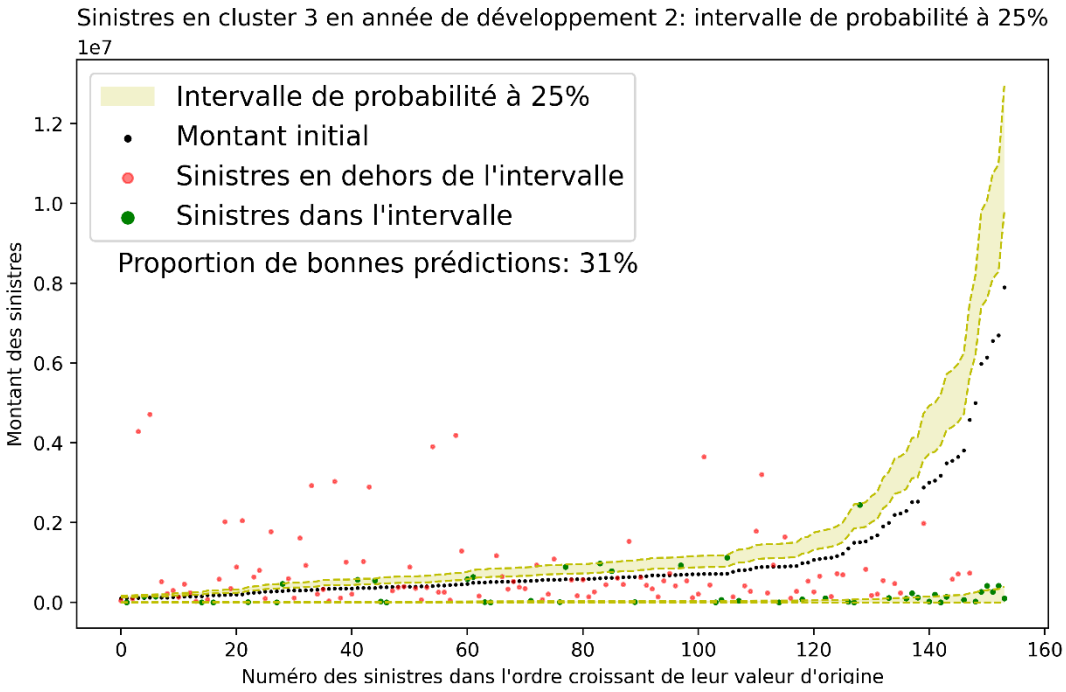


*Figure 39 - cohérence de l'intervalle de probabilité théorique à 25%*

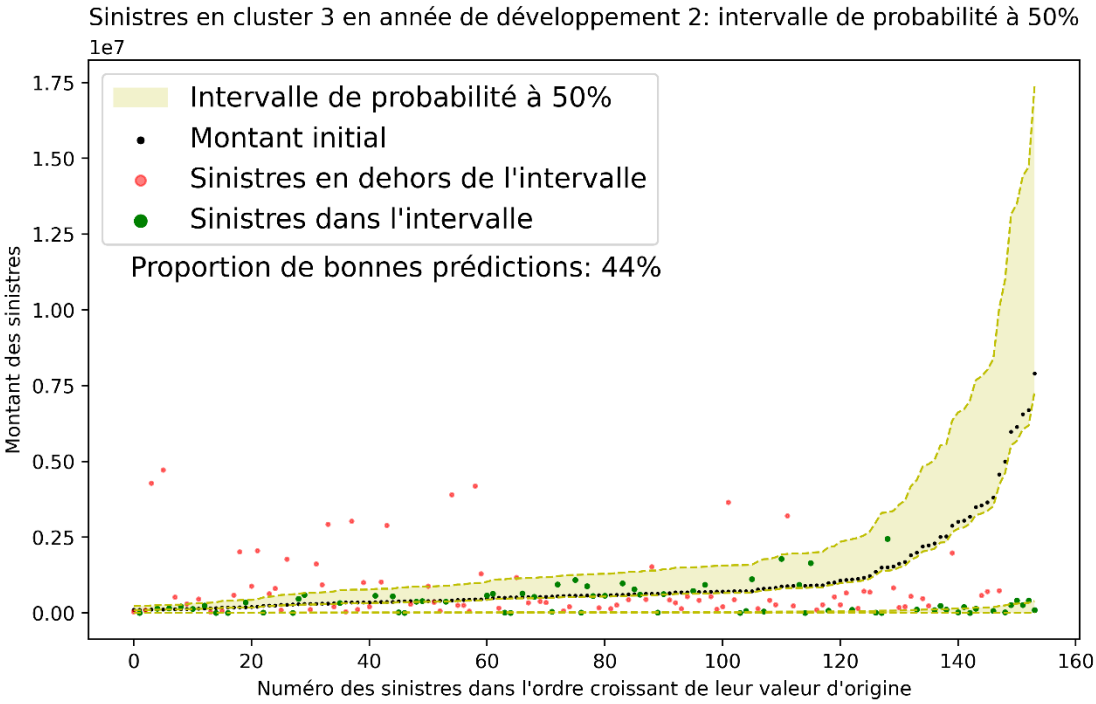


*Figure 40 - cohérence de l'intervalle de probabilité théorique à 50%*

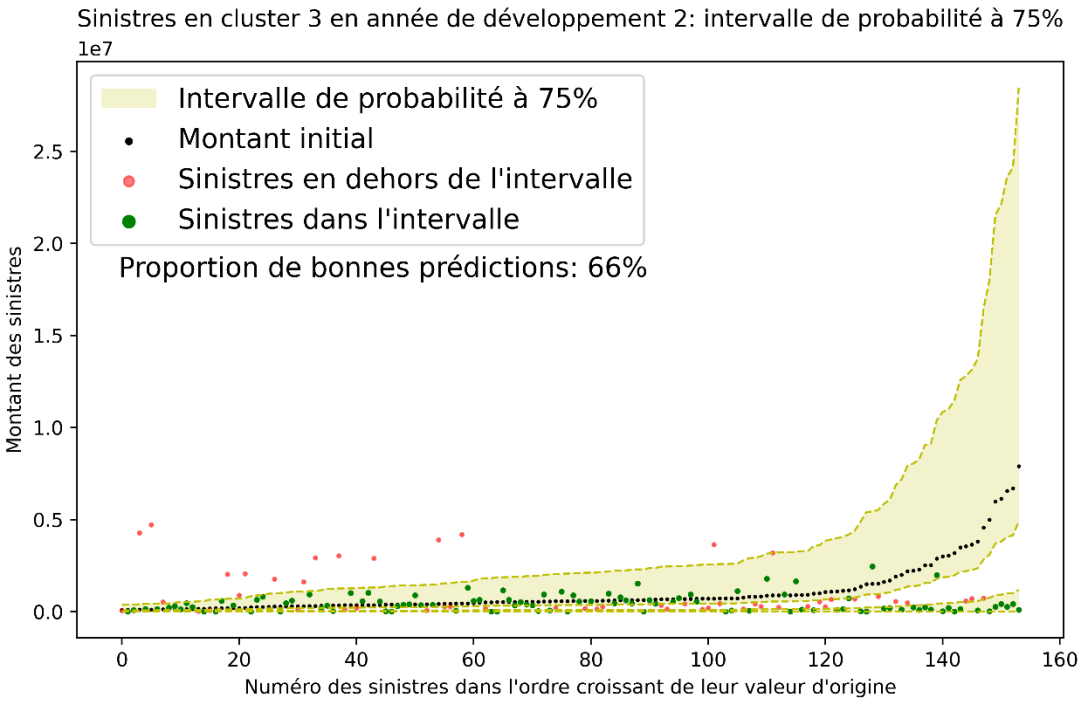


*Figure 41 - cohérence entre l'intervalle de probabilité théorique à 75%*

A travers ces trois exemples, on peut voir que la méthode donne de très bons résultats. Bien que les chiffres des bonnes prédictions soient légèrement inférieurs (sauf dans l'intervalle à 25%) à la valeur théorique, le modèle semble être satisfaisant pour cet exemple. L'ajout d'un deuxième intervalle venant capter les sinistres retombant à 10% ou moins de leur montant initial permet d'améliorer grandement la prédiction. Ainsi les intervalles prédits seront des doubles intervalles tels que leur somme fasse la grandeur de l'intervalle souhaité.

Pour analyser de manière plus globale les résultats du modèle, j'ai réalisé les calculs de la proportion d'observations dans les intervalles pour toutes les combinaisons de clusters, d'années de développement d'origine et d'années de développement finales.

Voici la boite à moustache (ou *Boxplot*) représentant les résultats :

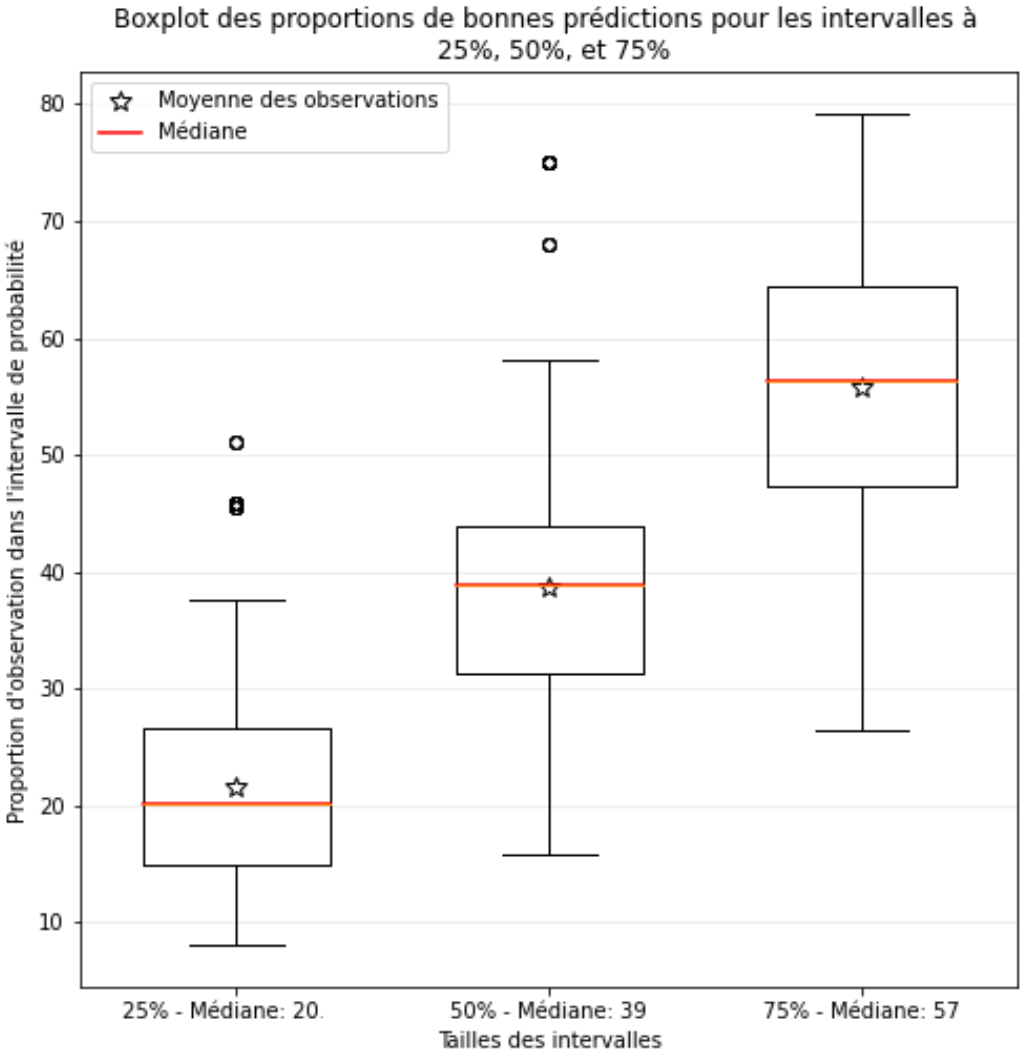


*Figure 42 - Boxplot des proportions de bonne prédiction pour les intervalles à 25%, 50% et 75%*

Les résultats du modèles finals sont assez semblables à ceux des modèles précédents. Il y a un écart de 20% par rapport à l'objectif (respectivement 22% et 24%) pour l'intervalle de probabilité à 25% (respectivement 50% et 75%). On peut considérer que ces estimations sont de bonnes estimations.

La prise en compte de de la masse de proba en 0 a notamment permis d'augmenter le taux de bonne prédiction pour l'intervalles à 25%. Cela fait que l'estimation du montant ultime d'un sinistre donnée par le modèle sera composée de deux parties : Une partie ou le sinistre voit son montant chuter et une autre partie où son montant évolue autour de la moyenne des évolutions.

Cet écart peut en partie s'expliquer par les simplifications effectuées pour rendre le modèle calculable. En effet, peut-être qu'avec plus de complexité pour décrire les hausses et les baisse du montant estimé, le modèle serait plus précis.

# Chapitre 5 : Résultats

L'objectif du chapitre est d'analyser les résultats obtenus par l'application de la méthode présentée dans ce mémoire. Ce chapitre permettra de définir les qualités, les défauts et surtout dans quelle mesure la méthode proposée répond aux attentes fixées au début du mémoire.

Dans un premier temps, les résultats seront comparés à ceux de la méthode classique. Cette comparaison permettra de comprendre les avancés permises par la méthode. Enfin, les quelques défauts relatifs à l'utilisation de cette méthode seront mis en exergues.

Dans un second temps, les résultats seront comparés avec les objectifs fixés au début de ce mémoire. Tout d'abord en étudiant la précision des intervalles prédits. Ensuite, en étudiant l'impact des résultats sur l'estimation de la fréquence et de la sévérité.

Dans un troisième temps, une critique de la méthode sera dressée. Tout d'abord, il conviendra de réfléchir à l'utilisation des résultats par les techniciens lors de la tarification. Puis, j'aborderai la question de l'obsolescence des calculs. Ces différentes questions sont indispensables pour donner des pistes d'amélioration.

## 5.1 Comparaison des résultats avec la méthode classique

### 5.1.1 Améliorations notables

La première chose à relever c'est que cette nouvelle méthode permet de donner beaucoup plus d'informations sur le développement possible en montant ultime du sinistre. Il est possible de dégager plusieurs degrés de lecture de ces informations :

- Il est toujours possible de ne s'intéresser qu'à l'évolution moyenne des sinistres
- Il est aussi possible de s'intéresser aux intervalles de probabilités des montants ultimes de sinistres.
- Enfin, il est aussi possible d'utiliser la totalité ou une partie de la fonction de répartition estimée pour garder le maximum d'informations données par la méthode. De cette manière, on obtient un ensemble de points représentant les futures situations en ultime possible, avec leur probabilité de survenance.

Il est évident que cette nouvelle méthode est une amélioration totale de la méthode précédente.

Tout d'abord, l'information possédée pour chaque sinistre est améliorée grâce à la création des métriques (3.1.2 Définition des nouvelles variables - Métriques). Cette amélioration permet de classer les sinistres d'une manière beaucoup plus précise que de seulement admettre que les sinistres d'une même cédante ont un comportement similaire. A partir de maintenant, l'entièreté des sinistres Français et Belge disponibles dans la base de données de QBE Re sont pris en compte dans le développement à l'ultime.

Enfin, la manière dont les sinistres sont développés est plus complexe. En effet, la méthode déterministe est remplacée par une méthode stochastique. Cette méthode stochastique est une véritable amélioration pour développer les sinistres.

### 5.1.2 Défauts relatifs

Comme nous l'avons vu, la méthode proposée dans ce mémoire peut être considérée comme meilleure que la méthode précédente. Cependant, il y a quelques points qui peuvent être considérés comme des défauts.

Le premier point problématique est l'aspect boite noire et complexe de la méthode. En effet, la méthode ne permet pas au technicien de voir tous les rouages du développement de sinistre lors de la tarification. Contrairement à la méthode de Chain Ladder, le technicien ne peut ni vérifier les calculs effectués ni ajuster les coefficients selon son jugement d'expert.

Le deuxième point problématique est que l'utilisation de l'estimation d'ultime donnée par cette nouvelle méthode n'est pas encore gérée par une méthode de tarification actuarielle classique. Pour l'instant la méthode de tarification permettant de prendre en compte ces nouvelles estimations du montant ultime de sinistre est celle développée le mémoire : *Study of the Body and the Tail* (Biber, 2021) [1].

Ces défauts ne sont pas rédhibitoires à l'utilisation de la nouvelle méthode.

## 5.2 Comparaison des résultats avec les objectifs fixés

Maintenant que nous avons vu que cette méthode est, de manière générale, une amélioration de l'ancienne méthode, il nous faut définir dans quelle mesure cette nouvelle méthode répond aux objectifs formulés tout au long de ce mémoire.

### 5.2.1 Précisions des résultats et mise sous forme d'intervalles

Le premier objectif de ce mémoire était d'améliorer l'estimation du montant ultime des sinistres. On peut dire que c'est le cas. Seulement, dans quelle mesure ces résultats sont-ils précis ? Cette précision peut-elle être considérée comme satisfaisante pour une utilisation en tarification ?

Comme nous l'avons vu précédemment, les résultats des intervalles de probabilités obtenus ne sont pas totalement satisfaisants. En effet, lorsque l'on étudie la *Figure 42 - Boxplot des proportions de bonne prédiction pour les intervalles à 25%, 50% et 75%*, on remarque une différence moyenne de 20 à 25% de la proportion d'observations comprises dans l'intervalle de probabilité par rapport la proportion attendue d'observation dans l'intervalle.

Pour rappel, les intervalles de probabilité de taille $I$ créés sont constitués de deux parties : une première partie comprenant les évolutions entre $[0; 0{,}10]$ et une deuxième partie comprenant les évolutions autour de la moyenne des évolutions $[M_-, M_+]$, telle que :

$$P(0{,}10) + P(M_+) - P(M_-) = I$$

La distribution des observations autour de l'espérance d'évolution est, elle, bien respectée, comme le montre ce graphique :

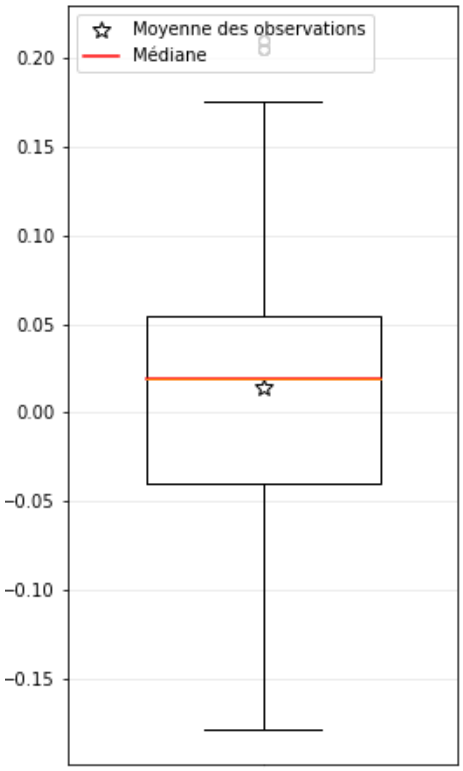


*Figure 43 - Boxplot des écarts absolus entre la proportion théorique de sinistres supérieur à l'espérance de l'évolution et la proportion observée*

Ce graphique montre bien que la probabilité que le montant ultime d'un sinistre soit inférieur ou supérieur à l'espérance de l'évolution est bien estimée. En effet, l'écart entre la proportion théorique et celle observée est inférieur à 1% en moyenne.

Cela vient alors compléter le fait qu'il y a 20% à 25% en moins d'observations dans les intervalles théoriques. Cela signifie qu'en réalité les intervalles proposés par ma méthode sont légèrement plus petit que ce qu'il devrait, que ce soit pour les évolutions à la hausse ou à la baisse.

Cette imprécision que j'ai relevée précédemment est donc d'une importance relative. En effet, les intervalles ont leur point central bien calibré mais leur taille est sans doute trop petite. Cela peut notamment s'expliquer par la simplification adoptée en vue de pouvoir réussir à calculer les fonctions de répartitions. De manière générale, les évolutions sont plus probables au plus elles sont proches de 1. Cela fait que les évolutions extrêmes (à la baisse ou à la hausse), qui sont peu probables, sont diluées par les faibles évolutions (plus probables) par la simplification adoptée

### 5.2.2 Fréquence

Cette nouvelle méthode a permis d'obtenir une plus grande précision sur l'estimations de la fréquence. La fréquence est souvent modélisée par une loi de poisson. Ainsi, le nouvel estimateur du paramètre $\lambda$ de la loi de poisson sur des sinistres développés avec la nouvelle méthode présentée dans ce mémoire peut être définie de cette manière :

On pose $X_i = (x_1, \dots, x_k)$, le $i^{ème}$ $sinistre\ développé\ en\ k\ points\ de\ situation\ ultime$ et D le montant de la priorité.

On a alors, en détournant l'estimateur par maximum de vraisemblance de la loi de poisson :

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} P(X_i = x_j) \times \mathbb{1}_{x_j > D}$$

De cette manière, l'estimateur $\hat{\lambda}$ obtenu est plus fin que celui obtenu lorsque les montants des sinistres sont développés de manière déterministe. De plus, bien que les intervalles de probabilité soient plus petits que ce qu'ils devraient, cela est tout de même une amélioration pour estimer la fréquence.

### 5.2.3 Sévérité

Avant l'écriture de ce mémoire, aucune méthode ne permettait d'estimer les paramètres d'une fonction de répartition à partir des montants ultimes probables.

C'est pour cela que l'écriture de ce mémoire c'est fait en collaboration avec BIBER.L. qui a écrit son mémoire *« Study of the Body and the Tail »* (Biber, 2021) [1]. Dans ce mémoire, une nouvelle méthode de modélisation de la sévérité des sinistres de réassurance est développée. Cette partie sera une présentation des résultats obtenus dans le mémoire de BIBER.L grâce à cette nouvelle manière de développer les sinistres.

L'auteur utilise des sinistres développés par la méthode que présentée dans ce mémoire, en vue de modéliser la fonction de répartition de sévérité associée. L'intervalle de probabilité retenu est celui à 75%. En effet cet intervalle permet aux sinistres de prendre une grande amplitude de valeur, tout en restant dans des valeurs probables.

Voici une représentation graphique (Biber, 2021) [1] des sinistres développés avec la méthode de ce mémoire, comparés à un développement déterministe :

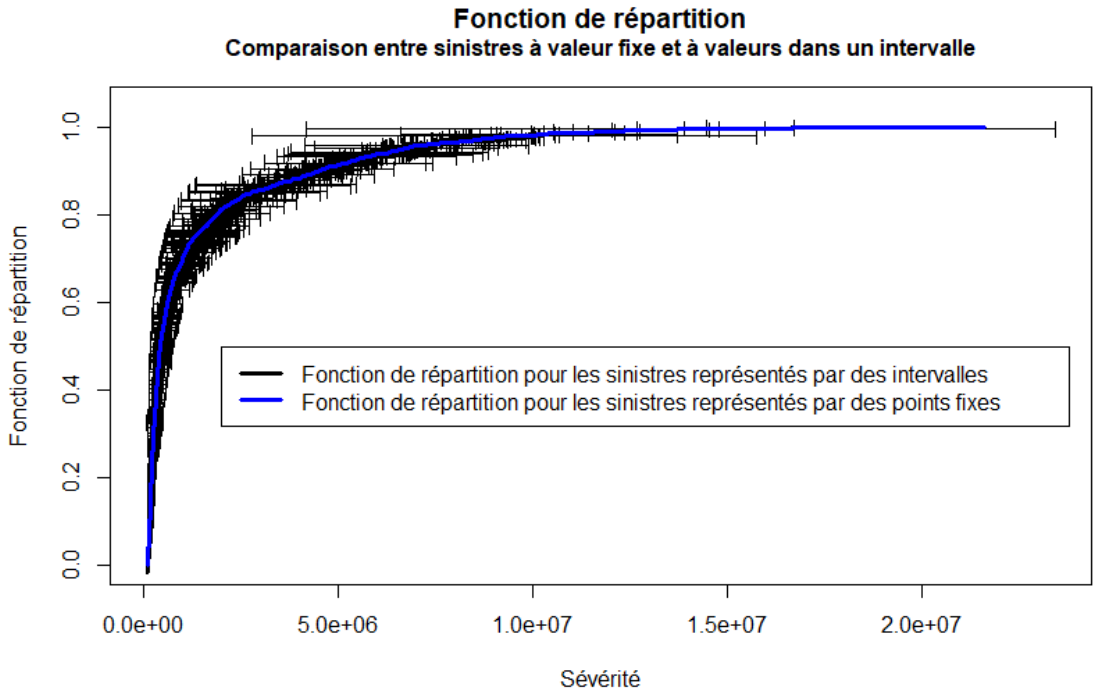


*Figure 44 - BIBER.L - Représentation graphique de la fonction de répartition de la sévérité*

On peut alors modéliser la *Tail* de la fonction de répartition (ici BIBER.L utilise une loi exponentielle). Voilà la représentation graphique de la modélisation proposée par BIBER.L. :

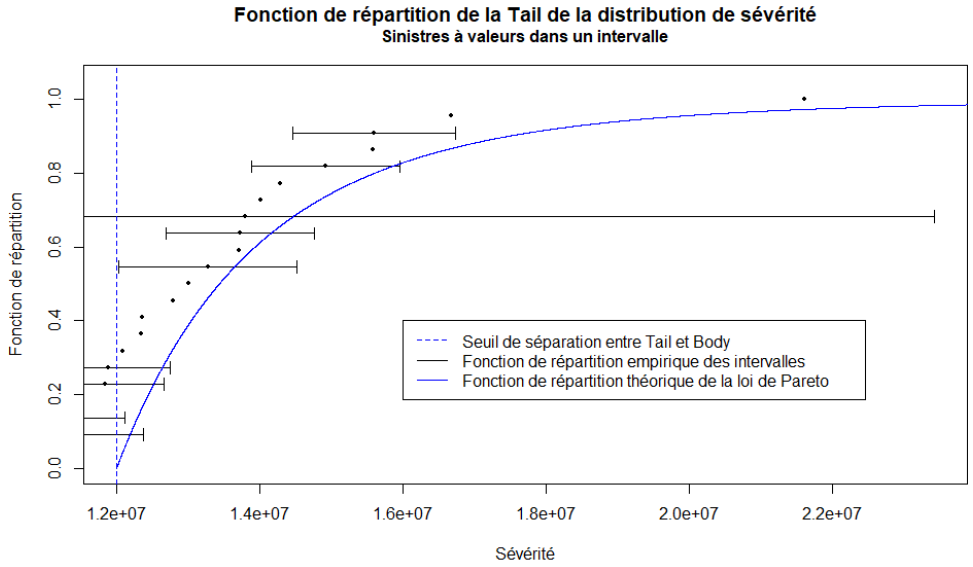


*Figure 45 - BIBER.L – Fonction de répartition de sévérité de la Tail modélisée*

Dans ce dernier graphique, il est intéressant de regarder les quatre premiers sinistres. Ces 4 premiers sinistres n'auraient pas dépassé le seuil de la *Tail* s'ils avaient été développés de manière déterministe. Cette nouvelle manière de développer les sinistres permet de prendre en compte la possibilité de ses sinistres de dépasser le seuil de séparation entre le *Body* et la *Tail*.

Pour modéliser le *Body* de la sévérité, on utilise des mélanges d'Erlang (Biber, 2021) [1]. Dans le cas de l'exemple précédent, un mélange de huit Erlang a été choisi. Ce mélange est alors représenté de la manière suivante :

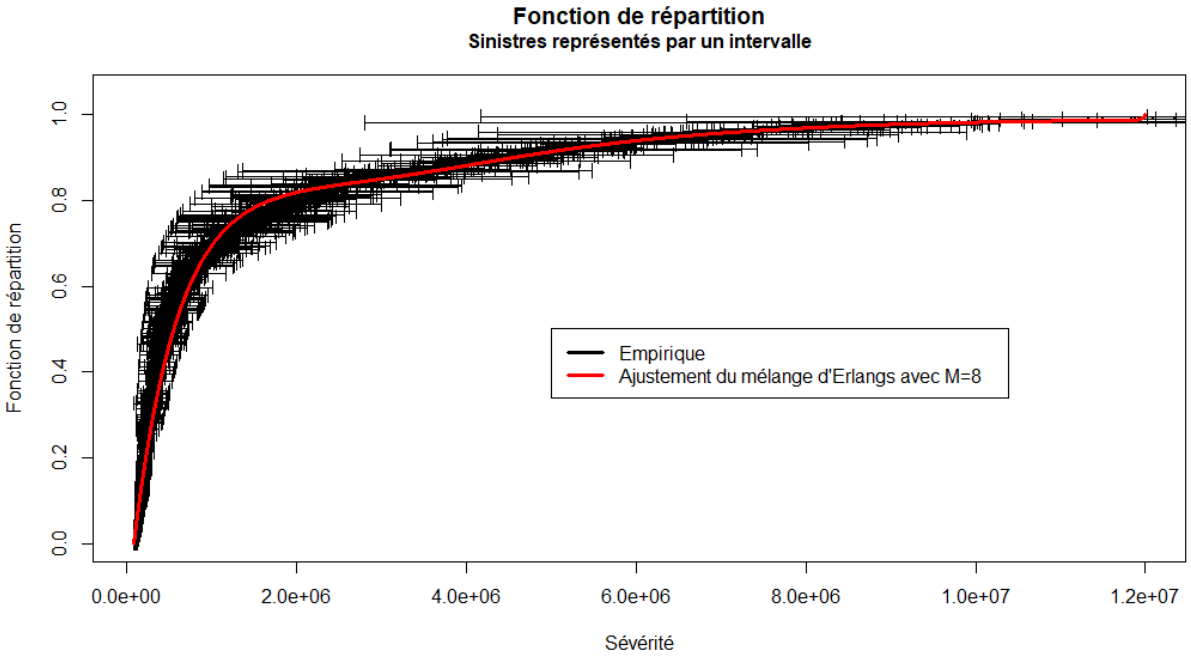


*Figure 46 - BIBER.L - Modélisation du Body via un mélange d'Erlang*

Finalement, cette nouvelle méthode de développement des sinistres remplie bien son rôle : elle donne beaucoup plus d'information sur les développements possibles et permet une modélisation de la fréquence et de la sévérité plus fine qu'auparavant. Notamment, en utilisant un mélange d'Erlang pour modéliser le *Body* de la fonction de répartition.

## 5.3 Critique de la méthode

### 5.3.1 Interprétabilité des résultats et marge de manœuvre pour les techniciens

Le principal problème de cette méthode de développement de sinistre est sa complexité. En effet, contrairement à Chain Ladder, la méthode repose sur plusieurs étapes qui ne sont pas facilement visualisables.

Dans un premier temps, il faut calculer les différentes variables venant mesurer le comportement du sinistre. Cette étape est déjà compliquée à visualiser pour un technicien à cause du grand nombre (13) de variables.

Dans un second temps, le clustering est effectué. Le clustering est pensé pour être réalisé une seule fois avec tous les sinistres. Par la suite, il peut être réutilisé pour chaque nouvelle tarification. Cependant ce clustering n'est pas aisément visualisable, toujours à cause du nombre de variables.

Dans un troisième temps, les fonctions de répartitions sont calculées. A cette étape, encore, l'actuaire en charge de la tarification ne peut pas voir les rouages des calculs effectués.

Ces étapes rendent la méthode de développement des sinistres complexe et opaque. C'est un problème car, en réassurance particulièrement, l'actuaire doit pouvoir comprendre le résultat de la méthode pour y confronter son expertise. Par ailleurs, il n'est pas rare qu'un actuaire modifie, toujours selon son expertise, les différents coefficients obtenus via Chain Ladder. Ceci en vue d'adapter les résultats aux connaissances acquises lors des études précédemment effectuées.

### 5.3.2 Obsolescence des calculs

Le second problème de cette méthode est l'obsolescence des calculs. Les différentes étapes de la méthode doivent être recalculées tous les ans pour mettre à jour la base de données et rendre la méthode de plus en plus performante.

Cela est un désavantage par rapport à la méthode de Chain Ladder, qui ne demande aucun calcul préalable sur les données. Cela rajoute donc une certaine quantité de travail pour que la méthode soit mise à jour chaque année. Il faut compter entre 1 et 2 heures de calculs pour faire tourner les différents algorithmes. Mais il faut aussi prendre en compte le temps d'implémentation des nouveaux résultats dans les différents outils de tarifications. Cela rend la méthode beaucoup plus contraignante dans son suivie que la méthode de Chain Ladder.

### 5.3.3 Sensibilités

En vue de pouvoir utiliser cette méthode d'estimation de la valeur des ultimes, il est important de pouvoir quantifier la sensibilité des résultats obtenus aux différents paramètres. Idéalement, il faudrait que cette méthode produise des résultats stables dans le temps pour que la tarification qui découle des résultats n'évolue pas de manière trop marquée d'un renouvellement à l'autre.

Dans un premier temps, la stabilité des intervalles de probabilité selon le nombre d'observation est étudiée. Tout d'abord à travers un exemple : l'intervalles de probabilité à 75% sur les sinistres du cluster 3 en année de développement 2. Dans cet exemple, un certain de nombre d'observations (10%, 25% et 50%) ont été retirées de la base de données. Ensuite, la méthode est réappliquée à cette nouvelle base de données pour obtenir les nouveaux intervalles de probabilité à 75%. Voici les résultats obtenus sous forme de graphique :

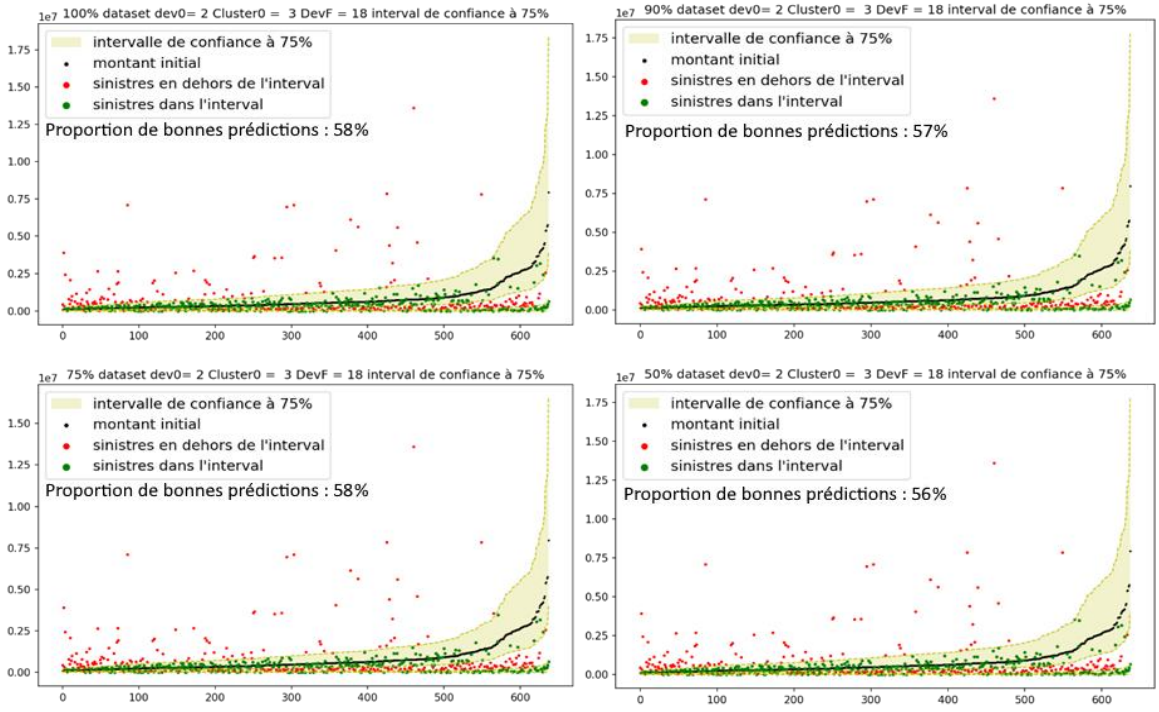


*Figure 47 - Comparaison des intervalles de probabilité selon la taille de la base de données*

| Taille de la base de données | Variation de la taille de l'intervalle par rapport à l'intervalle avec 100 des données |
|---|---|
| 90% de la base de données | -3% |
| 75% de la base de données | -5% |
| 50% de la base de données | -7% |

On peut remarquer que la différence graphique n'est pas flagrante. L'intervalle proposée par le modèle semble stable, les prédictions du modèle sont assez peu impactées par la variation de la taille de la base de données. Il en est de même pour la taille des intervalles, ceux-ci sont seulement en légère baissent.

Cette stabilité s'explique par la manière dont les données sont utilisées : la simplification décrite en 4.4.2 effectue une moyenne des évolutions en fonction de leur appartenance à un intervalle. Ainsi le fait qu'il y ai un grand nombre d'observations dans les clusters importants rend les résultats stables. Si certaines observations atypiques disparaissent de petits clusters alors l'impact sera plus important mais sur un nombre restreint de sinistre.

Pour appréhender au mieux la sensibilité de la méthode proposée les calculs de la méthode ont été relancés entièrement trois-cent fois en enlevant aléatoirement cent fois 5%, 25% et 50% de la base de données. Les variations de deux indicateurs ont été étudiées :

- Le pourcentage de sinistres compris dans les intervalles calculés.
- La répartition des sinistres autour de la moyenne des évolutions.

Cela permettra de comprendre si les intervalles sont centrés autour de l'évolution moyenne.
Voici les résultats des prédictions sur les bases de données réduites de 5%, 25% et, 50% :

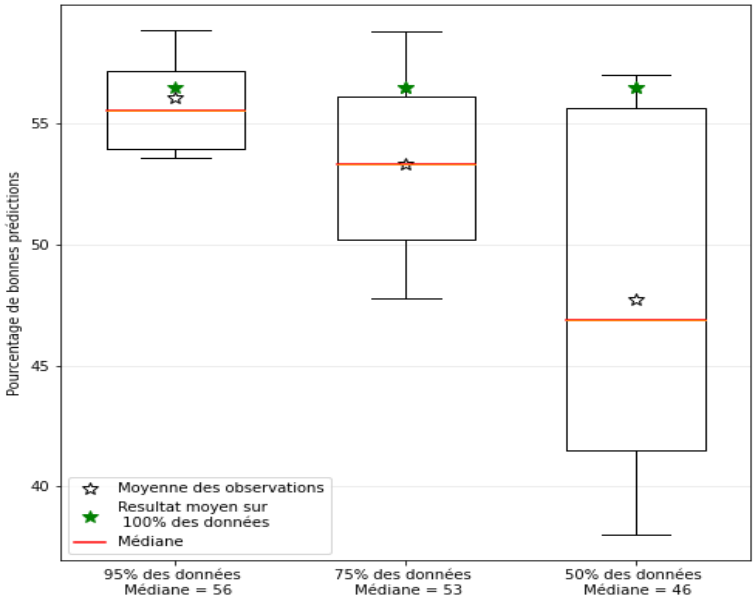


*Figure 48 - Répartitions des résultats des prédictions selon la taille de la base de données*

Ces tests de sensibilité des résultats montrent des résultats très intéressants. La première observation à faire est que la baisse de la précision liée à la baisse du nombre d'observations dans la base de données est contenue. En effet, les médianes de bonnes prédictions sont respectivement de 56, 53 et 46 pourcents, contre 57% pour la base de données complète. Soient des baisses respectives de 2, 7 et 19 pourcents.

Une deuxième observation à faire est que l'amplitude de la variabilité des résultats augmente en fonction de la quantité de données supprimées. Pour la base de données à 5%, la variabilité est faible. Pour la base de données à 75%, la variabilité du résultat n'est pas trop importante, sauf dans quelques cas extrêmes. En revanche la variabilité des résultats pour la base de données à 50% est très importante.

La stabilité de la base données à 95% est extrêmement importante pour considérer la viabilité de la méthode. Puisque le modèle a été construit à partir de 20 années d'observation de sinistres, l'ajout d'un exercice supplémentaire augmentera la base de données de 5%.

Voici les répartition moyenne des sinistres autours de la moyenne d'évolution sur les bases de données réduites de 5%, 25% et, 50% :

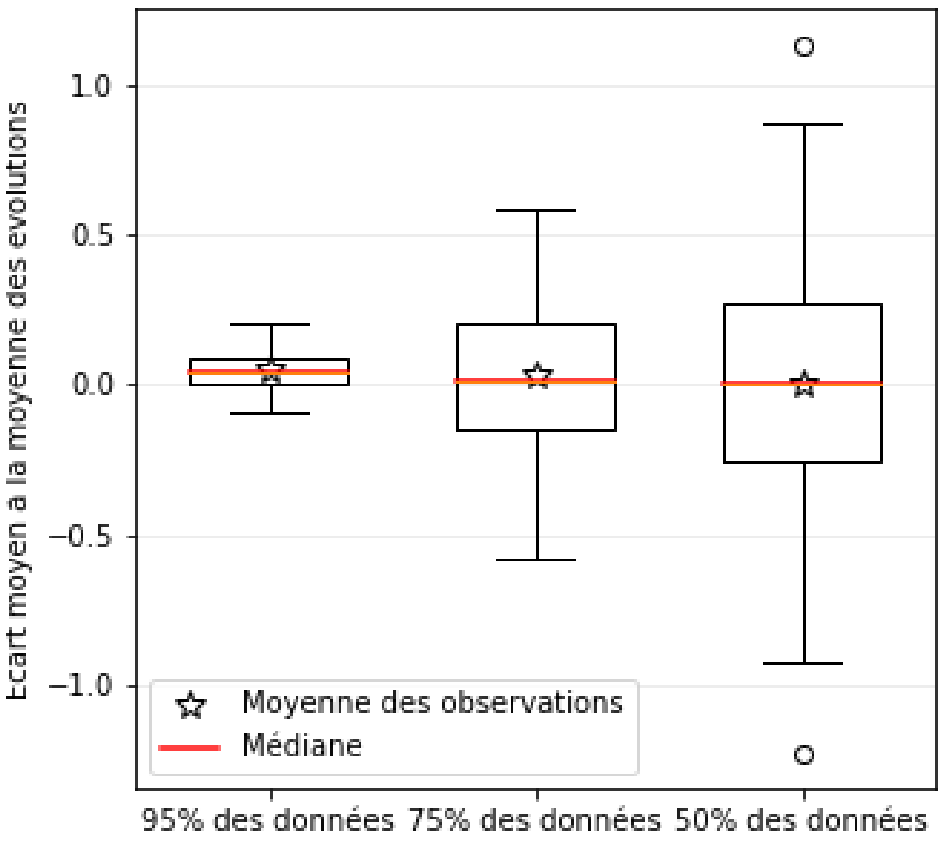


*Figure 49 - répartition moyenne des sinistres autours de la moyenne d'évolution sur les bases de données réduites*

On peut remarquer que les écarts moyens à la moyenne d'évolution sont, là aussi, satisfaisante. Les évolutions calculées sont donc bien centrées sur l'évolution moyenne. On peut donc conclure que si les intervalles calculés n'atteignent pas la précision visée (75% sur ces exemples), ils sont juste trop petits par rapport à l'évolution réelle des sinistres.

On a pu voir à travers ces deux tests de sensibilités que la méthode proposée dans ce mémoire est stable dans le temps. En particulier pour la base de données à 95%, représentant la prédiction du modèle sur les sinistres de l'exercice suivant.

Cependant, il convient d'énoncer la limite des tests de sensibilités effectués. Le modèle a seulement été recalculé cent fois par base de données réduite. Ce choix a été fait par soucis de temps de calcul. En effet, il faut compter 1 heure en moyenne pour réaliser le modèle. Il était donc compliqué de faire un test de sensibilité de plus grande envergure, alors que cela aurait été intéressant.

# Conclusion

Dans ce mémoire, nous avons pu comprendre le fonctionnement général de la tarification en réassurance. Cela nous a permis de comprendre l'importance du développement de sinistre dans le processus de tarification.

La méthode de Chain Ladder a pu être étudiée. C'est la méthode de développement classique des sinistres en réassurance. Cependant, elle souffre de défauts qui empêchent de penser que cette méthode soit optimale. Les développements proposés sont déterministes, cela ne permet pas d'obtenir d'informations sur les développements possibles des sinistres. De plus, nous avons aussi vu que le modèle de Mack, version stochastique de la méthode de Chain Ladder, n'est pas non plus satisfaisant. Les évolutions de montants proposées par ce modèle ne sont pas en accord avec celles observées.

Le fait que ces méthodes de développement de sinistres ne prennent en compte que l'historique du coût des sinistres nous est apparu comme une de leurs limites. C'est pour cela qu'une amélioration de la qualité des données a été mise en place. Elle consiste en la création de 13 nouvelles variables, nommées *Métriques*, se basant sur l'historique des montants estimés et des paiements liés à un sinistre.

Grâce à ces nouvelles variables, une classification non-supervisée des données a été réalisée. L'algorithme utilisé est celui de la classification ascendante hiérarchique et la méthode de Ward, ceci de manière à trouver les différents grands schémas de développement de sinistres existants et de les étiqueter. Ces étiquettes permettent de calculer les probabilités de transition entre les clusters. De cette manière, il devient possible de définir les différents chemins que peut prendre un sinistre (combinaison des différents clusters aux différentes années de développement), et de trouver la probabilité que ces chemins soient empruntés.

Enfin, nous avons calculé les évolutions historiques pour chaque transition à chaque année de développement. Ces évolutions historiques ont été simplifiées pour rendre possible de combiner chaque possibilité d'évolution entre elles. De cette manière, pour chaque chemin d'évolution possible, nous connaissons sa probabilité de survenance et l'évolution qu'il entraîne sur le sinistre.

En combinant toutes ces possibilités, il devient possible de calculer les fonctions de répartition des évolutions finales pour chaque cluster de chaque année de développement. Grâce à cette estimation, des intervalles de probabilité peuvent être estimés. Mais il est aussi possible d'utiliser directement chacune des possibilités d'évolution avec sa probabilité associée.

Finalement nous avons pu conclure que les estimations des montants ultimes sous forme d'intervalles de probabilité sont satisfaisantes. Elles sont légèrement moins précises que ce qui était attendu, mais elles restent assez précises pour être considérées comme viables. De plus la méthode peut être considérée comme stable. En adoptant des simplifications moins grossières, il serait sans aucun doute possible d'améliorer grandement la précision de l'estimation.

Cependant, la complexité de la méthode rend celle-ci opaque. C'est pour cette raison qu'il pourrait être intéressant de proposer aux actuaires en charge de la tarification lors du renouvellement de développer les sinistres avec la méthode présentée dans ce mémoire. Ainsi nous pourrions avoir un retour pratique sur cette méthode. Il serait aussi intéressant de mener des tests de sensibilité d'une plus grande envergure pour confirmer la stabilité.

Enfin, ces estimations de montants ultimes des sinistres ont pu être testées dans le mémoire de BIBER.L. Il en ressort que ces estimations sont suffisamment satisfaisantes pour être utilisées dans une nouvelle méthode de tarification que l'auteur a développée.

Cette nouvelle méthode d'estimation des montants ultimes des sinistres répond donc bien aux attentes fixées au préalable de ce mémoire. Le risque d'évolution du montant des sinistres est bien plus détaillé, et peut donc être bien mieux compris qu'auparavant. De plus, grâce à la nouvelle méthode de tarification proposée par BIBER.L, il est possible d'affirmer que cette nouvelle connaissance de l'évolution des sinistres servira en pratique à proposer une tarification encore plus adaptée au risque.

# Bibliographie

[1] Biber, L. (2021). *Study of the Body and the Tail.* QBE Re.

[2] Busson, E. (2012). *Evaluation du risque de provisionnement à 1 an : Adaptation de la méthode de Merz & Wutrich à des cas non standards.* DUAS - Université de Strasbourg.

[3] Mack, T. (1993). Distribution-free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin 23*.

[4] Mahon, J. B. (2005). *Transition Matrix Theory and Individual Claim Loss Development.* Casualty Actuarial Society

[5] Martin Ester, H.-P. K. (1996). *A Density-Based Algorithm for Discovering Clusters.* Institute for Computer Science, University of Munich.

[6] Moffat, M. (2018, 03 18). *Definition and Example of a Markov Transition Matrix*. Retrieved from thoughtCo: https://www.thoughtco.com/markov-transition-matrix-definition-1148029

[7] scikit-learn.org. (2020, 06 17). *Demonstration of k-means assumptions*. (scikit-learn) Retrieved Jun 17, 2020, from scikit-learn.org: https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_assumptions.html

[8] Ward, J. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*.

# Annexes

## Table des figures