Institut des ACTUAIRES

ISUP

Mémoire présenté le :
pour l'obtention du diplôme
de Statisticien Mention Actuariat
et l'admission à l'Institut des Actuaires

Par : Madame / Monsieur Latif Rouamba

Titre du mémoire : Estimation des sinistres tardifs du risque arrêt de travail

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

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

Membres présents du jury de la filière :

Signature :

Entreprise : Malakoff Humanis

Nom : Jing WANG

Signature :

Directeur de mémoire en entreprise

Nom : Anouar Pachiri

Signature :

Membres présents du jury de l'Institut des Actuaires :

Signature :

Invité :

Nom : Anouar Pachiri / Karim Firad

Signature :

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

Signature du responsable entreprise :

Signature du candidat :

Résumé

Les assureurs dans le but de maintenir leurs activités, se doivent de bien d'estimer les provisions pour faire face à leurs engagements futurs. Pour évaluer ces réserves ou provisions, ils utilisent des méthodes qui agrègent un nombre de données. Elles sont robustes et appréciées du fait de leur simplicité d'interprétation et d'utilisation. Bien que ces méthodes soient assez efficaces, nous nous retrouvons souvent avec une volatilité des provisions d'une année à l'autre.

Ces estimations se doivent de plus en plus précises et fiables notamment avec les nouvelles normes assurantielles et comptables comme Solvabilité 2 et IFRS17. Le montant à provisionner doit être estimer en vision $\ll$ Best Estimate$\gg$ permettant de couvrir aussi bien les sinistres déjà connus ou RBNS et les sinistres inconnus ou IBNR.

Avec le développement des systèmes de gestions des assureurs, de grandes quantités de données sont recueillies qui renseignent un maximum d'informations qui caractérisent les sinistres. Les assureurs se doivent de prendre en compte ces informations caractérisques pour mieux évaluer leurs charges. Le machine learning se veut une solution efficiente et adaptée à l'exploitation de ces données.

Ce mémoire a pour objet de réfléchir sur le provisionnement des risques Arrêt de Travail, à savoir l'incapacité et l'invalidité sur un portefeuille de Conventions Collectives Nationales, plus particulièrement l'estimation des sinistres tardifs. Dans un premier temps, nous présenterons le contexte de l'étude puis un rappel des méthodes de provisionnement classiques.

Dans une seconde partie,la méthode de provisionnement individuel retenue est présentée. Nous expliciterons les hypothèses et les correctifs par rapport à la méthode initiale de Baudry & Robert. Un rappel des notions sur le machine learning et l'algorithme XGBoost est proposé au lecteur.

Nous terminerons en présentant les données à notre disposition, le tout suivi de l'application de différentes méthodes pour déterminer le nombre de sinistres inconnus pour le risque invalidité et la provision invalidité pour le risque incapacité.

Mots-clés : Provisionnement, Arrêt de Travail, Sinistres tardifs, Incurred But Not Reported (IBNR), Reported But Not Settled (RBNS), Chain Ladder, Mack, Schnieper, Machine learning, XGBoost.

Abstract

Insurers need to properly estimate their reserves to meet future liabilities. To evaluate these reserves, they use methods that aggregate data. These methods are quite robust and very popular due to their simplicity of interpretation and use. Although these methods are quite effective, we often end up with quite volatile reserves.

With the new insurance and accounting standards such as Solvency 2 and IFRS17, insurers have a duty to determine their reserves more accurately and reliably. The amount of the reserve must be estimated in a Best Estimate vision, allowing to cover both the already known claims or RBNS and the unknown claims or IBNR.

With the development of insurers' management systems, large quantities of data are collected which provide a maximum of information characterizing claims. Insurers have to take this characteristic information into account when assessing their costs. Machine learning is an efficient and suitable solution for the analysis of all this data.

The purpose of this thesis is to reflect on the reserving of work stoppage guarantees, i.e. incapacity and invalidity on a portfolio of National Collective Agreements, more particularly the estimation of our late claims. First, we will present the context of our study and then a reminder of the classic provisioning methods.

In a second part, the individual reserving method used is presented. We will explain the hypotheses and the corrections compared to the initial method of Baudry & Robert. A reminder of the notions on machine learning and the XGBoost algorithm is proposed to the reader.

We conclude by presenting the data at our disposal, followed by the application of different methods to determine the number of unknown claims for the disability risk and the disability provision for the disability risk.

Keywords : Reserving, Temporary disability, permanent disability, Late claims, Incurred But Not Reported (IBNR), Reported But Not Settled (RBNS), Chain Ladder, Mack, Schnieper, Machine learning, XGBoost.

Note de Synthèse

Introduction

L'activité d'assurance se caractérise par son cycle de production inversé. Les assureurs se doivent de bien estimer leurs provisions pour faire face à leurs engagements futurs. Pour évaluer ces réserves ou provisions ils utilisent des méthodes qui agrègent leurs données. Elles sont assez robustes et très appréciées du fait de leur simplicité d'interprétation et d'utilisation. Bien que ces méthodes soient assez efficaces, nous nous retrouvons souvent avec des provisionnements assez volatils d'une année à l'autre. Ces méthodes agrégées telles que Chain-Ladder ne prennent pas en compte la multitude d'informations sur le sinistre et ne permettent pas de calibrer de façon précise nos provisions, notamment celles allouées aux sinistres tardifs.

Nous avons étudié le risque Arrêt de Travail relatif à des portefeuilles de Conventions Collectives Nationales, précisement ses deux sous-risques : l'Incapacité ou Incapacité Temporaire de travail et l'Iinvalidité ou Incapacité Permanente de Travail. La prise en charge des sinistres Arrêt de Travail est déterminée par décision de la Sécurité Sociale.

Ce mémoire propose une réflexion sur le provisionnement en arrêt de travail et la possibilité de séparation des provisions RBNS (sinistres connues) et de nos IBNR (sinistres inconnus). Les méthodes traditionnelles utilisées dans le calcul des réserves seront challengées par des méthodes de Schnieper et une méthode de provisionnement individuel. Les approches proposées sont inspirées des travaux de Schnieper(1991)[1] et de Baudry & Robert (2017) [2] qui ont la particularité de distinguer les RBNS et les IBNR.

En première partie, nous proposons une présentation du contexte des conventions collectives nationales, du risque arrêt de travail et du provisionnement via les méthodes classiques utilisées.

Dans une deuxième partie, la méthode inspirée des travaux de Baudry & Robert (2017) [2] sera présentée dans le cadre d'un provisionnement ligne à ligne. Elle utilisera est l'algorithme de gradient boosting XGBoost.

Enfin en dernière partie de ce mémoire, nous accorderons un grand soin aux traitements de notre base de données et des sinistres atypiques avant d'appliquer nos différentes méthodes pour chaque sous-risque.

Nous n'avons pas pu aborder l'aspect invalidité en attente dans notre étude.

Contexte

Une convention collective ou accord de branche concerne les salariés d'une même activité sectorielle. Les accords de branche peuvent prévoir des garanties prévoyance et/ou des garanties frais de santé. Ils ne peuvent plus imposer des organismes assureurs mais ils peuvent en revanche en proposer quelques-uns de manière facultative. Cela est une recommandation et non une obligation imposée aux entreprises.

Cette précision d'estimation des provisions est capital pour l'assureur.

En cas de sous-provisionnement, il sous-estime ses charges futurs et risque de ne pas pouvoir faire face à ses engagements futurs.

L'assureur se doit d'apporter des justifications et explications sociaux dans le cas des Conventions Collectives Nationales dans les cas de sous-provisionnement ou de sur-provisionnement.

Dans ces différentes situations, l'assureur donne une mauvaise image de lui aux partenaires sociaux et remet en cause son expertise et ses méthodes de calcul. Il sera amené à faire des probables redressements ou des hausses des taux de cotisations qui ne sont pas dans ce cas nécessaire.

Application au risque Incapacité Temporaire Total

Pour le risque Incapacité, nous considérons les sinistres survenus entre 2009 et 2016 de plusieurs CCN ayant des caractéristiques similaires. Ces sinistres sont clos au 31/12/2021. On comparera alors les estimations de provision réalisées avec les quatre méthodes Chain-Ladder, Mack, Schnieper et provisionnement individuel. L'incapacité est limitée à trois ans dans notre étude.

Nous rappelons que le provisionnement pour passage en invalidité n'est pas présenté dans ce mémoire.

Méthodes agrégées

Dans le cadre de notre étude, nous considérons des sinistres survenus entre 2009 et 2016 en arrêt de travail qui sont clos au 31/12/2021.
Nous nous plaçons au 31/12/2016. Notre objectif sera d'estimer les provisions pour ces sinistres et de faire la comparaison avec les montants de provisions réelles.
Pour éliminer le biais de l'actualisation qui diffère selon les années, nous avons annulé son impact en le ramenant à 0 chaque année.

Période de survenance	Provisions réelles	Provision Chain-Ladder	Provisions Schnieper
2009	0	0	0
2010	0	0	0
2011	24 887	1 115	-206
2012	10 488	1 916	13
2013	17 055	50 342	40 546
2014	947 005	1 228 861	1 002 577
2015	4 744 276	5 417 189	4 449 890
2016	13 138 527	14 894 023	12 624 545

TABLE 1 – Ecart entre la provision réelle et l'estimation Chain-Ladder et Schnieper

Sur le montant de provision total, Chain-Ladder obtient une erreur de plus de 14,4%. La méthode la plus répandue de provisionnement ne présente pas dans notre cas des résultats fiables étant donné que nos sinistres ne sont pas homogènes. Par contre, l'estimation de nos réserves par la méthode de Schnieper est meilleure a une erreur de 4% de prédiction. On a en 2011, deux sinistres très tardifs qui ne sont pas estimés par aucune de nos méthodes.

Méthode de provisionnement individuel

Cette méthode sínspire des travaux de Baudry & Robert et propose une séparation de modélisation de nos IBNR et nos RBNS. Pour la modélisation, nous utiliserons l'algorithme XGBoost qui est un algorithme ensembliste qui agrège des arbres et qui a la particularité de sommer le résultat de chaque arbre pour obtenir un arbre fiable. Nous utiliserons les informations de notre base de données pour enrichir le modèle.

Pour les sinistres survenus entre 2009 et 2016 et connus au 31/12/2016, nous essayerons de déterminer la charge à l'ultime. La provision obtenue sera celle de nos RBNS à laquelle s'ajoute la provision de nos sinistres inconnus pas encore déclarés que nous avons approcher grâce à une méthode de fréquence/sévérité. Nous fixons une hypothèse selon laquelle tous les sinistres sont réglées 6 années après leur survenance.

Période de survenance	Provisions réelles	Provision XGBOOST
2009	0	0
2010	0	0
2011	24 887	0
2012	10 448	3 437
2013	17 055	40 270
2014	947 005	617 121
2015	4 744 276	4 999 742
2016	13 138 527	9 674 298

TABLE 2 – Provisions individuelles

Conclusion

L'apport des algorithmes de machine learning dans le provisionnement en assurance peut être immense en valeur ajoutée comme la détermination de nos sinistres inconnus.

Dans notre étude sur les sinistres inconnus, l'absence de certaines informations a pénalisé le pouvoir prédictif de nos algorithmes. Le développement et la vulgarisation des données DSN seront un plus pour les assureurs pour mieux piloter leur portefeuille et améliorer des méthodes actuarielles comme la méthode de Schnieper ou les algorithmes de Machine Learning.

Application au risque Invalidité

Pour le risque Invalidité, nous considérons les sinistres survenus entre 2009 et 2016 de plusieurs CCN ayant des caractéristiques similaires. Cette subdivision aura pour but d'observer la précision de nos prédictions. Nous fixons. Nous chercherons à estimer le nombre de sinistre tardifs pour ces survenances. Pour cela, nous nous plaçons au 31/12/2016 et estimerons le nombre de sinistre vient les méthodes de Chain-Ladder et de Schnieper.

Survenance	Estimation Chain-Ladder	Estimation Schnieper	Observations au 31/12/2021
2009	0	0	0
2010	0	0	0
2011	0	0	1
2012	1	0	1
2013	1	1	3
2014	6	3	12
2015	16	12	11
2016	262	350	374

TABLE 3 – Estimation du nombre de sinisitres

Executive Summary

Introduction

The insurance business is characterised by its inverted production cycle. Insurers need to estimate their reserves to meet future liabilities. To evaluate these reserves or provisions they use methods that aggregate data. These methods are quite robust and are highly regarded for their simplicity of interpretation and use. Although these methods are quite efficient, we often end up with quite volatile provisions. Aggregate methods such as Chain-Ladder do not take into account the multitude of information about the claim and do not allow us to accurately calibrate our reserves, especially those allocated to late claims.

We have studied the risk of work stoppage relating to portfolios of National Collective Agreements, specifically its two sub-risks : temporary incapacity or incapacity to work and permanent incapacity to work. The coverage of work stoppage claims is determined by decision of the Social Security.

This paper proposes a reflection on the provisioning of work stoppage and the possibility of separating our RBNS provisions (known claims) from our IBNR (unknown claims). The traditional methods used in the calculation of reserves will be challenged by Schnieper's methods and an individual provisioning method.The proposed approaches are inspired by the work of Schnieper(1991)[1] and Baudry & Robert (2017) [2] which have the particularity of distinguishing between RBNS and IBNR.

In the first part, we propose a presentation of the context of the national collective agreements, the risk of work stoppage and its provisioning, and the classic methods used. In the second part, the method inspired by the work of Baudry & Robert (2017) will be presented within the framework of a line by line provisioning. It will use the gradient boosting algorithm XGBoost. Finally, in the last part of this paper, we will pay great attention to the treatment of our database and atypical claims before applying our different methods for each sub-risk.

We were unable to explore the aspect of pending disability in our study.

Context of the study

A collective bargaining agreement covers employees in the same sectoral activity. These may provide for provident and/or health care coverage. They can no longer impose an insurer, but they can propose one or more. This is a recommendation and not a choice imposed on companies.

This precision in estimating reserves is of capital importance to the insurer.

In the case of under-reserving, it underestimates its future liabilities and risks being unable to meet its future commitments.

The insurer must provide social justifications and explanations in the case of National Collective Agreements in situations of under-reserving or over-reserving.

In these different situations, the insurer portrays a negative image to its social partners, questioning its expertise and calculation methods. This may lead to likely adjustments or increases in contribution rates that are not necessary in such cases.

Application to tempority disability risk

For the Disability risk, we consider claims from 2009 to 2016 for several collective bargaining agreement with similar characteristics. We will then compare the provision estimates made with the four methods Chain-Ladder, Mack, Schnieper and individual provisioning. Disability is limited to three years in our study.

We remind you that we have not explore the pending disability in this study.

Aggregated methods

In the case of our aggregated methods, our claims incurred between 2009 and 2016 are stopped at a view as at 31/12/2016. The reserves will be estimated for claims incurred between 2009 and 2016 and compared to the actual reserve amounts observed at 31/12/2021.

Occurence	Reserving	Reserving with Chain-Ladder	Reserving with Schnieper
2009	0	0	0
2010	0	0	0
2011	24 887	1 115	-206
2012	10 488	1 916	13
2013	17 055	50 342	40 546
2014	947 005	1 228 861	1 002 577
2015	4 744 276	5 417 189	4 449 890
2016	13 138 527	14 894 023	12 624 545

TABLE 4 – Difference between the actual reserves and the Chain-Ladder and Schnieper estimates

On the total reserve amount, Chain-Ladder obtains an error of more than 14.4%. The most common method of reserving does not provide reliable results in our case, as our claims are not homogeneous. On the other hand, the estimation of our reserves by the Schnieper method is better with a prediction error of 4 %.

Individual reserving method

This method is inspired by the work of Baudry Robert which has proposed a separation of the modelling of our IBNR and RBNS. For the modelling, we will use the XGBoost algorithm, which is an ensemblistic algorithm that aggregates trees and has the particularity of summing the result of each tree to obtain a reliable tree. We will use the information in our database to enrich the model.

For claims that occurred between 2009 and 2016 and are known as of 31/12/2016, we will try to determine the ultimate expense. The resulting provision will be that of our RBNS. In this case, we set an assumption that all claims are settled 6 years after their occurrence.

Period of occurrence	Actual Reserves	XGBOOST Reserves
2009	0	0
2010	0	0
2011	24 887	0
2012	10 448	3 437
2013	17 055	40 270
2014	947 005	617 121
2015	4 744 276	4 999 742
2016	13 138 527	9 674 298

TABLE 5 – Individual reserves

For late claims, Baudry & Robert in their article propose to estimate IBNR by a frequency/severity method. We did not finalise this modelling, which proved to be rather complex.

Application to permanent disability risk

For the permanent disability risk, we consider claims that occurred between 2009 and 2016 for several collective bargaining agreement with similar characteristics. The purpose of this subdivision will be to observe the accuracy of our predictions. We set. We will try to estimate the number of late claims for these occurrences. To do this, we will take 31/12/2016 as the date and estimate the number of claims using the Chain-Ladder and Schnieper methods.

Occurence	Late claims withChain-Ladder	Late claims with Schnieper	Late claims on 31/12/2021
2009	0	0	0
2010	0	0	0
2011	0	0	1
2012	1	0	1
2013	1	1	3
2014	6	3	12
2015	16	12	11
2016	262	350	374

TABLE 6 – Estimation du nombre de sinisitres

Conclusion

The contribution of machine learning algorithms in insurance reserving can be immense in added value such as determining our unknown claims.

In our study on unknown claims, the absence of certain information penalized the predictive power of our algorithms. The development and popularization of DSN data will be a plus for insurers to better manage their portfolio and improve actuarial methods such as the Schnieper method or Machine Learning algorithms.

Remerciements

Je souhaite adresser mes remerciements les plus sincères à toutes les personnes qui ont apporté leur aide, et ont contribué, à des degrés divers, à l'élaboration de ce mémoire.

Je remercie tout particulièrement mes tuteurs d'alternance Yousra KHARROUBI & Anouar LACHIRI pour le soutien, les conseils, leurs implications et encouragements qui ont été précieux pour la réalisation de ce mémoire.

Je remercie également ma manager Jing WANG de m'avoir proposé ce sujet intéressant et de m'avoir permis d'adapter mon emploi du temps pour me consacrer à cette étude.

Je tiens à témoigner toute ma reconnaissance à mes collaborateurs du service souscription et comptes branches qui ont su m'accueillir, pour les agréables moments passés au sein de ce service, pour leur soutien et les différentes discussions sur mon sujet de mémoire qui m'ont permis de mieux développer mon analyse. Une mention spéciale pour Stanislas D'HUMIERES et Karim FIRAS pour leur aide et contribution dans ce mémoire.

J'adresse mes remerciements à l'ensemble du corps professoral de l'ISUP pour la qualité de leurs enseignements, en particulier Olivier LOPEZ mon tuteur pédagogique pour sa disponibilité et son aide pour l'élaboration de ce mémoire.

Finalement je remercie mes parents et mes frères pour leur soutien sans faille et leurs encouragements.

Table des matières

I Contexte et modélisations actuelles **14**

1 Contexte et objectif de l'étude **15**
1.1 Les Conventions Collectives Nationales . . . 15
1.2 Rappels de notions de l'arrêt de travail . . . 16
1.2.1 L'incapacité temporaire total de travail . . . 16
1.2.2 L'invalidité . . . 18
1.2.3 Notions de censures et troncatures . . . 20
1.3 Le provisionnement en Arrêt de travail . . . 21
1.4 Contexte de l'étude . . . 22

2 Méthodes de provisionnements agrégées **24**
2.1 Les triangles de développement . . . 24
2.2 Méthode de Chain Ladder . . . 25
2.3 Méthode de Bornhuetter-Ferguson . . . 26
2.4 Méthode de Mack . . . 27
2.5 Méthode de Schnieper . . . 29

II Provisionnement lignes à lignes **32**

3 Modèle de provisionnement individuel **34**
3.1 Construction du modèle . . . 34
3.2 Formalisme des bases de données pour le ML . . . 36
3.2.1 Découpage sur les périodes de développement . . . 36
3.2.2 Construction de nos bases d'entraînement . . . 38

4 eXtreme Gradient Boosting (XGBoost) **41**
4.1 Quelques notions importantes . . . 41
4.1.1 Qu'est-ce-que le Machine Learning . . . 41
4.1.2 Les méthodes d'apprentissage . . . 41
4.1.3 Méthodes ensemblistes . . . 42
4.1.4 La descente de Gradient . . . 43
4.1.5 Gradient Boosting . . . 44
4.2 L'algorithme XGBoost . . . 45
4.3 Mesure de l'incertitude de prédiction . . . 46

III Application de nos modèles de provisionnement sur nos données **47**

5 Les données utilisées pour notre étude **48**
5.1 Présentation de nos données et traitements . . . 48
5.1.1 Risque Incapacité . . . 48
5.1.2 Risque Invalidité . . . 50
5.1.3 Prise en compte de l'inflation . . . 51
5.2 Détermination du seuil séparant des sinistres graves . . . 53
5.2.1 Rappel de la Théorie des Valeurs Extrêmes . . . 53
5.2.2 Méthodes de détermination du seuil . . . 55
5.2.3 Choix de notre seuil . . . 58
5.3 Statistiques élémentaires et choix des variables . . . 59
5.3.1 Incapacité . . . 59
5.3.2 Invalidité . . . 62
5.3.3 Etude des corrélations . . . 63

6 Risque Incapacité **66**
6.1 Applications de nos méthodes de provisionnement classiques . . . 66
6.1.1 Résultats de la méthode de Chain-Ladder . . . 67
6.1.2 Résultats de la méthode de Schnieper . . . 68
6.2 Modèle de provisionnement individuel . . . 69
6.2.1 Résulats de l'apprentissage des RBNS . . . 69
6.2.2 Résultats de l'apprentissage des IBNR . . . 71
6.3 Comparaison des charges prédictes . . . 75

7 Application sur le risque Invalidité **76**
7.1 Résultats de nos méthodes classiques . . . 77
7.1.1 Chain-Ladder . . . 77
7.1.2 Modèle de Schnieper . . . 77
7.1.3 Méthode de provisionnement individuel . . . 78

Première partie

Contexte et modélisations actuelles

Chapitre 1

Contexte et objectif de l'étude

1.1 Les Conventions Collectives Nationales

Une convention ou accord de branche concerne les salariés d'une même activité sectorielle. Les conventions de branche sont conclues entre :

- d'une part, une ou plusieurs organisations syndicales de salariés reconnues représentatives au sein de la branche conformément à la loi ;
- d'autre part, une ou plusieurs organisations syndicales d'employeurs reconnues représentatives au sein de la branche conformément à la loi.

Les accords de branche peuvent prévoir des garanties prévoyance et/ou des garanties frais de santé. Ils ne peuvent plus imposer un organisme assureur mais ils peuvent en revanche en proposer un ou plusieurs. Cela est une recommandation et non d'un choix imposé aux entreprises.

Les organismes recommandés sont tenus de respecter les obligations suivantes :

- ils ne peuvent refuser l'adhésion d'une entreprise relevant du champ d'application de l'accord ;
- la présence d'un degré élevé de solidarité ;
- ils sont tenus d'appliquer un tarif unique ;
- ils sont tenus d'offrir des garanties identiques pour toutes les entreprises et pour tous les salariés concernés ;

Les problèmes rencontrés par les organismes assureurs recommandés sont :

- une instabilité des effectifs des entreprises : les entreprises étant libres d'entrer ou sortir chaque année, la population couverte risque d'être instable.
- un risque d'anti-sélection car les assureurs ont l'obligation d'accepter la souscription d'une entreprise de la branche même si présentant un risque élevé tandis que les entreprises peuvent quitter annuellement le régime si elles ont un intérêt.

- le sujet des chargements : ils sont généralement en deçà et ne couvrent pas les frais de gestion des organismes assureurs.

Il existe la labellisation qui est une procédure moins contraignante que la recommandation. Elle permet aux employeurs de prendre connaissance d'offres respectant les conditions de leur branche sans l'aléa complexe de la recommandation. Elle a l'avantage de simplifier les procédures des TPE et PME.

La labellisation n'est pas obligatoirement le résultat d'une négociation et le processus ne se déroule pas selon des règles précises. Les offres labellisées sont une sélection d'un ou plusieurs contrats répondant le mieux à un appel d'offres de la branche incluant des contraintes telles que le niveau de garanties.

1.2 Rappels de notions de l'arrêt de travail

Le régime d'assurance maladie permet pour un salarié en arrêt de travail pour maladie ou accident d'être bénéficiaire d'un revenu fonction de son salaire durant sa période d'absence. Ces indemnités sont versées par la Sécurité Sociale de façon partielle, complétée par l'employeur et l'organisme assureur complémentaire.
Le risque arrêt travail regroupe le risque incapacité et de celui invalidité encore appelé incapacité permanente.

1.2.1 L'incapacité temporaire total de travail

La définition de l'incapacité d'après l'article L321-1 du Code de la Sécurité Sociale est « L'assurance maladie comporte (...) l'octroi d'indemnités journalières à l'assuré qui se trouve dans l'incapacité physique constatée par le médecin traitant, (...) de continuer ou de reprendre le travail ; l'incapacité peut être également constatée, dans les mêmes conditions, par la sage-femme dans la limite de sa compétence professionnelle et pour une durée fixée par décret (...)»[3].

Elle peut être causée par une maladie ou un accident liée à la vie privée mais aussi par un accident ou une maladie lié à la vie professionnelle.

Sa durée est de trois ans maximum. Pour un individu en état d'incapacité il y a tois causes de sorties de l'état incapacité : le rétablissement, le passage à la retraite, l'invalidité.

Arrêt de travail lié à la vie privée

L'ouverture des droits de l'arrêt de travail varie en fonction de la durée de ce dernier.

- Pour un arrêt de travail inférieur à 6 mois : le sinistré doit avoir travaillé au moins 150 heures au cours des trois mois précédent l'arrêt ou avoir cotisé sur un salaire au moins égal à 1 015 fois le montant du SMIC horaire au cours des 6 mois précédant l'arrêt.

- Pour une durée supérieure à 6 mois : il doit justifier à la date de l'arrêt de 12 mois d'immatriculation en tant qu'assuré social auprès de l'Assurance Maladie ou avoir cotisé sur un salaire au moins égal à 2 030 fois le montant du SMIC horaire au cours des 12 mois précédant l'arrêt.

L'indemnité journalière est octroyée à la suite du délai de carence de 3 jours. Toutefois le délai de carence ne s'applique pas dans les cas suivants :

- Lorsqu'il s'agit d'une rechute en moins de 48h le délai de carence n'est pas appliqué au deuxième arrêt.
- si l'assuré est en affection de longue durée et que les arrêts de travail sont en rapport avec cette maladie, le délai de carence n'est retenu que pour le premier arrêt de travail.

Le montant des IJ est de 50% du salaire journalier de base et majoré à $66,66\%$ à partir du 31ème jour d'arrêt pour un assuré ayant au moins 3 enfants dans la limite de 1,8 fois le SMIC

Nombre d'enfants à charges	Période de versements des IJ	% du salaire
moins de 3	A partir du 4^{me} jour	50%
3 ou plus	du 4^{me} jour au $28^{me} jour$	50%
	A partir du $29^{me} jour$	$66,66\%$

TABLE 1.1 – Récap des indemnités journalières versées par la SS.- cas vie privée

Arrêt de travail lié à la vie professionnelle

L'article L411-1 du code de la Sécurité Sociale définit les termes d'accident du travail et de maladie professionnelle.[3]

Dans ce cas, aucune condition n'est nécessaire pour l'ouverture des droits et il n'y a pas de délai de carence. Le jour de survenance du sinistre est à la charge de l'employeur.

L'indemnité journalière est plus élevée que dans le cas d'un arrêt de travail lié la vie privée et s'élève à 60% du salaire journalier de base jusqu'au 28^{me} jour à 80% à partir du 29^{me}jour.

Période de versements des IJ	% du salaire
Du 1^{er} jour au 28^{me}	60%
A partir du 29^{me} jour	80%

TABLE 1.2 – Récap des indemnités journalières versées par la SS- cas vie pro.

1.2.2 L'invalidité

La définition de l'invalidité ou incapacité permanente d'après l'article L341-1 est la suivante : «L'assuré a droit à une pension d'invalidité lorsqu'il présente une invalidité réduisant dans des proportions déterminées sa capacité de travail ou de gain, c'est-à-dire le mettant hors d'état de se procurer un salaire supérieur à une fraction de la rémunération soumise à cotisations et contributions sociales qu'il percevait dans la profession qu'il exerçait avant la date de l'interruption de travail suivie d'invalidité ou la date de la constatation médicale de l'invalidité.»

Il existe 3 catégories d'invalidité :

- Première catégorie : l'assuré est capable d'exercer une activité remunérée,
- Deuxième catégorie : l'assuré est incapable d'exercer la moindre activité professionnelle.
- Troisième catégoroie : l'assuré en plus d'être incapable d'exercer une activité professionnelle a besoin d'une assistance d'un tiers pour effectuer les actes ordinaires de la vie.

Pour un assuré en invalidité, les causes de sorties de cet état sont le passage à la retraite,le décès et le rétablissement (en invalidité 1ere catégorie).

Les assurés ayant effectués 3 années en invalidité ou ceux ayant des taux d'incapacité très importants, passe en invalidité. Tout comme l'incapacité, les prestations versées en invalidité dépendent du motif de l'arrêt de travail.

Arrêt de travail lié à la vie privée

L'assuré doit remplir les conditions suivantes :

- Ne pas avoir l'âge légal de départ à la retraite soit 62 ans.
- Avoir sa capacité de travail réduite d'au moins 2/3.
- Être affilié depuis au moins 12 mois au moment du sinistre.
- Justifierau moins 600 heures de travail ou avoir cotisé un salaire au moins égal à 2030 fois le SMIC horaire au cours des 12mois qui précèdent le sinistre ou la constatation médicale.

Les montants des pensions invalidités sont recapitulés dans le tableau ci-dessous.

Catégorie de pension	Pourcentage du salaire moyen perçu pendant les 10 meilleures années d'activité	Montant mensuel minimum au 01/01/2022	Montant mensuel maximum au 01/01/2022
Catégorie 1	30%	297,20 €	1 028,40 €
Catégorie 2	50%	297,20 €	1 714,00 €
Catégorie 3	50% + majoration pour tierce *personne*[1]	297,20 €	1 714,00 € + 1 146,69 €

TABLE 1.3 – Montant des pensions d'invalidité lié à la vie privée

Arrêt de travail lié à la vie professionnelle

A la suite d'une maladie ou un accident de travail, lorsque l'état l'assuré se détériore et que sa capacité de travail est impacté il peut être déclarer en incapacité permanente à la suite d'une consultation avec un médecin conseil. Ce dernier lui fixe un taux d'Incapacité Permanente taux d'Incapacité Permanente qui dépend de la nature de l'infirmité, de l'âge, de l'état général, des facultés physiques et mentales et des aptitudes et qualifications professionnelles selon l'article L434-2 du code de la Sécurité Sociale[3]. La pension d'invalidité dépendra de ce taux :

- Si le taux est inférieur à 10% : l'indemnité est versée à l'assuré sous forme d'un capital forfaitaire fonction du taux d'incapacité. Son montant, fixé par décret, est forfaitaire et variable selon votre taux d'incapacité.

Taux d'incapacité permanente	Montant de l'indemnité en capital
1%	426,91 €
2%	693,90 €
3%	1 013,99 €
4%	1 600,43 €
5%	2 027,46 €
6%	2 507,64 €
7%	3 040,95 €
8%	3 628,07 €
9%	4 628,27 €

TABLE 1.4 – Montant de l'indemnité en capital versée selon le taux d'incapacité permanente (depuis le 01/01/2022)

- Si le taux est supérieur ou égal à 10% : l'assuré recevra une rente viagère mensuelle ou trimestrielle versée jusqu'au décès de la victime. Le montant de la rente est donné par la formule suivante :

  Salaire annuel de base$*((0,5 *$ part du taux IPP $< 50\%) + (1,5 *$ part du taux IPP $\geq 50\%))$

1.2.3 Notions de censures et troncatures

Sur une période d'observation, nos données ne sont pas totales précises. Il nous manque des information sur ce qui se passe avant le début ou après la fin de la période d'observation.

Troncature

La troncature concerne les assurés qui étaient déjà en arrêt avant le début de la période d'observation. Les informations entre le début du sinistre et le début de la période d'observation sont absents. Les assurés dans ce cas sont dits tronqués à gauche.

Nous considérons les sinistres survenus entre 01/01/2000 et 31/12/2021 ce qui nous enlève l'aspect troncature.

Censure

La censure concerne les assurés qui sont encore en arrêt de travail après la fin de la période d'observation, ils sont dits censurés à droite car la durée de leur sinistre n'est pas observée dans sa totalité. Aussi on considère comme censure à droite le cas de la résiliation de contrat avant la fin de la période de l'observation alors que l'assuré est toujours en incapacité.

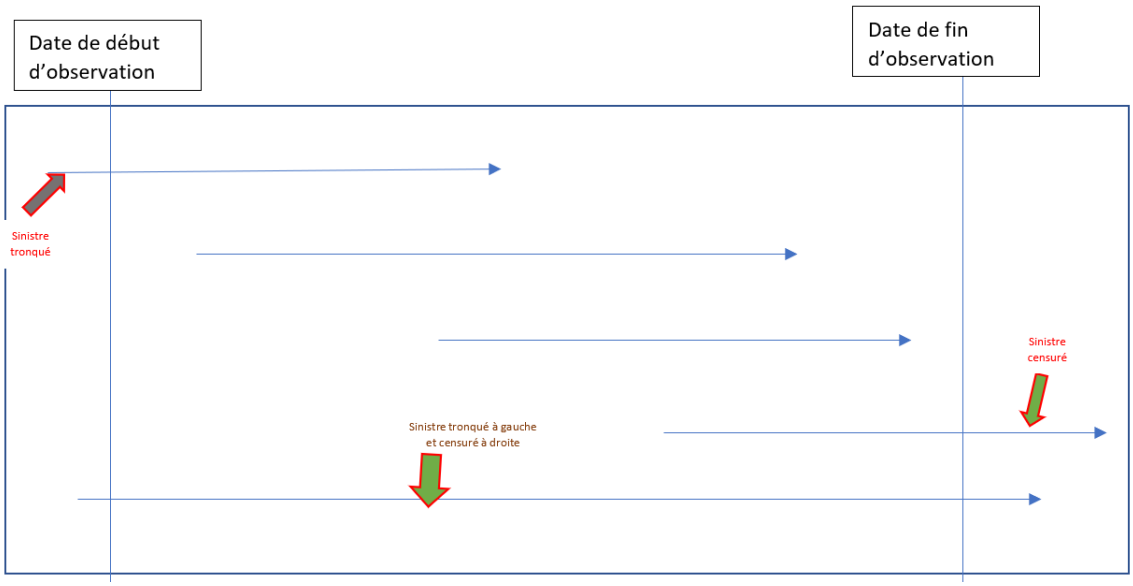


FIGURE 1.1 – Exemples de censures à droite et de troncature à gauche

1.3 Le provisionnement en Arrêt de travail

En assurance, du fait du cycle est inversé de production, l'assuré paye une prime et ne reçoit la prestation (probable en cas de sinistre) qu'à posteriori. Cette inversion du cycle rend impossible la détermination exacte de la richesse de l'assureur à un instant donné puisqu'elle ne connaît pas exactement ses engagements.

A chaque période d'arrêté, il comptabilise en provisions techniques la charge qu'il estime devoir verser dans le futur pour régler aussi bien les sinistres survenus connus que ceux dont il n'a pas encore connaissance jusqu'à la fin de la période.

La détermination des provisions est donc un impératif pour l'assureur pour survivre et pouvoir faire face à ces engagements.

Dans le cas de l'arrêt de travail,les principales provisions techniques que nous chercherons à constituer sont :

- Les provisions mathématiques (PM) qui représentent la partie complémentaire versée par l'assureur, ajustée au fait de rester ou non en arrêt de travail, sur toute la durée possible de l'arrêt. Année après année, les PM s'amenuisent pour se liquider en prestations. Si elles sont bien calculées, dans le cas d'un provisionnement en vision Best Estimate et non prudente l'année de survenance doit couvrir l'ensemble des versements effectués dans le temps pour une survenance donnée.

- les Provisions pour Sinistres A Payer(PSAP) ou RBNS :«Reported But Not Settled» qui font référence à des sinistres déclarés, toujours en cours. Le montant des prestations ultimes versées n'est pas encore connu à la date d'arrêté.

- les provisions IBNR ou Provisions pour Sinistres Inconnus (PSI) : signifiant «Incurred But Not Reported» correspondent dans notre étude aux sinistres déjà survenus mais dont l'assureur n'a toujours pas connaissance. Ces sinistres sont souvent déclarés plusieurs mois après leur survenance. Cela peut s'expliquer par plusieurs raisons :

  – le délai de franchise présent sur certains contrats peut inciter les personnes à une déclaration plus tardive. Il est fréquent de rencontrer des franchises de 90 jours en arrêt de travail, l'assuré victime d'un sinistre déclarera son litige qu'après ce délai de 90 jours.

  – Un manque de documents dans le dossier à fournir peut ralentir la déclaration du sinistre.

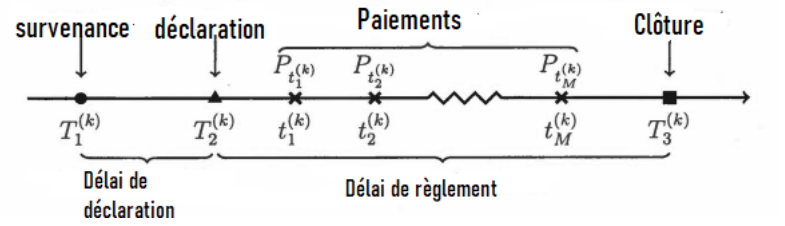


FIGURE 1.2 – Schéma des étapes d'un sinistre

1.4 Contexte de l'étude

Comme modalités de calcul des IBNR et des RBNS, il est spécifié dans le Code des Assurances qu'elles se doivent se baser sur des méthodes statistiques.

Le calcul de ces provisions se fait principalement par des méthodes actuarielles (ex : méthodes de cadencement à partir de triangles, études sur la fréquence/sévérité des sinistres) à partir d'observations historiques sur le portefeuille.

La détermination des provisions dans notre étude intervient dans le cadre des comptes de résultats annuels livrés aux partenaires sociaux et/ou coassureurs.

Que ce soit un compte comptable avec une vision annuelle ou une vision pluriannuelle avec un compte de survenance, nos calculs de provisions sont déterminants pour le pilotage du risque et aussi pour donner l'état des comptes.

Cette précision d'estimation des provisions est capital pour l'assureur. En cas de sous-provisionnement, il sous-estime ses charges futurs et risque de ne pas pouvoir faire face à ses engagements futurs.
Ainsi, en cas de sous-provisionnement ou de sur-provisionnement, l'assureur se doit d'apporter des justications et explications sociaux dans le cas des Conventions Collectives Nationales. Dans ces différentes situations, l'assureur donne une mauvaise image de lui aux partenaires sociaux et remet en cause son expertise et ses méthodes de calcul.

Par exemple, en cas de sur-provisionnement, il se risque à un redressement fiscal mais aussi à une incompréhension des partenaires sociaux dans le cas des Conventions Collectives Nationales. En effet sur un portefeuille où les provisions sont surestimées, on a une dégradation des résultats. Cette dégradation peut conduire à des probables redressements ou des hausses des taux de cotisations qui ne sont pas dans ce cas nécessaire.

Cette précision d'estimation des provisions est capitale pour l'assureur. En cas de sous-provisionnement, il sous-estime ses charges futurs et risque de ne pas pouvoir faire face à ses engagements futurs. Ainsi, puisqu'il n'est plus en mesure de respecter ses engagements auprès de ses clients il risquerait dans notre cas de perdre la recommandation.

A l'inverse, en cas de sur-provisionnement, il se risque à un redressement fiscal mais aussi à l'incompréhension des partenaires sociaux dans le cas des Conventions Collectives Nationales. En effet sur un portefeuille où les provisions sont surestimées, on a une dégradation des résultats. Cette dégradation peut conduire à des probables redressements ou des hausses des taux de cotisations qui ne sont pas dans ce cas nécessaire. Dans cette situation, l'assureur donne une mauvaise image de lui aux partenaires sociaux et remet en cause son expertise et ses méthodes de calcul.

L'arrêt de travail présente des problématiques de provisionnement depuis quelques années. En effet, sur le risque incapacité on observe un sur-provisionnement et sur le risque invalidité au contraire un sous-provisionnement.

Pour résoudre cette problématique, de nombreuses études ce sont penchées sur les tables du BCAC qui pourraient être la cause à ces problèmes de provisionnement.

Dans notre étude, nous envisagerons la problématique autrement en nous penchant sur les méthodes de calcul de nos provisions.

L'objet de ce mémoire est de confronter les méthodes classiques de provisionnement, existantes bien avant les progrès technologiques que nous avons de nos jours, à des méthodes novatrices et récentes qui pleines de promesses aux résultats très fiables. Une estimation plus exacte des IBNR et RBNS sera le but ultime de ce mémoire.

Chapitre 2

Méthodes de provisionnements agrégées

Ces méthodes dites agrégées utilisent des triangles de données agrégées. Parmi ces méthodes, nous avons la méthode de Chain-Ladder qui est la plus populaire chez les actuaires et la méthode de Mack.

2.1 Les triangles de développement

Ces méthode se basent sur les cadences de développement estimées à partir de triangles de liquidation. Ces triangles contiennent les informations sur l'historique des sinistres dans le portefeuille regroupées par période de survenance et par période de développement. Elles nous donnent une vision agrégée et globale. Dans un triangle, Les notations sont les suivantes :

- i correspond à l'indice des périodes de survenance des sinistres, avec $i = 1, \ldots, n$;
- j correspond à l'indice des périodes de développement des sinistres, avec $j = 1, \ldots, n$;
- $X_{i,j}$ correspond au montant des règlements ou le nombre de règlements effectuées la période j pour les sinistres survenus à la période i ;

Considérons un triangle contenant des informations sur les sinistres (paiements ou nombre de sinistres) :

$i \backslash j$	0	1	$\cdots$	j	$\cdots$	$n-1$	n
0	$X_{0,0}$	$X_{0,1}$	$\cdots$	$X_{0,j}$	$\cdots$	$X_{0,n-1}$	$X_{0,n}$
1	$X_{1,0}$	$X_{1,1}$	$\cdots$	$X_{0,0}$	$\cdots$	$X_{1,n-1}$	
$\vdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\ddots$		
i	$X_{i,0}$	$X_{i,1}$	$\cdots$	$X_{i,j}$			
$\vdots$	$\cdots$	$\cdots$	$\ddots$				
$n-1$	$X_{n-1,0}$	$X_{n-1,1}$					
n	$X_{n,0}$						

FIGURE 2.1 – Triangle de liquidation

Nous notons $C_{i,j} = \sum X_{i,j}$ le montant ou le nombre cumulé de la période de survenance i et de la période de développement j.

Dans la suite de cette partie, nous étudierons deux méthodes agrégées déterministes : la méthode Chain-Ladder et la méthode de Bornhuetter-Ferguson. Enfin, nous nous étudierons la méthode stochastique de Mack encore appélée Chain-Ladder stochastique.

2.2 Méthode de Chain Ladder

La méthode de Chain Ladder est une méthode déterministe permettant de déterminer les éléments de sinistres futurs (coût ou nombre de sinistre) à partir de ceux ayant eu lieu. Elle est très répandue dans le domaine de l'actuariat du fait de sa simplicité d'interprétation et d'implémentation.

Hypothèses

Cette méthode se repose sur l'hypothèse suivante : Les facteurs de développement sont supposés indépendants de la période de survenance i autrement $\forall j = 0, \ldots, n-1$ fixé :

$$f_j = \frac{C_{i,j+1}}{C_{i,j}} = \ldots = \frac{C_{n-j,j+1}}{C_{n-j,j}}, \forall i$$

.
Cette hypothèse se fonde sur une proportionnalité de nos périodes de survenance. En pratique ces égalités ne sont qu'approximativement vérifiées. On choisit donc un facteur de développement commun :

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

$\forall i$.
Les paiements cumulés sont déterminés à partir de la formule :

$$\widehat{C}_{i,j} = \prod_{k=j-1}^{n-i} \widehat{f}_k C_{i,n-i}$$

Les réserves ultimes RBNS+IBNR sont ainsi estimées pour la période de survenance i par :

$$\widehat{R}_i = \widehat{C}_{i,n} - C_{i,n-i}$$

La réserve ultime pour l'ensemble des sinistres est donnée par :

$$\widehat{R} = \sum \widehat{R}_i$$

Avantages

Cette méthode a le bénéfice d'être très bien documentée car utilisée préférentiellement par les actuaires.

Elle se base sur des hypothèses fortes comme la loi des Grands Nombres ce qui a permis d'utiliser

ses bases pour d'autres méthodes comme l'estimation stochastique de Mack.

Son utilisation est simple car elle utilise des inputs, ici des triangles, que l'on peut obtenir sans grand retraitements des données. Les résultats obtenus sont facilement interprétables et vulgarisables.

Inconvénients

Les hypothèses de la méthodes sont difficilement toujours valides. Une absence de linéarité entre les coefficients est souvent apparue ce qui remet en cause l'indépendance entre les périodes de survenance.

Également, la méthode n'est pas complète et nécessite des ajustements faites par l'actuaire pour tenir compte des facteurs influents sur les sinistres comme par exemple l'inflation. En effet, les valeurs de la dernière année de survenance du triangle découlent du produit de n-1 estimations de facteurs ; cela montre une forte sensibilité à la première valeur estimée. Il suffit que l'année la plus récente n'ait plus du tout les mêmes caractéristiques que les années d'historique pour que les prédictions soient erronées et pas du tout représentatives de l'année considérée. Aussi, les derniers coefficients de développement peuvent être très instables car ils ne sont basés que sur deux valeurs ce qui rend l'estimation très peu robuste. Cela introduit un biais qui n'existerait pas dans le cas d'une modélisation ligne à ligne.

2.3 Méthode de Bornhuetter-Ferguson

La méthode de Bornhuetter- Ferguson [**?**], développée en 1972 est une méthode déterministe. Elle a la particularité d'estimer l'ultime en y intégrant des informations externes au triangle de développement. Contrairement à la méthode de Chain-Ladder, celle de Bornhuetter-Ferguson utilise des cadences de développement ce qui la rend plus robuste.

On définit la cadence $\gamma_{i,j}$ pour la période de survenance i et la période de développement j par :

$$\forall (i,j) \in [0,n]^2, \gamma_{i,j} = \frac{C_{i,j}}{C_{i,n}}$$

Hypothèses

(H1) : Les sinistres cumulés $(C_{i,j})$ sont indépendants suivant la période de survenance i ; (H2) : Il existe des paramètres $\alpha_i et \gamma_j$ tel que $\forall i,j \in [0,n]^2$:

$$C_{i,j} = \alpha_i * \gamma_j$$

Estimations

On estime les cadences de développement par :

$$\gamma_j^* = \frac{1}{\prod_{k=j}^{n} \frac{C_{i,j+1}}{C_{i,j}}}$$

L'estimation de la charge est donnée par :

$$C_{i,j} := C_{i,n-i} + (\gamma_j^*) - \gamma_{n-i}^*) * \alpha_i^*$$

avec les α_i^* sont les estimateurs externes de nos charges.

Avantages et inconvénients

La méthode de Bornhuetter-Ferguson apporte plus de robustesses grâce aux informations dans les α_i. Cependant, ces informations externes ne sont pas uniquement représentatives de nos charges.

2.4 Méthode de Mack

Il s'agit de la version stochastique de la méthode de Chain-Ladder introduite en 1993 par R.Mack. [4].

A l'inverse du modèle de Chain Ladder, celui introduit par Mack en 1930 est non-paramétrique et ne fait aucune hypothèse de distribution sur les composantes du triangle de paiements cumulés.

Principe

Ce modèle s'applique sous les hypothèses suivantes : *Hypothèse 1. L'indépendance des années de survenances Les variables aléatoires $C_{i,j}$ et $C_{i,k}$ sont indépendants, pour tout $i \neq k$.

Hypothèse 2. Pour tout j=0,...,n il existe un paramètre f_j tel que :

$$\mathrm{E}[C_{i,j+1}|C_{i,1}, \ldots, C_{i,}n] = f_j * C_{i,j+1}$$

Hypothèse 3. Il existe des facteurs $\sigma_1, \ldots, \sigma_n$ tels que pour tout $i = 0, \ldots, n$:

$$\mathrm{Var}[C_{i,j+1}|C_{i,j}, \ldots, C_{i,0}] = \sigma_j^2 C_{i,j}$$

Sous ces trois hypothèses, nous avons conditionnellement à notre triangle T :

$$\mathrm{E}[C_{i,n}|T] = \prod_{k=j-1}^{n-i} f_k C_{i,n-i}$$

Les estimateurs des facteurs f_j sont estimés par :

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

La charge ultime est estimée similairement à la méthode de Chain-Ladder par la formule :

$$\widehat{C}_{i,j} = \prod_{k=j-1}^{n-i} \widehat{f}_k C_{i,n-i}$$

et

$$\widehat{R}_i = \widehat{C}_{i,n} - C_{i,n-i}$$

L'estimation du paramètre σ_j^2 est pour $j = 1, \ldots, n-1$:

$$\widehat{\beta_j^2} = \frac{1}{n-j-1} \sum_{i=0}^{n-j-1} C_{i,j} (\frac{C_{i,j+1}}{C_{i,j}} - \lambda_j)^2$$

Pour estimer σ_j^2 pour j=n :

- Si $\widehat{f}_n = 1$, alors $\widehat{\sigma_n^2} = 0$
- Sinon, $\widehat{\sigma_n^2} = min(\frac{\widehat{\sigma_{n-1}^2}}{\widehat{\sigma_{n-2}^2}}; min(\widehat{\sigma_{n-1}^2}; \widehat{\sigma_{n-2}^2}))$

Analyse des erreurs de prédictions

A partir de ces estimateurs, il est possible d'estimer l'erreur de prévision ce qui nous apprend sur la qualité de notre prédiction. Dans notre cas, nous choisirons le MSEP (Mean Square Error Prediction) c'est-à-dire la distance moyenne entre l'estimateur des réserves de chaque année de survenance $\widehat{R}_i$ et la véritable valeur R_i

$$MSEP(\widehat{R}_i) = \widehat{C}_{i,n}^2 * \sum_{n-i}^{n-1} \frac{\widehat{\sigma}^2}{\widehat{\lambda}_j^2} * (\frac{1}{\widehat{C_{i,j}}} + \frac{1}{\sum_{i=1}^{n+1} C_{i,j}})$$

On en déduit l'erreur moyen de tout le portefeuille par :

$$MSEP(\widehat{R}_i) = \sum_{i=1}^{n} [MSEP(\widehat{R}_i + 2\widehat{C}_{i,n}^2 * (\sum_{k=i+1}^{n} \widehat{C}_{k,n}) * \sum_{j=n+1-i}^{n-1} \frac{\widehat{\sigma}_j^2}{\widehat{f}_j^2 * \sum_{l=1}^{n-j} C_{l,j}}]$$

Avantages

Cette méthode a le mérite d'estimer l'erreur de prédiction ce qui donne l'incertitude la charge ultime.

Inconvénients

La méthode est excessivement sensible aux facteurs de développement. Dans le cas d'une année atypique, une seule irrégularité dans les facteurs peut perturber l'estimation de la volatilité. De plus, elle donne l'erreur de prédiction mais ne permet pas d'obtenir la distribution

de la provision. L'assureur peut alors estimer le niveau de prudence de la provision à l'aide de Mack uniquement en fixant la loi de distribution par lui-même.

Cette méthode présente enfin l'inconvénient d'être difficilement interprétable et peut conduire à de mauvais résultats dans le cas de données non régulières.

2.5 Méthode de Schnieper

Cette méthode a l'avantage d'utiliser l'exposition. Contrairement à Chain-Ladder où les sinistres latents sont proportionnels à ce qui est connu, l'idée de Schnieper [1] est la suivante :

- Les sinistres inconnus sont proportionnels à l'exposition ;
- Les sinistres déclarés non clôturés sont proportionnels à la charge actuelle.

Nous définissons les variables aléatoires :

- $(N_{i,j})_{1\leq i,j\leq n}$ qui représente la charge des nouveaux sinistres survenus à la période de survenance i et déclarés à la période de développement j ;
- $(D_{i,j})_{1\leq i,j\leq n}$ qui représente la baisse de la charge estimé des sinistres déclarés les années précédentes.

On construit le modèle en considérant $D_{i,1} = 0\ \forall i = 1,\dots,n$ et $D_{i,j} = C_{i,j-1}\ \forall i = 1,\dots,n$ et $\forall j = 2,\dots,n$. Les charges s'écrivent ainsi :

$$C_{i,1} = N_{i,1}$$

$$C_{i,j+1} = C_{i,j} + N_{i,j+1} - D_{i,j+1}$$

La connaissance de deux triangles parmi C, N et D permet d'en déduire le troisième. On considère l'ensemble $H_k = \{N_{i,j}; D_{i,j} | i + j \leq k + 1\}$.

L'exposition

On note $Expo_i$ l'exposition à l'année i. Elle correspond au nombre de contrats souscrits/en cours dans le portefeuille. Le nombre de sinistres tardifs évolue naturellement avec l'exposition en ce sens que si l'exposition du portefeuille augmente plus on est exposé à plus de sinistres tardifs.

Sur notre portefeuille de CCN, nous n'affilions pas en prévoyance pour l'instant. Des travaux futurs sur la DSN dont les résultats sont déjà exploitables pourront permettre d'affiner nos résultats de façon plus juste.

Dans notre étude pour identifier l'exposition, nous déterminons un effectif théorique moyen en déterminant le rapport entre les cotisations brutes encaissées sur l'année i sur les cotisations moyennes de l'année i par personne. Cet effectif théorique sera alors notre exposition puisque nous considérons tous les contrats d'une CCN comme identique.

Le modèle repose sur les hypothèses suivantes :

Hypothèse 1

Les variables aléatoires $(N_{i_1,j}, D_{i_1,j})_{1\leq i_1,j\leq n}$ et $(N_{i_2,j}, D_{i_2,j})_{1\leq i_2,j\leq n}$ sont inépendants pour $i_1 \neq i_2$

Hypothèse 2

Pour $j = 1, \leq, n$, il existe $\lambda_j \geq 0$et pour $1 \leq j \leq n-1$, il existe $\delta_j \leq 1$ tels que

$$\mathrm{E}(N_{i,j}|H_{i+j-2}) = \lambda_j * Expo_i$$

et

$$\mathrm{E}(D_{i,j}|H_{i+j-1}) = \delta_j * C_{i,j}$$

Hypothèse 3

Pour $j = 1, \leq, n$, il existe $\tau_j^2 \geq 0$ et $\sigma_j^2 \geq 0$ tels que

$$\mathrm{Var}(N_{i,j}|H_{i+j-2}) = \tau_j^2 * Expo_i$$

et

$$\mathrm{Var}(D_{i,j}|H_{i+j-1}) = \sigma_j^2 * C_{i,j}$$

L'hypothèse 2 suppose que les sinistres inconnus dépendent donc uniquement de l'exposition et du facteur de développement et que la variation des charges est fonction que des charges précédentes et du facteur δ_j.

Estimation

Les estimateurs sans biais de nos paramètres sont définis par :

$$\widehat{\lambda}_j = \frac{\sum_{i=1}^{n+1-j} N_{i,j}}{\sum_{i=1}^{n+1-j} Expo_i}$$

et

$$\widehat{\delta}_j = \frac{\sum_{i=1}^{n+1-j} D_{i,j+1}}{\sum_{i=1}^{n+1-j} C_{i,j}}$$

Ces estimateurs combinés sous nos hypothèses nous permettent de déterminer pour $i+j \geq n+1$:

$$\widehat{N}_{i,j} = \widehat{\lambda}_j * Expo_i$$

$$\widehat{D}_{i,j} = \widehat{\delta}_{j-1} * C_{i,j-1}$$

$$\widehat{C}_{i,j} = X_{i,j-1} * (1 - \widehat{\delta}_j) + \widehat{\lambda}_j * Expo_i$$

Nous mesurons l'erreur de nos paramètres pour $1 \leq j \leq n-2$ grâce aux estimateurs sans biais de σ_j^2 et τ_j^2 :

$$\widehat{\sigma}_j^2 = \frac{1}{n-j} * \sum_{i=1}^{n-j+1} Expo_i * (\frac{N_{i,j}}{Expo_i} - \widehat{\lambda}_j^2)$$

$$\widehat{\tau}_j^2 = \frac{1}{n-j-1} * \sum_{i=1}^{n-j} C_{i,j} * (\frac{D_{i,j+1}}{C_{i,j}} - \widehat{\delta}^2)$$

Conclusion du Chapitre

Les méthodes expliquées dans ce chapitre sont celles majoritairement utilisées par les actuaires à travers le monde. Elles ont montré leur robutesse mais aussi des limites sur la volatilité entre les années. Nous allons dans le chapitre suivant nous intéresser à un autre type de provisionnement en vogue depuis quelques années, le provisionnement individuel.

Deuxième partie

Provisionnement lignes à lignes

Introduction au provisionnement individuel

Pour déterminer les provisions, les actuaires ont généralement recours à des approches que l'on peut qualifier de classiques, basées sur des triangles de paiements. Cette idée est présentée dans l'article du groupe de travail ASTIN[5]. Dans cet article recensant les méthodes de provisionnement actuarielles dans le monde, il revient que la méthode de Chain Ladder est la plus plébiscitée suivie de ces extensions (Méthodes de Mack et Bornhuetter-Ferguson). Le succès de ces méthodes est conséquence de leur simplicité d'exécution et d'interprétabilité mais également des résultats obtenus qui s'approchent de la réalité.

Toutefois, le monde actuariel s'intéresse de plus en plus au provisionnement individuel.Les premiers travaux sur le sujet sont réalisés dans le cadre stochastique. Arjas (1989)[6] propose une modélisation des durées de sinistres à partir de martingales. Cette approche de modéliser la survenance des sinistres est reprise par Norberg[7] qui luiconsidére des processus de poisson.Haastrup Arjas (1996)[8] mènent des travaux similaires à ceux de Norberg[**?**] plus poussés en considérant néanmoins une modélisant bayésienne non-paramétrique. Toujours sur le stochastique. Antonio Plat (2010) [9] se basent sur les travaux de Norbert [7] et Haastrup Arjas[8] et proposent une application sur des garanties responsabilités civiles du portefeuille d'une compagnie d'assurance.

Zhao et al. (2009)[10] puis Zhao et Zhou (2010)[11] se consacrent à un modèle alternatif de sinistres individuels avec une approche semi-paramétrique pour prédire les provisions IBNR.

Wüthrich (2016)[12] utilise des arbres de régressions pour modéliser les provisions RBNS et IBNyR grâce à des réseaux de neurones. En 2017, Wüthrich(2017)[13] propose une estimation du nombre de paiements restants des RBNS à l'aide l'algorithme CART.

Lopez et al. (2016)[14] proposent quant à eux une méthode d'estimation des poids de Kaplan-Meier par arbres CART et l'appliquent à des données d'assurance, notamment pour du provisionnement individuel.

Baudry Robert (2017)[2] adoptent une approche non-paramétrique pour estimer les provisions RBNS et IBNR l'algorithme Extra-Tree.

Dans la suite de notre étude, nous allons estimer le nombre de sinistres tardifs en AT ainsi que leur durée par des méthodes de GLM et une variante de l'approche de Pietro-Parodi. Ensuite, nous allons nous inspirer des recherches de Baudry Robert [2] en appliquant un pivot sur les garanties considérées, dans notre cas l'arrêt de travail.

Chapitre 3

Modèle de provisionnement individuel

Dans ce chapitre, nous présenterons la méthode que nous avons retenu pour estimer nos provisions au niveau individuel. Cette approche est inspirée en premier des travaux de Baudry & Robert(2017) qui eux avaient effectués un cas d'étude théorique sur l'assurance de téléphone. Nous avons tenté de l'adapter au mieux au risque arrêt de travail.

Plutôt que d'utiliser les méthodes classiques de provisionnement qui agrègent nos données, nous tenterons de prédire le montant total de prestation versée sur les sinistres individuels en utilisant l'apprentissage statistique. Notre objectif est de modéliser le comportement des sinistres individuels aussi précisément que possible. Nous voulons que nos prévisions de sinistres ne puissent pas être distinguées des sinistres réels au niveau individuel, à la fois en termes de valeur attendue et de variance. Si nous atteignons cet objectif, nous pouvons établir des valeurs attendues et des niveaux de confiance pour les sinistres individuels. Nous pouvons agréger les réclamations individuelles pour déterminer la valeur attendue et les niveaux de confiance pour la réserve totale. Il y a beaucoup d'autres informations que nous pouvons recueillir à partir des prédictions des sinistres individuels, mais ne nous emballons pas.

3.1 Construction du modèle

Pour fixer le cadre de notre étude, considérons les évènements suivants :

- T_1 : la date de survenance du sinistre ;
- T_2 : la date de déclaration c'est la date à laquelle l'assureur réceptionne de la demande d'indemnisation contenant des informations sur le sinistre individuel, comme la cause exacte de l'accident. On considère $\Delta_{max,r}$ comme la durée maximum qu'on a pour déclarer un sinistre. Dans la pratique, ce délai maximum de report du sinistre n'est pas souvent appliqué chez Malakoff Humanis. Il nous a été donné de remarquer des sinistres déclarés 3 années après leur survenance. Nous considérerons $\Delta_{max,r} = +\infty$;
- t_1,t_2,...,t_M les dates de paiements de notre sinistre où M représente la dernière période développement.
- T_3 la date de clôture du sinistre : dans le cas d'un passage en invalidité, il s'agit également d'une fermeture du sinistre incapacité. On définit $\Delta_{max,s}$ tel que $T_3 - T_2 < \Delta_{max,s}$.Sur le risque incapacité de l'arrêt de travail$\Delta_{m_{ax,s}}$=1095 jours soit 12 trimestres.

Nous notons aussi $(P_t)_(T_2 \leq t < T_3)$ les paiements cumulés de nos sinistres entre t_1 et t_M.

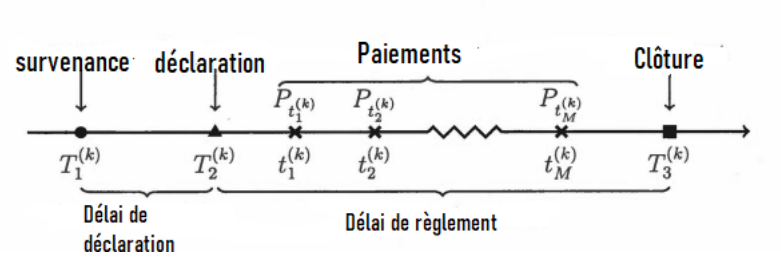


FIGURE 3.1 – Vie sinistre

Les informations caractérisant le sinistre autre que les paiements sont matérialisées dans la variable aléatoire E_t.
Rappelons que $a \wedge b$= min(a,b) qui désigne le minimum entre a et b. Nous nous plaçons à l'instant t. Les sinistres survenus avant t sont répartis en deux classes.

- Si $T_1 < t < T_2$, le sinistre i est survenu mais n'est pas encore déclaré à la mutuelle. Ce type de sinistre désigne les ≪Incurred But Not Reported (IBNR)≫.Nous ne disposons pas des informations spécifiques à chaque sinistre, mais nous utiliserons les facteurs de risque ainsi que des informations externes (informations recueillies dans les sinistres similaires) pour prédire son paiement cumulé. Nous définissons donc l'$IBNR_t$ à l'instant t de la manière suivante :

  $$IBNR_t = \mathrm{E}[P_{T_3} 1_{T_1<t} | A_t$$

  avec A_t={t<T_2,$(E_u)_{0\leq u\leq t}$}

- Si $T_2 < t < T_3$, le sinistre i a été déclaré mais le sinistre est toujours en cours de règlement.Dans cette situation, nous avons beaucoup plus d'informations sur le sinistre et l'incertitude de la prédiction dans l'évaluation finale diminue.Comme cette demande n'est pas complètement réglée, on l'appelle ≪ Reported But Not Settled (RBNS)≫ non réglée (RBNS). Nous pouvons utiliser la date de déclaration, la date de survenance, les facteurs de risque, les informations sur l'historique des sinistres ainsi que des informations externes pour prédire le montant cumulé des sinistres. A l'instant t, dans leur article Baudry & Robert [2] définissent :

  $$RBNS_t = E[P_{T_3} - P_t | B_t]$$

  avec

  $$B_t = \{T_2 < t \leq T_3, (E_u)_{0\leq u\leq t}, (I_u)_{T_2\leq u\leq t}, t_1 \leq t \leq t_M\}$$

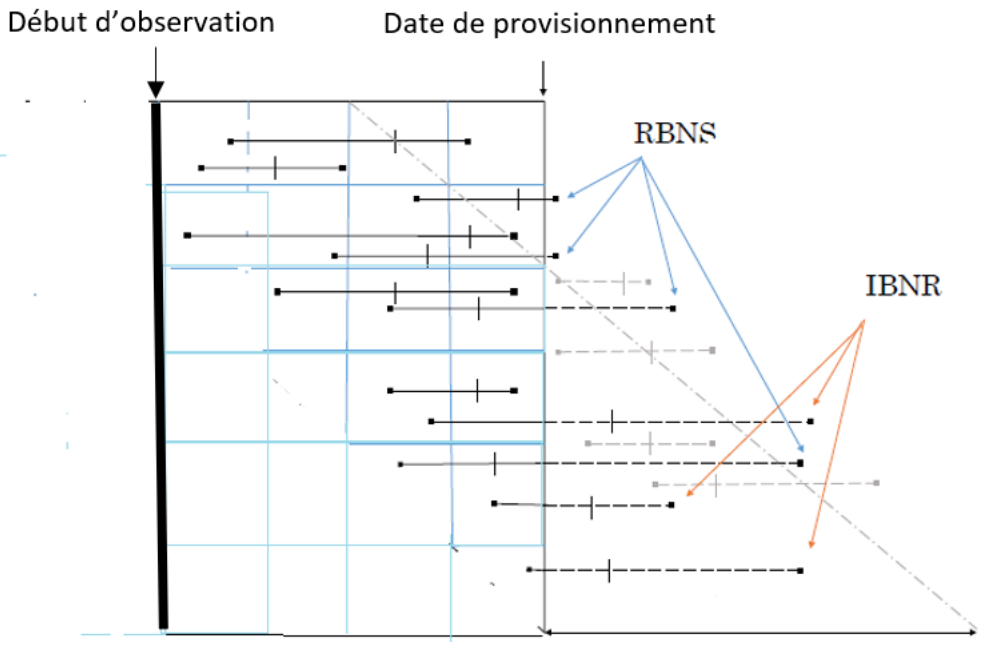


FIGURE 3.2 – Représentation graphiques des IBNR et RBNS

3.2 Formalisme des bases de données pour le ML

3.2.1 Découpage sur les périodes de développement

Dans cette partie, nous reprendrons les notations de Baudry & Robert de la section 2 et 3 que nous avons adaptés aux réalités de nos données et nos besoins en terme de prédiction.

Rappelons que la date de provisionnement est d_i. Nous allons considérer une grille de temps $d_j = \delta * j$, où $j \geq 0$ et δ est un pas de temps choisi (un trimestre dans le cas de l'incapacité et une année en invalidité). Pour j une période de développement, nous considérons la période $[d_{i+j-1}, d_{i+j}]$ pour $j \geq 1$ et définissons :

- les RBNS par :

$$RBNS_{d_{i,j}} = \mathbb{E}[P_{d_{i+j}} - P_{d_{i+j-1}} 1_{T_1 < d_i} | B_{d_i}]$$

et on obtient ainsi que

$$RBNS_{d_i} = \sum_j RBNS_{d_{i,j}}$$

Nous souhaitons déterminer la charge ultime en vision 31/12/2021 et donc cette modélisation paiement par paiement ne convient pas. Nous optons pour :

$$RBNS_{d_{i,j}} = \mathbb{E}[P_{d_{i+j}} 1_{T_1 < d_i} | B_{d_i}]$$

- Pour les IBNR, ils sont obtenus en faisant une modélisation fréquence/sévérité :

$$IBNR_{d_{i,j}} = IBNR_{freq_{i,j}} * IBNR_{gravit_{i,j}}$$

et on obtient ainsi que

$$IBNR_{d_i} = \sum_j IBNR_{d_{i,j}}$$

Les IBNR sont le résultat du rapport multiplicatif de la fréquence d'apparition des IBNR par leur coût. A la date de provisionnement d_i, on a comme montant de provisions ICR_{d_i} :

$$ICR_{d_i} = IBNR_{d_i} 1_{d_i < T_2} + RBNS_{d_i} 1_{d_i \geq T_2}$$

Pour estimer $RBNS_{d_{i,j}}$, $j = 0, 1, \ldots$ d'une CCN spécifique , nous utiliserons les variables suivantes :

  - les facteurs de risques associés évalués en t_i : ce sont les variables qui seront utilisées pour notre prédiction ;
  - les informations extérieures au sinistre (considération du comportement d'autre sinistre).

$$(E_{T_1}; F_{t_i}; E_{T_2})$$

Pour l'estimation des $IBNR_{d_{i,j}}$ j avec $j = 0, 1, \ldots$ pour une CCN donnée, nous utiliserons :

- les facteurs de risque ;
- les informations extérieures.

$$(F_{t_{i,p}}; E_{T_{1,p}})$$

En absence de la date de début de contrat pour chaque assuré, nous ne considérons pas cette variable pourtant utilisée dans l'article de Baudry & Robert.

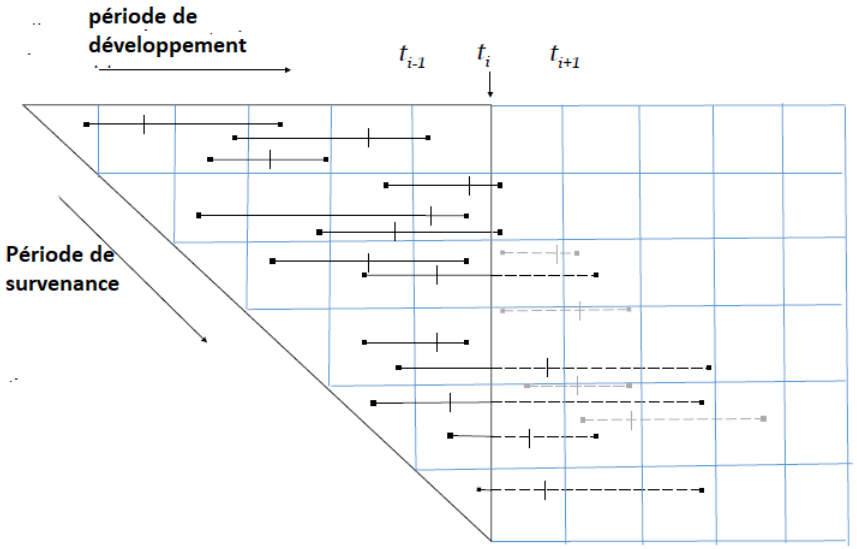


FIGURE 3.3 – Représentations graphiques des IBNR et RBNS

3.2.2 Construction de nos bases d'entraînement

Les données d'entrainement pour les réserves RBNS seront constituées de tous les sinistres déclarés avant la date de provisionnement d_i choisie. A ces données ont été jointes l'exposition du sinistre correspondant à la différence entre d_i et la date de souscription du contrat. Nous avons décidé de considérer l'ensemble des sinistres ayant une même date de souscription au 01/01/2009. Des caractéristiques du sinistre sélectionnées au préalable seront utilisées comme variables explicatives dans le calcul des réserves RBNS.
Un ensemble de données d'entraînement nous permettra de construire un modèle des exigences en matière de réserves RBNS aux dates de provisionnement historiques. Mais ce dont nous avons réellement besoin, c'est d'une vue de la réserve RBNS à la date d_i. Pour le calculer, nous devons créer un ensemble de données de test qui contient les valeurs des caractéristiques explicatives à chaque période de paiement de sinistre.

La création de l'ensemble de données IBNR suit un processus similaire à celui de RBNS mais est un peu plus complexe car tout contrat actif peut donner lieu à un sinistre IBNR. Nous passerons par une méthode de fréquence/sévérité. L'exposition nous permet de déterminer une estimation la totalité des contrats engagés. Nous cherchons une fréquence d'apparition des IBNR i.e prédire la probabilité d'un contrat de se transformer en IBNR. Cette fréquence multipliée par le coût de attendu en IBNR un sinistre nous donnerait le montant IBNR. Nous espérons déterminer le montant de provisions nécessaires pour les IBNR avec une précision significative. Cette méthode est un peu similaire à la tarification où on fixe le tarif en multipliant le coût moyen par la fréquence de sinistre.

A ce jour, Malakoff Humanis ne dispose pas d'une base de données d'affiliation en prévoyance pour le périmètre des CCN.Pour réaliser cette modélisation, nous avons décidé de simuler des contrats.

Nous allons donc approcher la réserve des IBNR par le produit du coût moyen de prestation

par la fréquence de sinistres IBNR. Le coût moyen de prestation sera calculé en à partir de la durée moyenne en arrêt par le montant l'indemnité journalière.

La variable cible de nos prédictions est le montant de paiement sur l'année. Nous avons pour objectif d'arriver à déterminer pour chaque sinistre le montant final qui en d_i n'est pas connu.

Pour former nos bases d'entrainement et de test pour évaluer nos RBNS et IBNR, nous considérons d_i la date de notre provisionnement qui sera le 31/12/2016, k qui représente le type de modèle et j la période de développement du sinistre. Le paramètre k a été laissé fixe à 1, son rôle n'étant pas explicite.

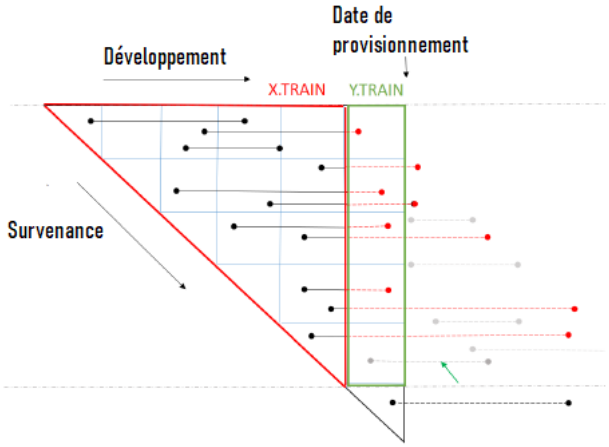


FIGURE 3.4 – Données de train X_{train} Y_{train} j=1

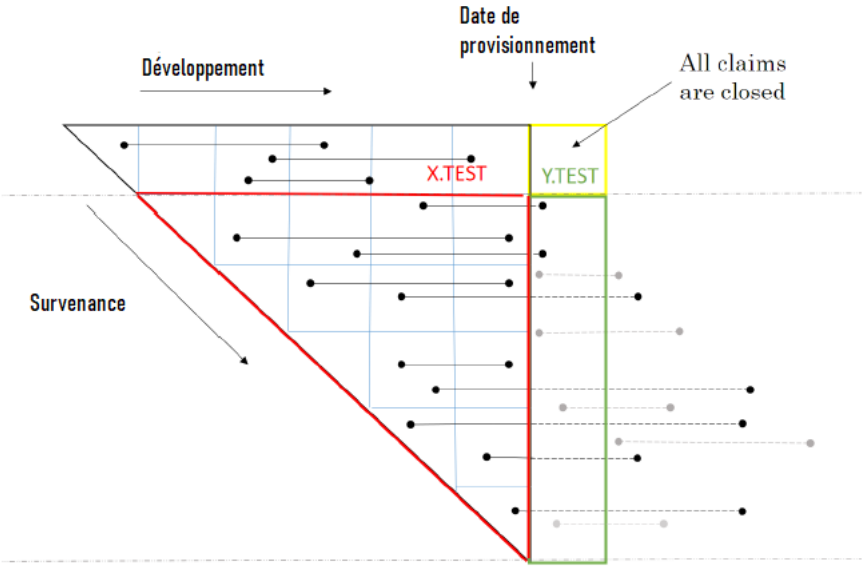


FIGURE 3.5 – Données de test X_{test} et Y_{test} j=1

Nos considérations

Nous nous permettons d'apporter des modifications pour concilier nos données à l'approche de Baudry & Robert.

- Pour simplifier le modèle dans le cas des IBNR, nous considérons nos sinistres comme étant à paiement unique. La date de premier paiement est considérée comme la date de paiement unique.
- Nous simulons des contrats d'affiliation en groupe semi-fermé, i.e nous considérons qu'il n'y a pas de sorties et qu'il n'y a que les entrées.
- En absence des dates de souscription et de sortie, nous faisons l'approximation que tous les sinistres ont un contrat en cours du 01/01/2009 au 01/01/2199.

Chapitre 4

eXtreme Gradient Boosting (XGBoost)

Dans ce chapitre, nous présenterons le fonctionnement et les hypothèses de l'algorithme de XGBoost utilisé pour notre modèle de Machine Learning.

4.1 Quelques notions importantes

4.1.1 Qu'est-ce-que le Machine Learning

Le machine learning ou apprentissage automatique peut être défini comme la capacité d'un programme à apprendre sans que cette modification ne soit explicitement programmée.On peut ainsi opposer un programme classique, qui utilise une procédure et les données qu'il reçoit en entrée pour produire en sortie des réponses, à un programme d'apprentissage automatique, qui utilise les données et les réponses afin de produire la procédure qui permet d'obtenir les secondes à partir des premières.

Comme ingrédients du machine learning nous avons :

- les données : qui permettront à l'algorithme d'apprendre ;
- l'algorithme d'apprentissage : qui est la procédure que l'on fait tourner sur ces données pour produire un modèle. On appelle entraînement le fait de faire tourner un algorithme d'apprentissage sur un jeu de données.

Ces deux ingrédients sont inter dépendants car nous ne pouvons pas avoir un bon algorithme d'apprentissage sans jeu de données qui ne soit pertinent, c'est le concept de ≪ Garbage in, Garbage out≫. Également, un modèle appris avec un algorithme inadapté sur des données pertinentes ne pourra pas être de bonne qualité.

4.1.2 Les méthodes d'apprentissage

L'apprentissage automatique selon consistent en l'étude des algorithmes qui permettent aux programmes de s'améliorer automatiquement, par expérience. Ceci consiste à examiner des données, appelées observations, pour en tirer des conclusions et des règles de décisions permettant d'effectuer des prédictions. L'objectif principal de cette approche est que ces prédictions ne soient pas exclusives aux observations utilisées pour l'apprentissage, mais plutôt généralisables sur de nouvelles données de même nature. Les règles et les conclusions déduites forment un

programme appelé prédicteur. L'algorithme permettant de générer un tel prédicteur est appelé algorithme d'apprentissage.

On distingue trois types d'algorithmes : les algorithmes supervisés et non supervisés.

- Les algorithmes supervisés consiste à apprendre un prédicteur, capable de résoudre une tâche, à partir de l'observation des exemples de solutions. Un exemple est un couple formé par une instance, et la sortie correspondante lorsque la tâche est effectuée correctement. L'algorithme supervisé vise à apprendre ce lien via différents exemples, pour ensuite être capable de prédire une valeur de sortie à partie des données d'entrées seules.
- Les algorithmes non supervisés n'intègrent pas cette notion d'entrée/sortie. Ils visent plutôt à créer des groupes homogènes de données. Ces algorithmes sont souvent moins performants.
- Les algorithmes semi-supervisés sont à mi-chemin entre les algorithmes supervisés et non-supervisés, car ils utilisent à la fois des données étiquetées et non étiquetées pour former généralement une petite quantité de données étiquetées et une grande quantité de données non étiquetées. Les systèmes qui utilisent cette méthode sont capables d'améliorer considérablement la précision de l'apprentissage.

4.1.3 Méthodes ensemblistes

L'≪Ensemble Learning≫ est un concept qui a pour idée est de former plusieurs modèles utilisant le même algorithme d'apprentissage. On y utilise la combinaison de plusieurs modèles individuels pour former un modèle plus fort et précis. Cette méthode permet au modèle de gagner en précision et en stabilité avec une variance moindre. Il existe deux grandes méthodes ensemblistes :le Bagging et le Boosting.

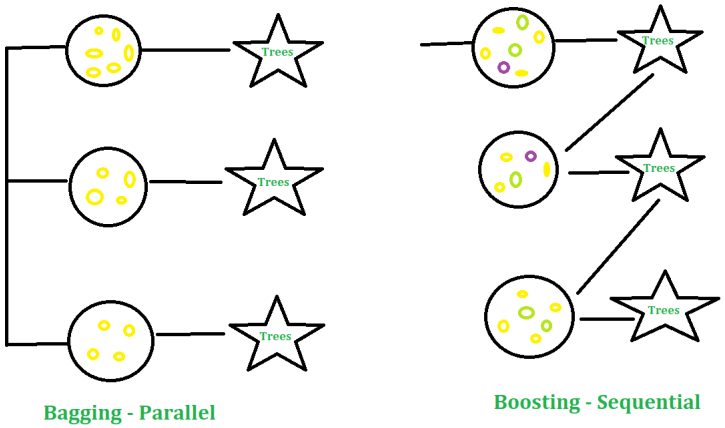


FIGURE 4.1 – Méthodes ensemblistes

Le Bagging

Le Bagging qui regroupe un ensemble de méthodes introduit par Brieman[1996]. Le terme vient de Bootstrap Aggregating. C'est une approche en parallèle car on a la réalisation de

plusieurs sous-ensembles indépendants.

Les principales étapes du bagging sont les suivantes :

- On prend un nombre X d'observations des données de départ avec remise ;
- On sélectionne aléatoirement K de nos P variables ;
- Notre modèle est entrainé sur cette portion aléatoire
- On répète les étapes précédentes N fois ;

Le résultat avec le plus de votes (le plus fréquent) devient le résultat final de notre modèle. Dans le cas de régression, on prendra la moyenne des votes de tous les arbres.

Le Boosting

Le Boosting est une méthode ensembliste permettant de prendre en compte le phénomène de sur-dispersion en combinant plusieurs classifieurs. A la différence des méthodes de Bagging, les prédicteurs ne sont plus créés de manière indépendantes. En effet, chaque arbre apprend des informations issues des arbres précédents. Bien que le fonctionnement dépende des différents algorithmes, le principe de fonctionnement reste globalement identique. Les algorithmes de Boosting se concentrent sur un apprentissage plus lent d'un même jeu de données.

Un premier arbre est expérimenté sur la base d'étude et permet d'obtenir un premier modèle. Un nouvel arbre est alors entrainé avec pour variable objectif, les résidus de l'arbre précédant. Ce nouvel arbre cherche ainsi à acquérir les informations non captées par le premier modèle. Cet arbre supplémentaire est ainsi ajouté au modèle afin de permettre un ajustement des résidus. En boosting, chaque votant est appelé un votant faible. Il est défini comme un votant de performances légèrement meilleures à celui qui choisit une étiquette de manière aléatoire.

L'apprentissage des votants faibles et des poids qui leurs sont associés est effectué itérativement. À chaque itération t, l'algorithme de boosting utilise les données préparées à l'itération précédente $t-1$pour appeler l'apprenant faible. L'objectif de cet appel est de produire un votant faible susceptible d'améliorer le vote de majorité actuel $f^{(t-1)}$. Ensuite, le poids associé à ce votant est calculé par l'algorithme du boosting. Finalement, les données pour la prochaine itération sont modifiées en fonction des erreurs effectuées par le vote de majorité actuel f(t).

4.1.4 La descente de Gradient

On se place dans un espace hilbertien muni du produit scalaire $< .,. >$ et la norme $||.||$. Considérons une fonction f différentiable. Pour x $\in \mathbb{E}$ on associe $f(x)$ différentiable. Le gradient de f est définie par :

$$\nabla f(x,y) = (\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y})$$

La descente de gradient est un algorithme d'optimisation qui permet de trouver le minimum de n'importe quelle fonction convexe en convergeant progressivement vers celui-ci. Il est utilisé dans l'apprentissage supervisé et permet de réduire la fonction de coût. Soit x_0 le point de

départ appartenant au domaine de définition de la fonction f, un seuil $\epsilon > 0$ et un pas η_0. L'algorithme de descente de gradient se présente comme suit :

Algorithm 1 Descente de Gradient

$k \leftarrow 0$
while $||\nabla f(x_k)|| > \epsilon$ **do**
  On calcule $\nabla f(x_k)$
  On calcule le pas optimal η_k
  $X_{k+1} \leftarrow x_k - \eta_k * \nabla f(x_k)$
  $k \leftarrow k + 1$
end while
return x_k

4.1.5 Gradient Boosting

C'est une méthode d'agrégation des modèles.Dans le cas d'arbres à agréger construit une série de modèles de façon à ce que chaque modèle soit ajouté à la combinaison pour améliorer la prédiction. ConsidéronsB_k l'arbre de l'étape k et $f_k = \sum_{i=1}^{k} B_i$ le modèle à l'étape k.
Soit X l'ensemble de nos données explicatives et Y la variable réponse. Nous prenons comme fonction de coût la fonction des moindres carrés définie de la manière suivante :

$$J = \sum_{i=1}^{m} j(Y_i, f(X_i))$$

avec $j(a, b) = \frac{(b-a)^2}{2}$ Le modèle Gradient Boosting se fait alors de la manière suivante :

- Nous construisons un premier arbre de décision B_1,visant à prédire Y à partir de X en d'autres termes $B_1(X_i) = Y_i + (B_1(Xi) - Y_i)$. Les résidus sont définis par $res_1 = -\frac{\partial J}{\partial f(X_i)} = Y_i - f_1(X_i)$;
- Pour combler les manques du premier arbre, on construit B_2 qui vise à prédire les résidus de res_1. A la construction de notre arbre, le modèle est maintenant égal à $f_2(X) = B_1(X) + B_2(X)$ avec $B_2(X_i) = (Y_i - f_1(X)) + (f_2(X_i) - Y_i)$. Nos résidus à cette étape seront donc $res_2 = -\frac{\partial J}{\partial f(X_i)} = Y_i - f_2(X_i)$
- On réitère cette procédure suivant la descente de gradient. A l'étape k on construit B_k qui cherche à prédire $res_{k-1} = -\frac{\partial J}{\partial f(X_i)} = Y_i - f_{k-1}(X_i)$

4.2 L'algorithme XGBoost

Dans cette section, nous présenterons cet algorithme dans le cadre des arbres de régression avec ces spécificités et le concept de XGBoost pour éviter cette conception de boite noire qu'on lui donne.

XGBoost pour eXtreme Gradient Boosting est un algorithme développé par Chen& Guestrin(2016)[15]. C'est un algorithme du type Gradient Booosting Machine qui a la particularité d'avoir sa complexité qui dépend des valeurs des paramètres à calibrer pour l'optimisation de l'algorithme ce qui réduit les coûts de calcul. Son implémentation est parallèle, ce qui permet de minimiser le temps d'entraînement comparé à celui du Gradient Boosting Machine. Aussi, de manière similaire au Gradient Boosting Machine, XGBoost construit des arbres en séries en vue de minimiser le biais, tout en contrôlant la variance.

La fonction objectif

Le XGBoost caractérisé par sa fonction objectif qui est une mesure des performances des paramètres. Une fonction objectif est fdéfinie comme composée d'une fonction de perte et d'un terme de régularisation :

$$Obj() = L() + \Omega()$$

où Obj est la fonction objectif ; L est une fonction de perte deux fois différentiables ; K le terme de régularisation

- L est une fonction de perte deux fois différentiables qui mesure la précision de rédiction du modèle.
- Ω le terme de régularisation K a pour objectif de contrôler la complexité des modèles et leur stabilité.

Les hyperparamètres du XGBoost

L'algorithme de XGBoost possède de nombreux paramètres. Les hyperparamètres du XGBoost ne se calibrent pas tous en même temps. Nous nous sommes focalisés sur les paramètres suivants à savoir :

- nrounds : qui est le nombre d'itérations ou le nombre de modèles individuels agrégés.
- eta : qui correspond à la vitesse de convergence de la descente de gradient.
- min_child_weight : la taille minimale d'un noeud, c'est un critère d'arrêt.
- max.depth : la profondeur maximale d'un arbre, c'est un critère d'arrêt.
- colsample_bytree : la part des variables explicatives prisent en compte aléatoirement pour la construction de chaque arbre.

Nous aurons la possibilité d'appliquer nos différentes méthodes de provisionnement sur les données arrêts de travail à notre disposition.

4.3 Mesure de l'incertitude de prédiction

Il est important de comparer les résultats de nos estimations aux données réelles. Afin de mesurer la robustesse et la prédiction de nos régressions,nous avons recours à des indicateurs statistiques. En apprentissage supervisé, les indicateurs les plus usuels sont :

- **L'erreur quadratique moyenne (MSE)**

$$\text{MSE} = \frac{1}{n}\sum_{k=1}^{n}(y_i - \widehat{y}_i)^2$$

- **L'erreur absolue moyenne (MAE)**

$$\text{MAE} = \frac{1}{n}\sum_{k=1}^{n}|y_i - \widehat{y}_i|$$

- **L'erreur relative** : qui correspond à une mesure de sur-estimation

$$\text{RE} = \frac{\sum_{k=1}^{n}(y_i) - \sum_{k=1}^{n}(\widehat{y}_i)}{\sum_{k=1}^{n}(y_i)}$$

Nous utiliserons surtout l'erreur relative dans nos comparaisons.

Conclusion

Dans cette partie, nous avons présenté l'approche que nous avons retenu pour estimer nos provisions individuelles et l'algorithme performant de boosting qui sera utilisé à savoir le XGBoost.

Troisième partie

Application de nos modèles de provisionnement sur nos données

Chapitre 5

Les données utilisées pour notre étude

Ce chapitre a pour objectif de présenter les données à notre disposition et les traitements effectués.

5.1 Présentation de nos données et traitements

Les données à notre disposition sont issues des systèmes de gestion de Malakoff Humanis. Elles comprennent des sinistres prévoyance collective des entreprises du type TPE PME. L'historique des sinistres des portefeuilles en gestion directe du 01/01/2000 au 31/12/2021 s'y trouve.

Les données ont été traités en tenant compte des réglementations en vigueur sur l'utilisation des données dont le RGPD.

Des pré-traitements ont été effectués :

- Sommer les prestations payées pour une même clé de sinistre sur la même période d'indemnisation ;
- Supprimer les doublons de lignes de prestations ;
- Corriger les dates de fin de période d'indemnisation.

Nous nous intéressons aux sinistres arrêts de travail. Pour le risque incapacité, la mensualisation a été écartée de notre étude. Un croisement des données incapacités et invalidité permettra de répertorier les assurés qui étaient en incapacité maintenant passés en invalidité.

5.1.1 Risque Incapacité

Sur le périmètre des CCN, nous disposons de 502 139 lignes de sinistres 61 254 sinistres différents en incapacité survenus sur la période 2000 au 31 décembre 2021.

Pour chaque ligne de sinistre nous avons :

- La date à laquelle le sinistre est survenu ;
- La date de réception de la demande par Malakoff Humanis qui est la date de déclaration de sinistre ;

- La période d'indemnisation du sinistre ;
- Etat du sinistre (clôturé ou en cours de paiement) ;
- Le montant brut payé comme prestations sur la période l'indemnisation ;
- La date comptable du règlement de la période d'indemnisation ;
- Des informations sur l'assuré (Date de sinistre, numéro assuré, catégorie socioprofesionnelle,sexe,etc.) ;
- Des informations sur le contrat en question (la franchise, le produit d'assurance, la garantie, ...)
- La cause du sinistre ;

A partir de ces données, nous créons les variables suivantes :

- l'âge d'entrée en l'arrêt de travail de l'assuré ;
- la date de début de période d'indemnisation qui correspond au premier jour d'indemnisé ;
- La date de fin d'indemnisation qui correspond au dernier jour indemnisé pour un sinistre ;
- On détermine le nombre de jour indemnisé ;
- La durée de déclaration qui correspond à la différence entre la date de déclaration et la date de survenance du sinistre ;
- La durée de règlement qui correspond à la différence entre la date comptable et la date de déclaration ;
- Le montant de prestations reçues par l'assuré à la date comptable. Cette variable sera d'intérêt dans notre modèle d'apprentissage ;
- Une clé de sinistre composée de la concaténation du numéro de sinistre, du numéro assuré et de la date de survenance du sinistre ;
- Le montant total des prestations reçues par l'assuré pour un même sinistre ;
- Le montant de l'indemnisation journalière qui correspond au rapport de l'indemnisation totale perçue sur le nombre de jour indemnisé.
- Le trimestre de survenance ;

Montant des prestations

Nous avons décidé de supprimer les sinistres ayant comme total de l'indemnisation un montant négatif. Ces sinistres au nombre de 10 sont écartés de notre étude ils sont vraisemblablement des erreurs de gestion.

L'âge d'entrée en arrêt de travail

Dans nos données nous avons des assurés dont l'âge d'entrée en incapacité se situe entre 3 et

79 ans,très vraisemblablement des données aberrantes pour les valeurs extrêmes. Nous gardons les assurés dont l'âge d'entrée se trouve entre 18 et 66 ans.

Durée des sinistres

Pour le risque incapacité, bien que la durée limite théorique est de 1 095 jours, il arrive d'observe des sinistres qui durent plus longtemps. Ces sinistres peuvent s'expliquer par un retard des démarches administratives dans le cas d'un passage en AT. Les sinistres de plus de trois années ne seront pas considérés.

Les sinistres atypiques

Nous avons décidé de retirer les sinistres atypiques de notre base d'étude. En effet, ces derniers provoquent une volatilité plus accrue dans l'estimation de nos provisions. La variable fixant le seuil des graves sera le montant de l'indemnisation journalière en incapacité et le montant annualisé des rentes en invalidité. Le seuil des sinistres atypiques est fixée à 135 €. L'étude conduisant à la détermination de ce seuil de gravité sera explicitée en détail plus bas.

On se retrouve 58 247 sinistres soit 5,16% de sinistres initiaux écartés de l'étude. Ces sinistres graves représentent 16,34% des montants de sinistres totaux.

5.1.2 Risque Invalidité

Nous avons sur le périmètre des CCN 192 326 lignes de sinistres invalidité au 31/12/2021 pour 6 218 sinistres.
Nous disposons des variables suivantes pour chaque ligne de sinistre :

- La date à laquelle le sinistre est survenu ;
- La date de réception de la demande chez la mutuelle qui est la date de déclaration de sinistre ;
- La période d'indemnisation du sinistre ;
- état du sinistre (clôturé ou en cours de paiement) ;
- Le montant de rente payée à l'assuré sur la période en question ;
- La date d'octroi de la rente ;
- La date de survenance du sinistre ;
- La date comptable du règlement de la période d'indemnisation ;
- Des informations sur l'assuré (numéro assuré, catégorie socioprofesionnelle,sexe,...) ;
- Des informations sur le contrat (franchise, le produit, la garantie, ...) ;
- La cause de l'accident ;

Ces informations nous ont permis de créer les variables suivantes :

- la clé du sinistre qui sera composé du numéro assuré et de la date d'octroi de la rente ;

- une variable répertoriant par 1 les assurés de notre portefeuille qui sont passés de l'incapacité à l'invalidité et 0 sinon ;
- l'âge d'entrée en l'arrêt de travail de l'assuré ;
- le montant total de la rente versée à l'assuré pour un même sinistre ;
- le montant annualisé de la rente.

Age d'entrée

Pour le risque invalidité, l'âge d'entrée sera limité entre 18 et 61 ans comme stipulée par la règlementation.

Traitement Montant

Les sinistres ayant un montant total brut versé négatif ont été supprimés.

Période d'observation des sinistres

Nous avons retenu les arrêts survenus à partir 01/01/2000. Ainsi pour

Gestion des sinistres atypiques

Une étude des valeurs extrêmes sur le montant des prestations annualisés a permis de fixer le seuil des sinistres graves à 50 000€.

Nous nous retrouvons avec 190 347 lignes de sinistres, soit $1,03\%$ de données supprimées.

5.1.3 Prise en compte de l'inflation

Les prestations de nos données sont des prestations de base et ne comportent donc pas les revalorisations.Nous avons fait le choix de ne pas sélectionner les prestations totales comprenant les revalorisations car historiquement dans le contexte des CCN, les revalorisations n'étaient automatiques chaque année.

Il était fréquent que la revalorisation se fasse selon l'évolution de la valeur d'un point propre à la CCN. L'évolution de ce point était votée en Commission Paritaire. Il arrive que le point n'évolue pas pendant plusieurs années.

Cependant, ces dernières années la plupart des recommandations pour les conventions collectives ont opté pour une évolution selon la valeur du point de l'Agirc Arrco.

Pour éviter les disparités et garder nos prestations sur le même référentiel, nous décidons de conserver les prestations de base que nous revaloriserons.

Pour appliquer de l'inflation, nous proposons d'utiliser les travaux de la Direction de l'Animation de la Recherche, des Études et des Statistiques (Dares) [16]. La Dares propose des fiches pour la majorité des CCN avec le salaire moyen au fil des années. Tous les montants sont revalorisés sous le référentiel 2016.

Soit I_i l'indice pour l'année de survenance i. On considère que :

$$I_{2009} = 100$$

et pour l'année suivante i+1 on obtient :

$$I_{i+1} = I_i * (1 + r_{i+1,i})$$

Le taux d'inflation entre une année m et une année k est donnée par :

$$r_{m,k} = \frac{Salairemoyen_m}{Salairemoyen_k}$$

Ainsi pour ce sinistre i, on déterminer le montant de prestations inflaté :

$$P_i^{inflat}(m,k) = \frac{I_{2018}}{I_{m+k}} * P_i(k,m)$$

Dans notre cas d'étude, quand nous étudierons un cas de as-if sur nos triangles pour voir l'impact de l'inflation.

5.2 Détermination du seuil séparant des sinistres graves

La gravité n'est pas toujours la même dans un portefeuille. Aussi, plus un sinistre est grave plus il a du poids au niveau du coût du sinistre. Ces sinistres dits lourds faussent la modélisation de nos coûts. Il est donc nécessaire de modéliser séparément nos sinistres graves des autres sinistres qui n'ont pas le même comportement.

Nous déterminerons dans cette partie à partir de quel montant un sinistre peut-être considérer comme atypique.

L'approche retenu pour analyser ces sinistres dits atypiques consiste à s'intéresser au comportement des sinistres à partir d'un seuil. Pour déterminer ce seuil nous aurons recours à la théorie des valeurs extrêmes mais aussi à une analyse des quantiles.

Le montant des indemnités journalières dépend du niveau de revenus de l'assuré, ce sera la variable que nous utiliserons pour caractériser les sinistres atypiques dans le cas de l'incapacité. En invalidité, nous nous intéresserons au montant la rente annuelle.
Désignons par J la variable aléatoire qui caractérise nos sinistres,i.e le montant des indemnités journalières en incapacité ou le montant de la rente en invalidité.

Dans la suite nous considérons $J_1, \ldots, J_n$ une suite de n variables aléatoires indépendantes et identiquement distribuées, avec pour même fonction de répartition F où n est la taille des données .

Nous rappellerons tout d'abord des notions importantes de la Théorie des Valeurs Extrêmes et les méthodes sur la détermination d'un seuil.

5.2.1 Rappel de la Théorie des Valeurs Extrêmes

L'objectif derrière cette théorie est d'étudier le comportement asymptotique des queues de distrubution.

Considérons nos n variables aléatoires iid $J_1, \ldots, J_n$ et notons $F = P(J_i \leq J)$ et $M_n = \max(J_1, \ldots, J_n)$. Etudier la loi du maximum M_n nous donne le comportement de la queue de notre distribution.

Le théorème de Fisher-Tippet nous définit la loi asympotique de ce maximum M_n.

Théorème de Fisher-Tippet

F appartient au domaine d'attraction de H_ζ si et seulement si :

$$\exists (a_n) > 0 et (b_n) telque \forall x \in \mathbb{R}, \lim_{n \to +\infty} F^n(a_n x + b_n) = H_\zeta(x)$$

$$\iff \exists (a_n) > 0 et (b_n) telque \frac{M_n - b_n}{a_n} \longrightarrow_{n \to +\infty} Z$$

, avec Z qui a pour fonction de répartition H.

$H_\zeta(x)$ est la fonction de répartition des valeurs extrêmes généralisées, appelée GEV (Gene-

ralized Extreme Value). On a :

$$H_\zeta(x) = \begin{cases} si\zeta \neq 0, \exp[-(1+\zeta x)_+^{\frac{-1}{\zeta}}] \\ si\zeta = 0, \exp(-e^{-x}) \end{cases}$$

La nature de la distribution est déterminée à partir de la valeur du shape parameter ζ :

- Si $\zeta > 0$, On a une distribution Fréchet ou à queue lourde à valeurs positives ;
- si $\zeta = 0$: la distriution est de Gumbel ou à queue légère ;
- Si $\zeta < 0$: la distribution est de Weibull ou à queue lourde à valeurs négatives ;

Nous résumons cela dans le tableau ci-dessous :

ζ	Domaines d'attractions	Exemples
> 0	Fréchet	Cauchy ; Pareto
$= 0$	Gumbel	Gaussien ; Gamma
< 0	Weibull	Uniforme ; Béta

TABLE 5.1 – Récapitulatif des domaines d'attraction

Cette approche sur la GEV entraine une perte d'information contenue dans les autres grandes valeurs de l'échantillon si celles ci sont inférieures au maxima.

Nous envisagerons à la place d'un maximum de considérer un seuil. Pour ce faire, nous aborderons la méthode de Peak Over Thresholds (POT) proposée par Pickands (1975)[17].

Méthodes Peak Over Threshold

Pour ces méthodes, nous nous intéressons aux valeurs extrêmes à partir d'un certain seuil et non simplement le maximum M_n. Il faut donc observer toutes les valeurs prises par J au delà d'un seuil u élevé. La difficulté de cette méthode consiste à choisir ce seuil u.

On note $Y = J - u$. La fonction de répartition des excès au-delà du seuil u est définie par :

$$\forall y \geq 0, F_u(y) = P(X - u < y | X > u) = \frac{F(u+y) - F(u)}{1 - F(u}$$

F est approximée par $G_{\zeta,u}$ pour u assez grand en concordance au théorème de Pickands où :

$$G_{\zeta,\sigma(u)}(y) = \begin{cases} \zeta \neq 0, 1 - (1 + \zeta * \frac{y}{\sigma(u)})^{\frac{-1}{\zeta}}] \\ \zeta = 0, 1 - \exp(-\frac{y}{\sigma(u)}) \end{cases}$$

5.2.2 Méthodes de détermination du seuil

Nous rechercherons en premier lieu le domaine d'attraction de la loi sous-jacente de nos observations. Pour cela, nous avons recours à un QQ-plot exponentiel grâce à l'estimateur :

$$H_{k,n} = X_{n-k,n} * (\frac{1}{k}\sum_{i=1}^{k} log(X_{n-i+1,n}) - log(X_{n-k,n}))$$

La courbe est croissante ce qui permet de conclure que sous-jacente de nos observation appartient bien au domaine d'attraction de Fréchet.

Nous poursuivrons en déterminant le seuil à travers les méthodes de l'excès moyen, l'estimateur de Hill et la méthode des écarts inter-quantiles.

Mean Excess

Cette méthode permet de déterminer notre seuil u.La fonction $e(u) = \mathbb{E}[Y|J > u]$ est appelée la fonction des excès moyen.Son estimation empirique est définie par :

$$\widehat{e_n}(u) = \frac{\sum_{i=1}^{n}(J_i - u)_+}{\sum_{i=1}^{n} \mathbb{1}_{J_i>u}}$$

Pour déterminer notre seuil adéquat u, nous traçons le mean excess plot de J, c'est-à-dire l'ensemble des points (u,$\widehat{e_n}(u)$, puis nous choisirons le seuil u de façon à ce que $\widehat{e_n}(u)$ soit quasi-linéaire pour $x \geq u$.

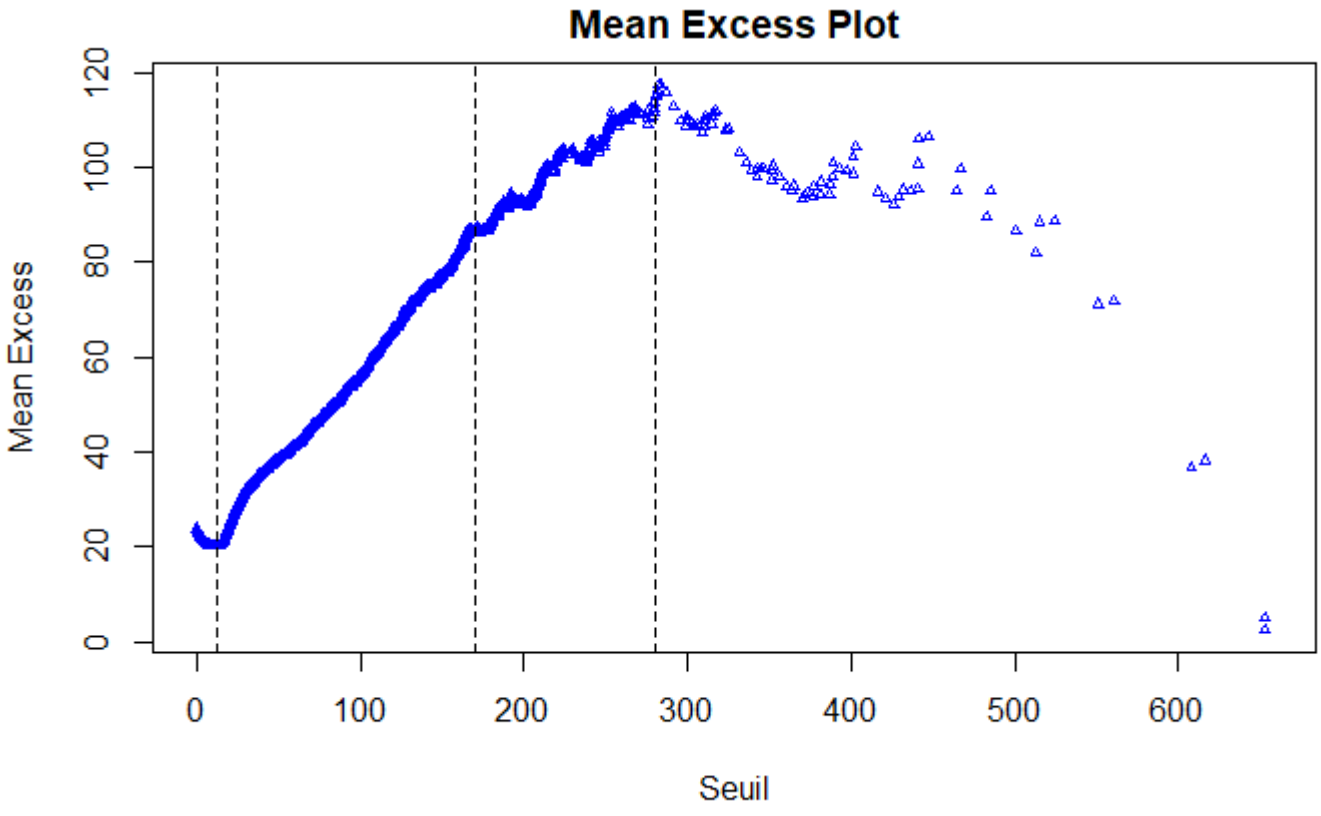


FIGURE 5.1 – Estimation des excès moyen - Incapacité

Pour le risque incapacité, les trois seuils identifiés correspondent chacun à une entrée dans une zone de stabilité à savoir 13€, 170€ et 280€.

Cependant, le seuil fixé à 13€ est inférieur à la moyenne des indemnités journalières. Nous jugeons que ce seuil est trop faible pour caractériser nos sinistres graves.

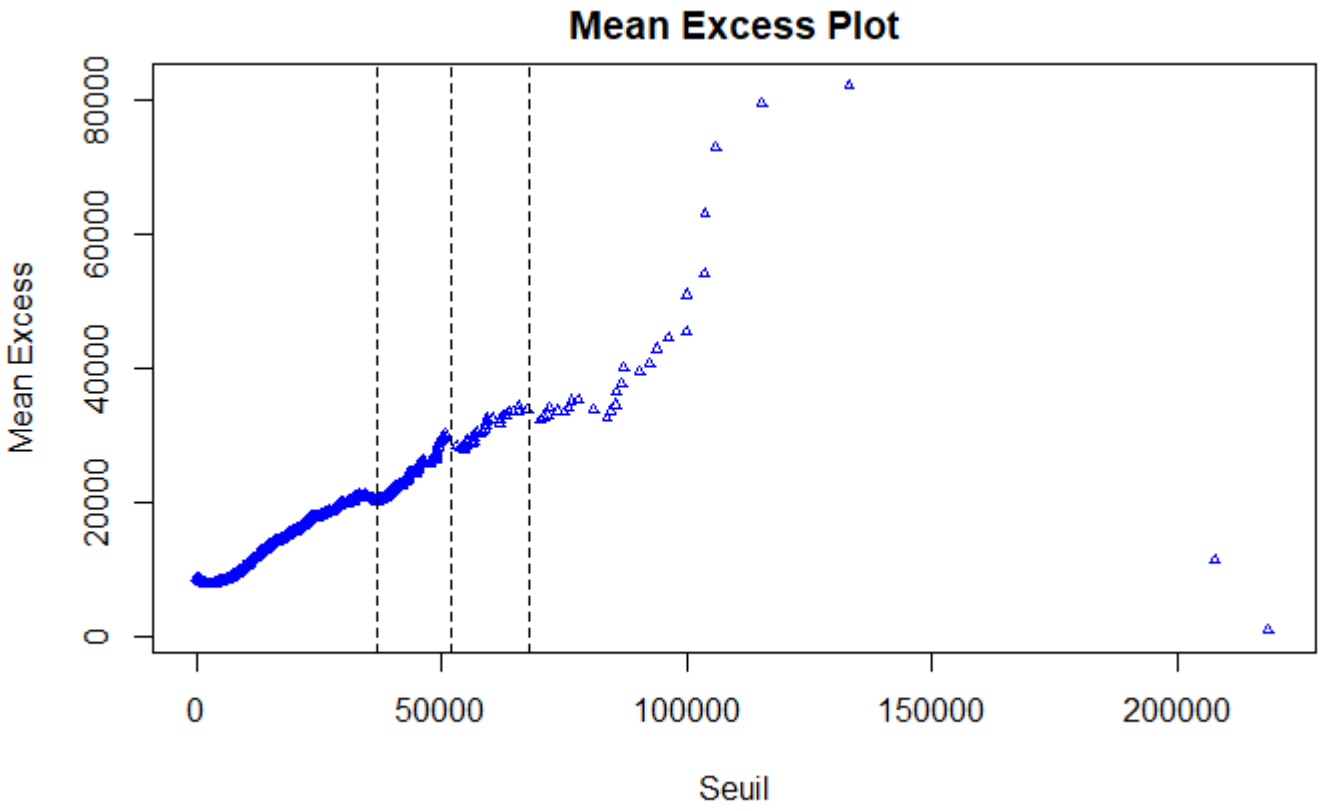


FIGURE 5.2 – Estimation des excès moyen - Invalidité

Pour le risque invalidité, on observe trois seuils pour les montants de rente annualisées pour 37000€, 53 000€ et 68 000€.

L'estimateur de Hill

Autre méthode répandue, elle se base sur les statistiques d'ordre de nos observations. On considère $X_{1,n}, \ldots, X_{n,k}$ nos statistiques d'ordre.

Il s'agit de sélectionner graphiquement le nombre d'excès au-delà duquel la valeur de l'indice de queue devient stable, c'est-à-dire quand l'estimation devient plus robuste. L'estimateur de Hill est donné par :

$$\widehat{\zeta}^{hill}(u) = \frac{1}{k}\sum_{i=1}^{k} \ln\left(\frac{X_{n-i+1,n}}{X_{n-k,n}}\right)$$

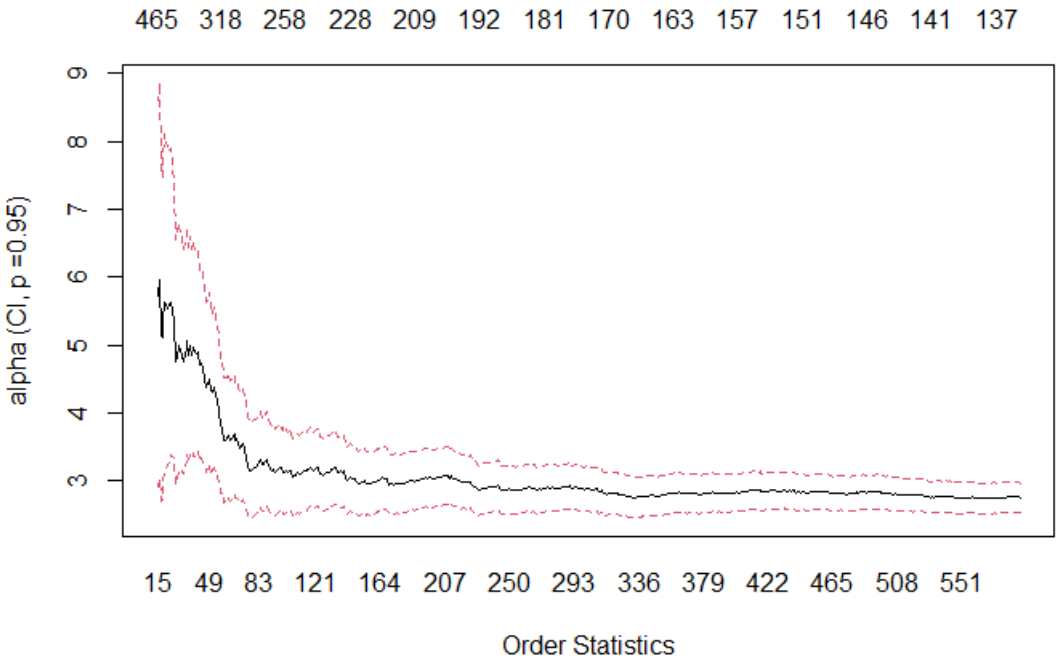


FIGURE 5.3 – Estimation de Hill - Incapacité

Pour l'incapacité, on a une stabilisation pour un seuil entre 180€ et 170 €.

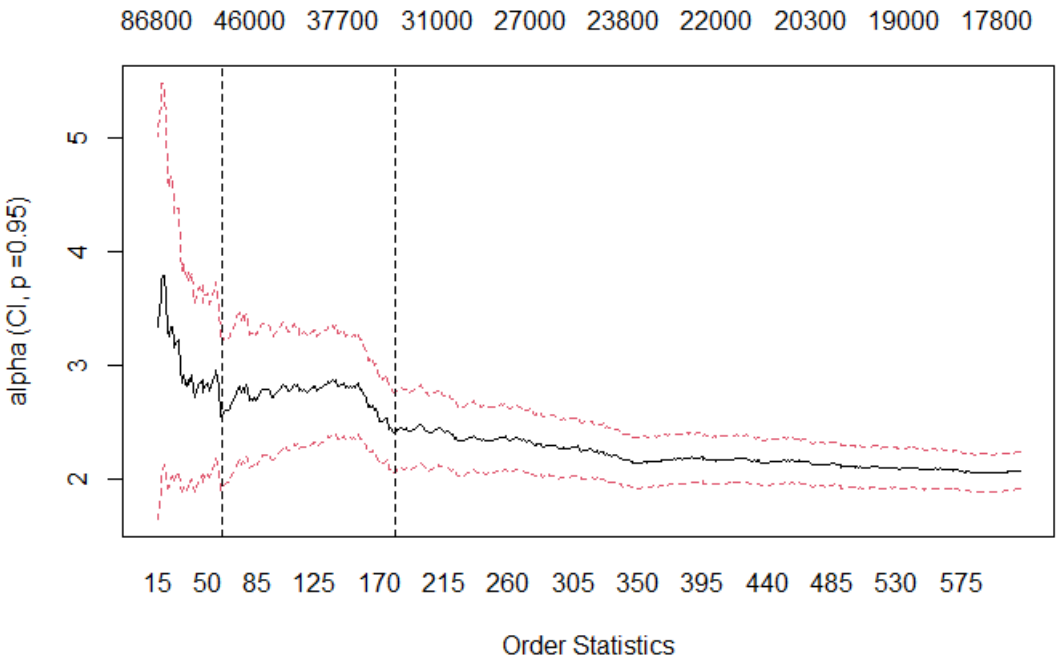


FIGURE 5.4 – Estimation de Hill - Invalidité

Nous identifions trois seuils pour le risque invalidité 35 800 €, 50 300 € et 64 300€.

La méthode des écarts inter-quantiles

Pour détecter nos pics de variations, nous utiliserons l'écart inter-quantile. Elle indique une plage d'appartenance du seuil des graves.

Nous étudierons les écarts de quantiles entre 90% et 100% par des pas de 0.5%

Niveau (%)	95	95,5	96	96,5	97	97,5	98	98,5	99
Quantile	70,5	74,17	78,18	83,05	88,72	95,65	104,56	116,44	135,4
Ecart IQ (%)		5.2	5.4	6.2	6.8	7.81	9.3	11.4	16.3

TABLE 5.2 – Variations interquantiles- Incapacité

Le passage du quantile à 98% au quantile à $98,5\%$ dénote une variation de $11,3\%$ qui est environ 3 points au-dessus des précédents écarts. Le même constant est observable pour le passage du quantile à $98,5\%$ au quantile à 99%. Il est ainsi considéré que le seuil des graves par cette méthode est compris entre 104,56€ et 135,4€.

Niveau (%)	95	95,5	96	96,5	97	97,5	98	98,5	99
Quantile	25537.6	26856.8	28452.5	305230	32670.5	37126.6	40076.3	44087.2	50012.5
Ecart IQ (%)		5.2	5.9	7.3	7	13.6	7.9	10	13.4

TABLE 5.3 – Variations interquantiles- Invalidité

Le passage du quantile à 97% au quantile à $97,5\%$ affiche un écart de $13,6\%$ qui est environ 6 points au-dessus des précédents écarts. Le même constant est observable pour le passage du quantile à $98,5\%$ au quantile à 99%. Dans ce cas le seuil des sinistres graves est compris entre 32 670€ et 50 012€.

5.2.3 Choix de notre seuil

Nous reprenons les seuils identifiés suivant les méthodes.

Incapacité

MEP	Hill	IQ
13		
		104
170	175	135
280		

TABLE 5.4 – Choix du seuil

Nous prenons comme seuil 135 € qui correspond au minimum des trois méthodes implémentées. On obtient 1% des sinistres comme graves par rapport à ce seuil ce qui représentent 1% du cout global.

Invalidité

MEP	Hill	IQ
37 000	35 800	32 700
53 000	50 300	50 000
68 000		

TABLE 5.5 – Choix du seuil

Nous prenons comme seuil 50 000 € qui correspond au minimum des trois méthodes implémentées. On obtient 1% des sinistres comme graves par rapport à ce seuil ce qui représentent 1% du cout global.

5.3 Statistiques élémentaires et choix des variables

5.3.1 Incapacité

Le nombre de sinistres incapacité par année de survenance est en constante évolution sauf pour l'année 2021. La date de fin d'observation étant fixée au 31/12/2021 et le délai de déclaration des sinistres étant souvent long peuvent expliquer ce retard.

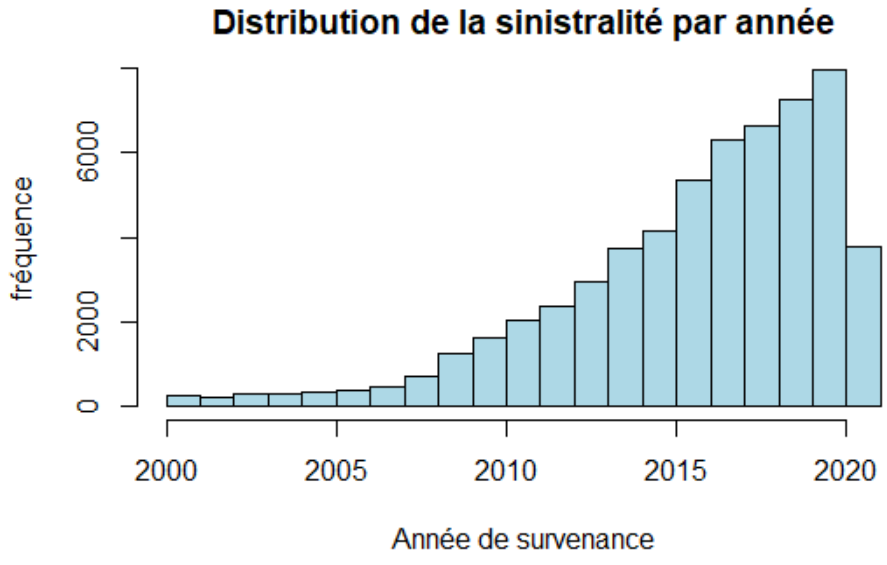


FIGURE 5.5 – Distribution de la sinistralité par année

Concernant le délai de déclaration, qui correspond à la différence de la date de réception et celle de survenance, on assiste à une réduction drastique de ce délai même si en 2007 et 2008 sont atypiques par rapport à la tendance globale (fusion Malakoff et Médéric sûrement). Il faut rappeler que l'assuré n'est pas tenu de déclarer son sinistre avant la fin de la franchise.

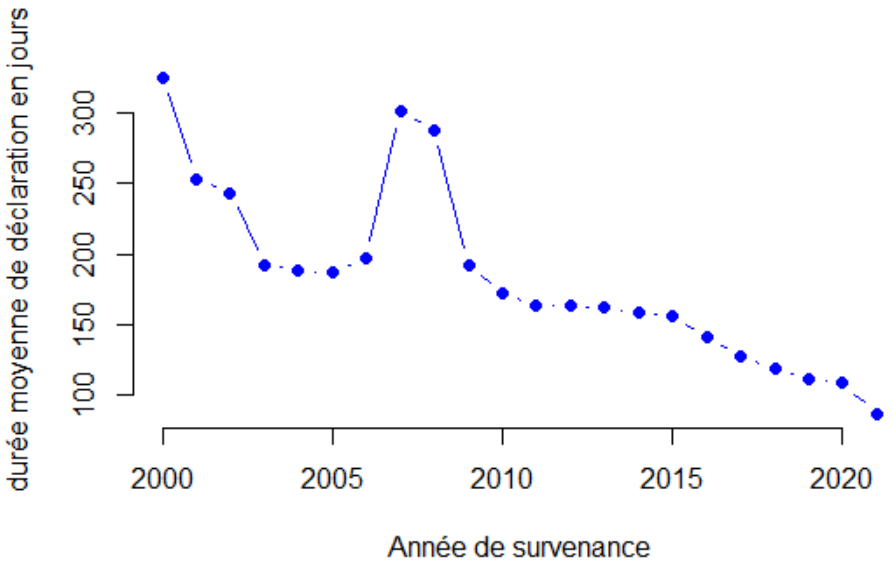
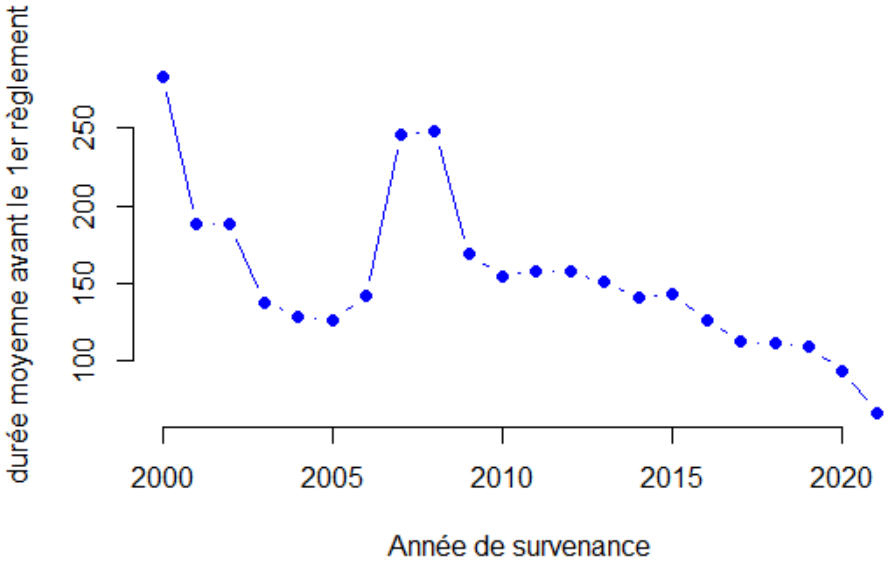


FIGURE 5.6 – Durée moyenne de déclaration en jours par année de survenance

La durée de règlement qui elle est obtenue en faisant la différence entre la date du premier versement et la date de début de droits elle a également diminué au fil des années. Les années 2007 et 2008 sont à nouveau atypiques. Nous ne prenons pas 2007 et 2008 dans notre étude de cas pour éviter de biaiser notre étude.



Nous avons très peu de sinistres avant l'âge de 26 ans et après 60 ans. Ceci s'explique par le passage en retraite pour les grands âges et pour les jeunes par le fait qu'il rentre dans la vie active généralement à 23ans.

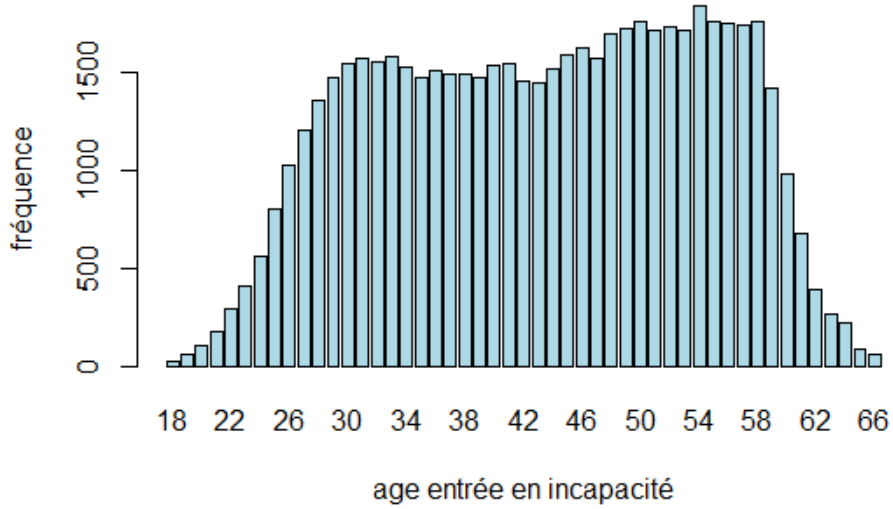


FIGURE 5.7 – Fréquence d'entrée en incapacité par âge

La durée moyenne en incapacité en mois des arrêts non censurés est croissante jusqu'à l'âge de 57 ans avant de décroître tandis que celle des sinistres censurés est très volatil.

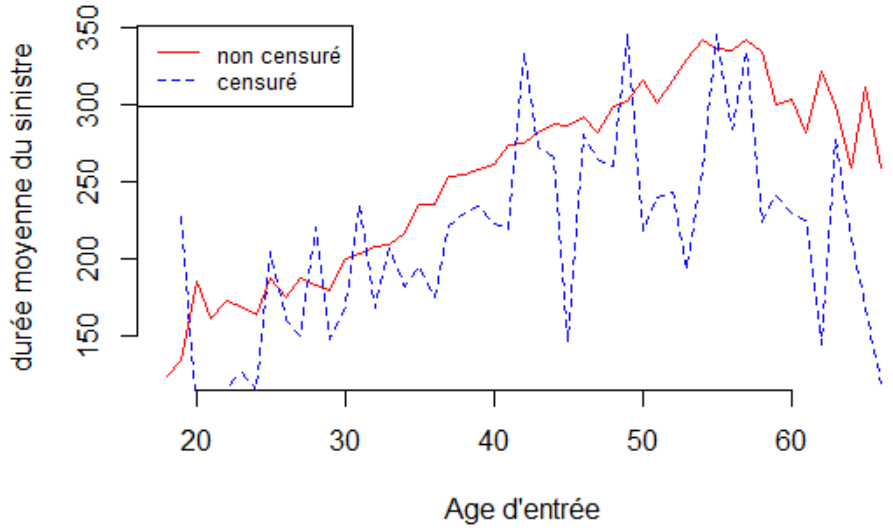


FIGURE 5.8 – Durée moyenne des sinistres par âge

5.3.2 Invalidité

Les sinistres invalidité qui étaient précédemment en incapacité dans le portefeuille de représentent près de 70% de nos arrêts invalidité. Nous avons donc presque 30% de nos arrêts en invalidité qui proviennent de reprise de passifs.

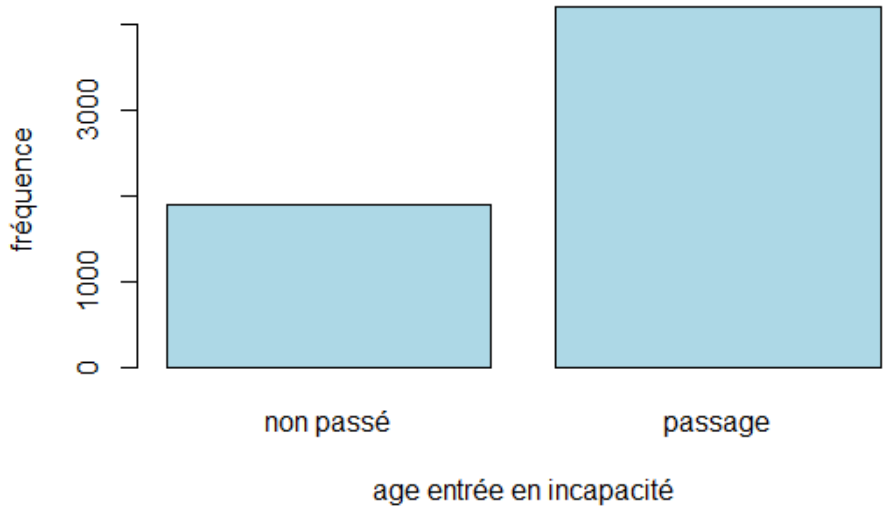


FIGURE 5.9 – Nombre de passage en invalidité

Sur la répartition par âge , 75% de la population a 46 ans et plus. Le minimum se situe à 23 ans et les assurés entre 23 et 35 ans ne représentent que $5,2\%$.

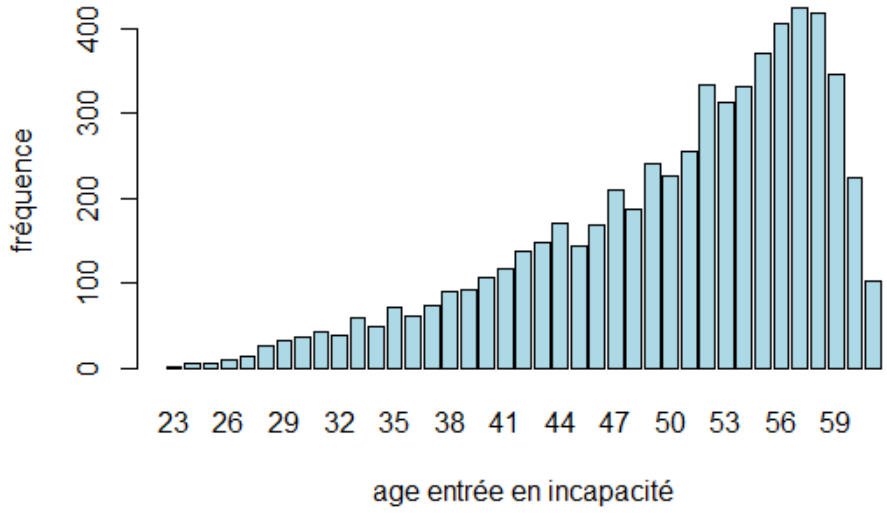


FIGURE 5.10 – Répartition par âge en invalidité

Dans notre portefeuille, la part des sinistres en cours est de $77,2\%$.

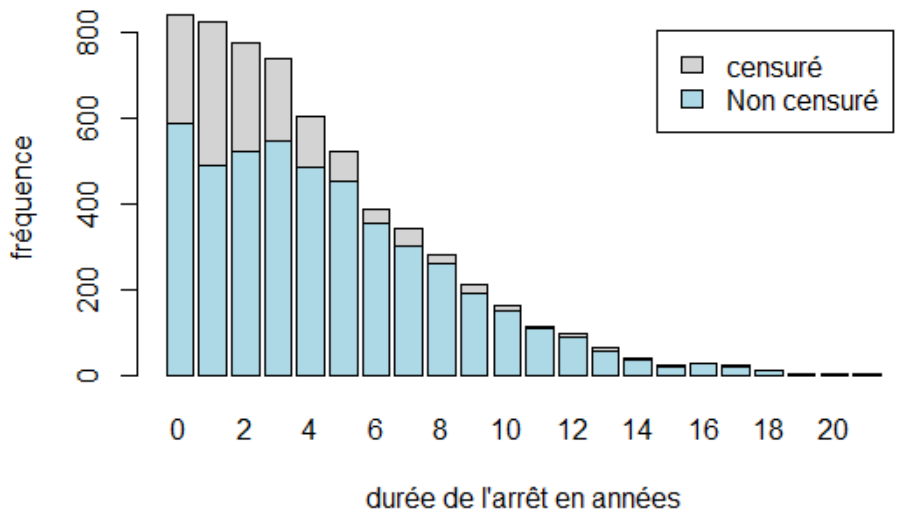


FIGURE 5.11 – Durée en invalidité

5.3.3 Etude des corrélations

Nous étudierons la corrélation entre nos variables dans ce point. Cela nous permet de déterminer l'intensité de la liaison qui peut exister entre ces variables

Cette mesure d'association est faite à grâce au V de Cramer. Nos variables n'étant pas toutes qualitatives, nous avons effectué des retraitements afin de regrouper les variables quantitatives en classes.

Rappelons que le V de Cramer se base sur le χ^2 de Pearson. Considérons les notations suivantes : représentant respectivement le nombre de lignes e cl t de colonnes du tableau de contingence. n le nombre de données contribuant au calcul des valeurs du khi2

- l réprésente le nombre de sinistres ;
- p le nombre de variables descriptives ;
- n représente le nombre de données contribuant au calcul des valeurs du ^{2}.

$$V_{Cramer} = \sqrt{\frac{\chi^2}{n * [(l \wedge p) - 1]}}$$

Plus V est proche de zéro, plus il y a indépendance entre les deux variables étudiées. Il vaut 1 en cas de complète dépendance puisque le ^{2}est alors égal au^{2}max (dans un tableau 2×2, il prend une valeur comprise entre -1 et 1).

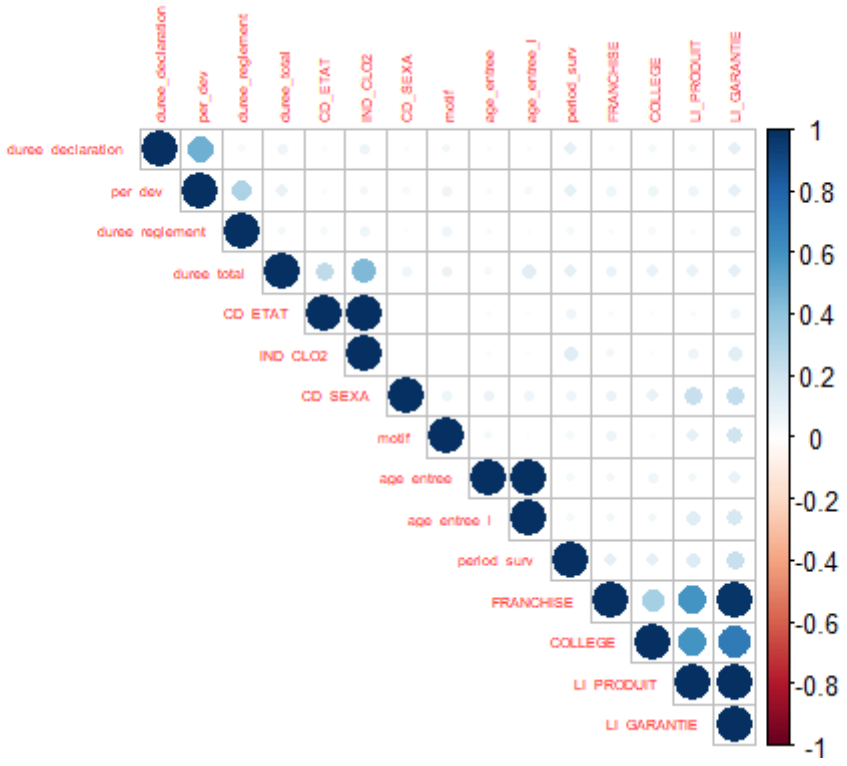


FIGURE 5.12 – Matrice de V de Cramer

Nous observons que le V de Cramer entre la variable franchise du contrat et la variable libellée de garantie est proche de 1. Ces deux variables sont fortement liées. Le type de garantie a donc une influence la franchise. Cela est en cohérence avec les observations faites précédemment. Pour les autres variables testées, nous n'identifions pas de liaisons fortes entre les variables, le V de Cramer étant largement inférieur à 0,5.

Conclusion du chapitre

Dans les parties qui suivront, nous présenterons et analyserons les résultats obtenus avec nos différents modèles de prédictions.

Pour l'expérimentation, nous nous intéresserons à un regroupement de CCN ayant des caractéristiques à peu près similaire. Aussi, ce regroupement qui représente une grande part du portefeuille en prévoyance du de Malakoff Humanis.

Pour le risque incapacité, notre objectif sera de déterminer au mieux la charge ultime de nos sinistres. Pour ce qui est du risque invalidité, nous tenterons de prédire le nombre de sinistres tardifs par année de survenance.

Nous analyserons nos résultats par risque d'abord pour le risque incapacité et ensuite le risque invalidité.

Chapitre 6

Risque Incapacité

Nous avons décidé pour notre application de provisionnement de considérer un regroupement de CCN aux mêmes caractéristiques, ce qui prouvera la robustesse des méthodes et leur précision. Nous considérons donc les arrêts de travail de notre cas d'étude survenus entre le 01/01/2009 et 31/12/2016. Les montants de nos prestations sont anonymisés dans des soucis de confidentialité.

Sur la période d'observation nos sinistres ont la particularité d'avoir la même franchise, 90 jours. Ce sera l'exemple parfait pour étudier le pouvoir prédictif de nos méthodes de provisionnement sur les contrats à franchise longue, qui constituent la majorité des contrats prévoyance.

Le choix a donc été fait de considérer l'ancienneté en années comme notre période de développement.

$$\text{Periode developpement} = AA_{COMPTABLE} - AA_{SURVENANCE}$$

Ce choix a été fait car sur les contrats à franchise longue, le paiement d'un sinistre i l'assuré doit souvent fournir des justicatifs de santé qui prouvent qu'il est toujours en arrêt de travail avant de percevoir ses indemnités. Cela occasionne souvent des retards comptables.

6.1 Applications de nos méthodes de provisionnement classiques

Dans cette partie, nous présenterons les résultats du calcul des réserves par les méthodes de provisionnement classiques. Nous comparerons les montants de provision estimés pour les sinistres ayant une année de survenance entre 2009 et 2016. Nous nous plaçons à la date du 31/12/2016 pour appliquer nos méthodes de provisionnement. Nos résultats seront comparés aux montants des réserves réelles observées au 31/12/2021.

Ainsi au 31/12/2016, le triangle de charges cumulés se présente comme suit :

Année	0	1	2	3	4	5	6	7
2009	539880	3322702	4799527	5288 295	5339199	5339199	5339 819	5339819
2010	1041102	4191579	5 651 715	6 205 193	6 206 383	6 207 303	6 208 461	
2011	1342106	4975146	6 526 649	7 189 812	7 236 881	7 236 881		
2012	1558725	6548404	8 682 451	9 418 043	9 439 388			
2013	2080096	7968018	10 401 906	11 206 532				
2014	2766120	9702486	12 949 394					
2015	3501842	11728892						
2016	3295826							

TABLE 6.1 – Triangle des charges cumulés cas d'étude

6.1.1 Résultats de la méthode de Chain-Ladder

Avant de pouvoir appliquer la méthode de Chain-Ladder sur nos données, il convient de vérifier ses hypothèses. Ce travail est présenté en annexe du document.

La réserve de provisions est estimée est de 21 593 446€ pour ce portefeuille de sinistres. Comme énoncé dans la partie théorique, elle est obtenue en déduisant les montants connus au 31/12/2016 des ultimes projetés.

En récapitulatif nous avons comme charge :

Période de survenance	Provision réelles	Provisions Chain-Ladder
2009	0	0
2010	0	0
2011	24 887	1 115
2012	10 488	1 916
2013	17 055	50 342
2014	947 005	1 228 861
2015	4 744 276	5 417 189
2016	13 138 527	14 894 023

TABLE 6.2 – Estimation des provisions Chain-Ladder

La provision via la méthode Chain-Ladder surestime au global les provisions de 14,4%. La captation par cette méthode n'est pas idéale.

6.1.2 Résultats de la méthode de Schnieper

Nous avons le triangle cumulé suivant auquel on associe l'exposition. Celui-ci se décompose en la matrice N et D.

Survenance	$N_{i,1}$	$N_{i,2}$	$N_{i,3}$	$N_{i,4}$	$N_{i,5}$	$N_{i,6}$	$N_{i,7}$	$N_{i,8}$
2009	539 880	1 280 565	750 003	259 935	30 104	0	1 244	0
2010	1 041 102	1 301 899	773 260	379 139	936	0	2 326	
2011	1 342 106	1 482 593	750 688	488 926	41 586	0		
2012	1 558 725	1 765 720	840 110	371 189	19 438			
2013	2 080 096	2 033 890	1 153 465	477 353				
2014	2 766 120	2 601 319	1 471 914					
2015	3 501 842	2 849 782						
2016	3 295 826							

TABLE 6.3 – Triangles des sinistres nouveaux $N_{i,j}$

Survenance	$D_{i,1}$	$D_{i,2}$	$D_{i,3}$	$D_{i,4}$	$D_{i,5}$	$D_{i,6}$	$D_{i,7}$	$D_{i,8}$
2009	0	- 1 502 258	- 726 82	- 228 833	- 20 800	0	624,1	0
2010	0	-1 848 577	-686 877	-174 339	-254	-920	1 167	
2011	0	-2 150 448	-800 814	-174 237	-5 483	0		
2012	0	-3 223 959	-1 293 937	-364 403	-1 907			
2013	0	-3 854 033	-1 280 423	-327 272				
2014	0	-4 335 047	-1 774 994					
2015	0	-5 377 268						
2016	0							

TABLE 6.4 – Triangle des variations $D_{i,j}$

Les estimateurs d'erreurs pour chaque période de développement sont données par

j	1	2	3	4	5	6	7	8
$\widehat{\lambda}_j$	8,24	8,06	0,53	0,37	0,05	0	0	0
$\widehat{\delta}_j$	-1,73	-0,31	-0,08	$-3{,}2\times10^{-3}$	$-4{,}9\times10^{-5}$	$-1{,}54\times10^{-4}$	0	

TABLE 6.5 – Estimations des paramètres de Schnieper

Les réserves sont les suivantes :

La réserve totale est de $\widehat{R}$= 18 117 365 ce qui représente un écart de -4,2% par rapport aux provisions réelles. On a une meilleure captation des provisions et prestations futurs à partir de 2013 qu'avec la méthode de Chain-Ladder en plus d'avoir une séparation des sinistres non déclarés. L'utilisation des expositions est très intéressantess pour déterminer nos provisions. Nos provisions estimées sont néanmoins inférieures aux réserves réelles ce qui nous déplait.

i	$\widehat{C}_{i,n}$	$\widehat{R}_i$
1	5 339 819	0
2	6 208 461	0
3	7 236 675	- 206
4	9 439 401	13
5	11 247 077	40 546
6	13 951 970	1 002 577
7	16 178 782	4 449 890
8	15 920 371	12 624 545

TABLE 6.6 – Résultats provisions de Schnieper

6.2 Modèle de provisionnement individuel

Dans cette partie, nous illustrons notre méthode de provisionnement individuel sur le cas d'étude d'arrêt incapacité présentée plus haut.

Nous utiliserons l'algorithme d'apprentissage XGBoost décrit les parties précédentes pour l'apprentissage statistique. Dans son article original, Baudry a utilisé le paquet extraTrees, une variante des forêts aléatoires. Nous avons opté pour le XGBoost à cause de sa robustesse.

Le calibrage de notre modèle de XGBoost passera par :

- ajuster les hyperparamètres afin de sélectionner des hyperparamètres xgboost plus optimaux ;
- sélectionner les variables qui expliquent au mieux nos sinistres. Les variables utilisées pour calibrer notre apprentissage permettent de gagner en performance dans la construction de nos arbres boostés. Ainsi le gain d'une variable augmente avec son utilisation dans le processus d'apprentissage.

Pour les trois modélisations à suivre, le choix du nombre optimal d'arbres (nrounds) a été fait en utilisant une cross-validation en appelant la fonction xgb.cv avec qui s'arrête au nombre optimal.

6.2.1 Résulats de l'apprentissage des RBNS

Nous avons considéré que chaque sinistre a une durée d'existence de six ans au maximum. Les sinistres ayant déjà eu lieu et clôturés sont utilisés dans notre base d'entrainement. Ceux toujours en cours dont nous voulons évaluer la charge totale seront marqués en cours jusqu'à la fin la 6e année probable de traitement du sinistre.
L'inspection du coût des montants pour nos données de test ci-dessous nous permet de choisir la fonction objective "reg :tweedie".

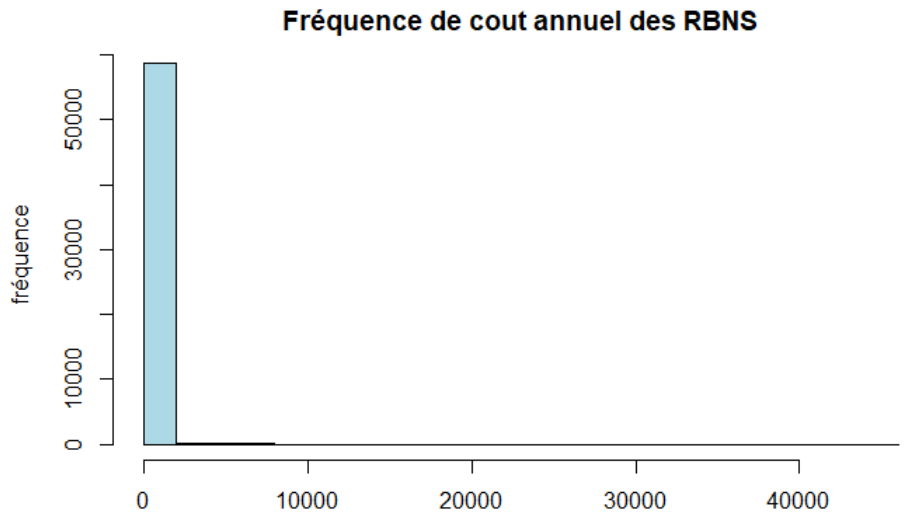


FIGURE 6.1 – Fréquence des coûts annuels RBNS

Les prédictions de notre modèle sont représentées dans le tableau si dessous :

Survenance	Prédiction	Réalité
2012	3 437	10 488
2013	1 893	1 440
2014	570 969	933 762
2015	4 798 191	4 520 625
2016	5 788 960	7 708 336

TABLE 6.7 – Prédictions des RBNS

Notre estimation est en deçà des règlements, elle prévoit 11 163 450,22€ soit un écart de -15,4%.
Une limite d'estimer nos sinistres incapacité est nous ne provisionnons pas le risque invalidité sous-jacent probable.

Sur cette application les variables les plus présentes sont principalement le moment cumulé des prestations déjà versées et la période de développement j par rapport à notre date de provisionnement.

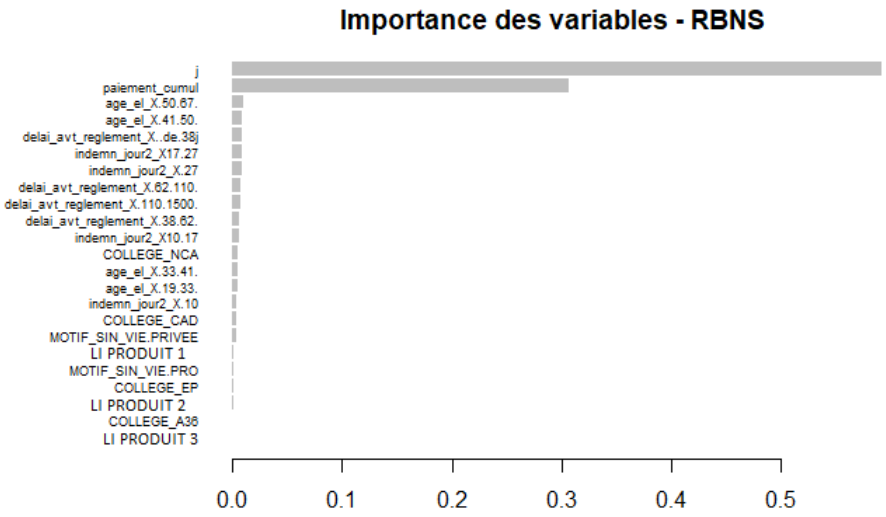


FIGURE 6.2 – Variables d'intérêt RBNS

6.2.2 Résultats de l'apprentissage des IBNR

Pour estimer nos IBNR, nous optons pour une méthode fréquence/sévérité. Nous considérons les sinistres comme étant à paiement unique. Cette dernière sera la date du dernier paiement. Nous nous plaçons au 31/12/2016 comme date de provisionnement.

Nos données d'entraînement sont formées des sinistres tardifs observés après les différentes dates de provisionnement entre 2009 et 2015.

Pour établir un modèle rigoureux, nous rappelons que nous simulons des sinistres fictifs sur l'hypothèse des données de population dans notre portefeuille.

Pour notre modèle de fréquence, nous choissisons comme fonction objective "count :poisson" sur la base de la fréquence des sinistres.

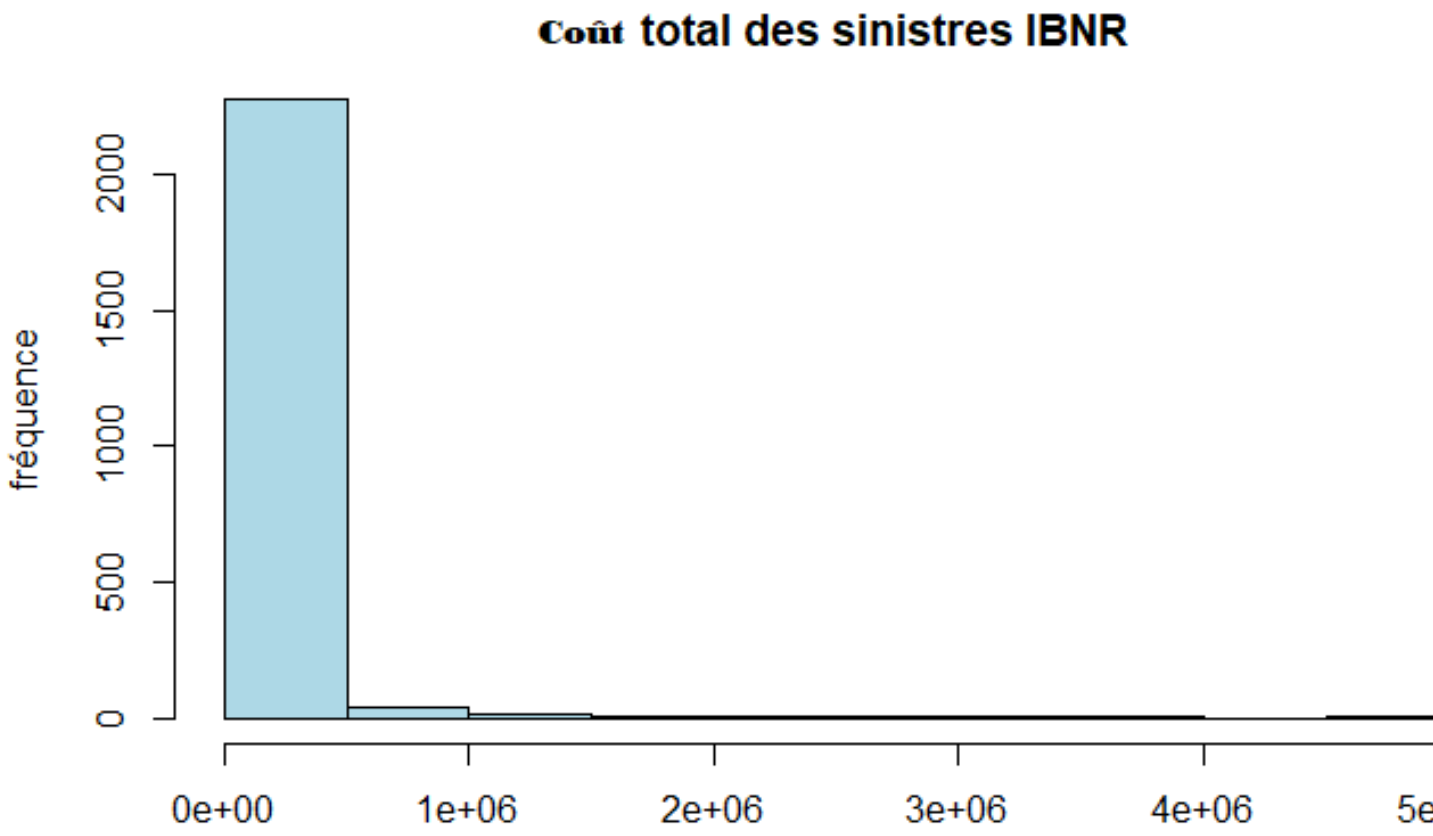


FIGURE 6.3 – Fréquence des IBNR

Nous obtenons un écart de -8% qui à la vue de nos considérations et hypothèses est un résultat encourageant.

Survenance	Prédiction	Réalité
2013	3	2
2014	5	4
2015	44	27
2016	464	526

TABLE 6.8 – Prédictions de la fréquence

Les variables significatives de notre modèle de fréquence sont ici :

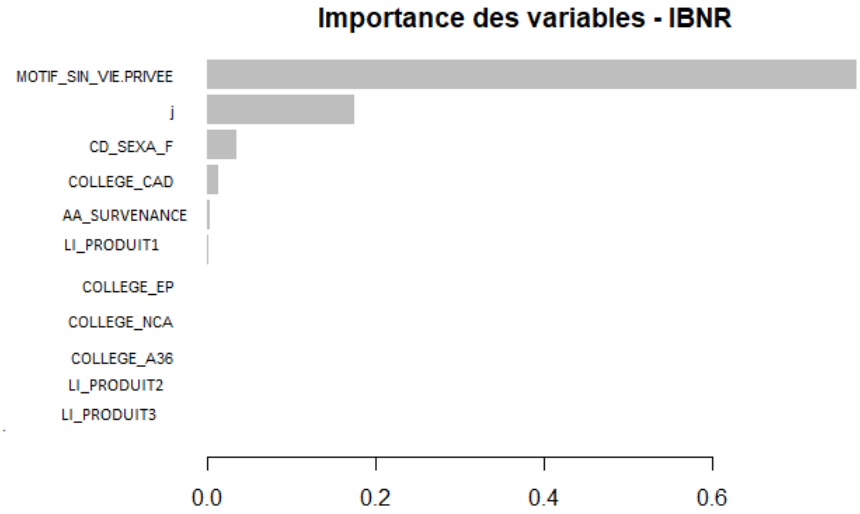


FIGURE 6.4 – Fréquence des coûts annuels RBNS

Pour notre modèle de sévérité nous obtenons Comme pour la fréquence de survenance des

Survenance	Prédiction	Réalité
2011	0	24 887
2013	38 377	15 615
2014	46 152	13 243
2015	201 551	223 872
2016	3 885 338	5 430 191

TABLE 6.9 – Prédictions de la sévérité

IBNR, nous avons choisi la fonction objective "reg :tweedie" à partir de la distribution des coûts totaux des IBNR.

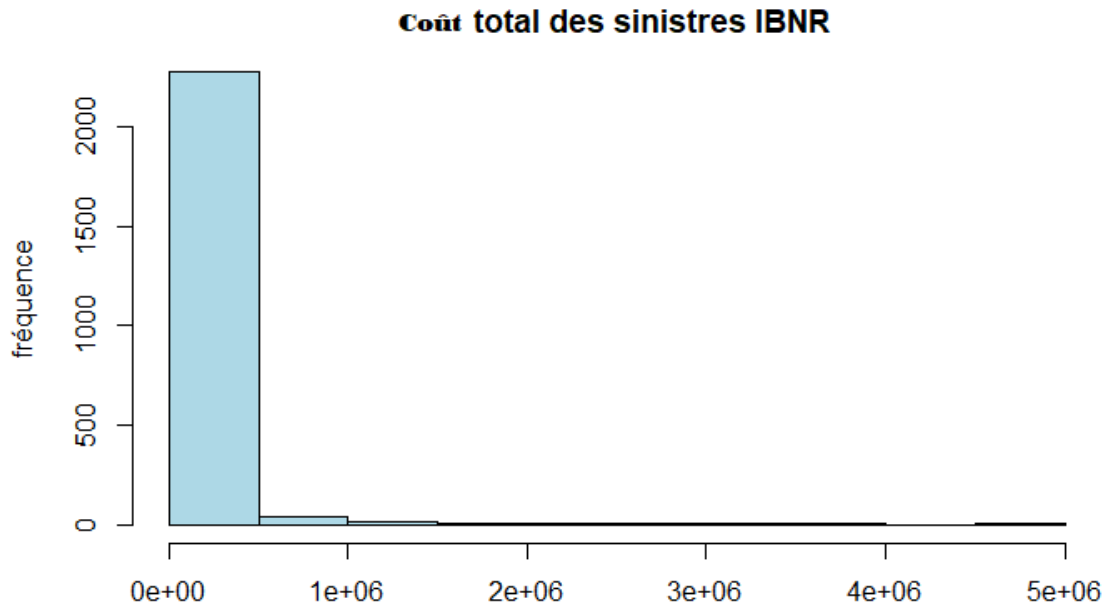


FIGURE 6.5 – Fréquence des coûts IBNR

Les variables significatives pour notre algorithme sont :

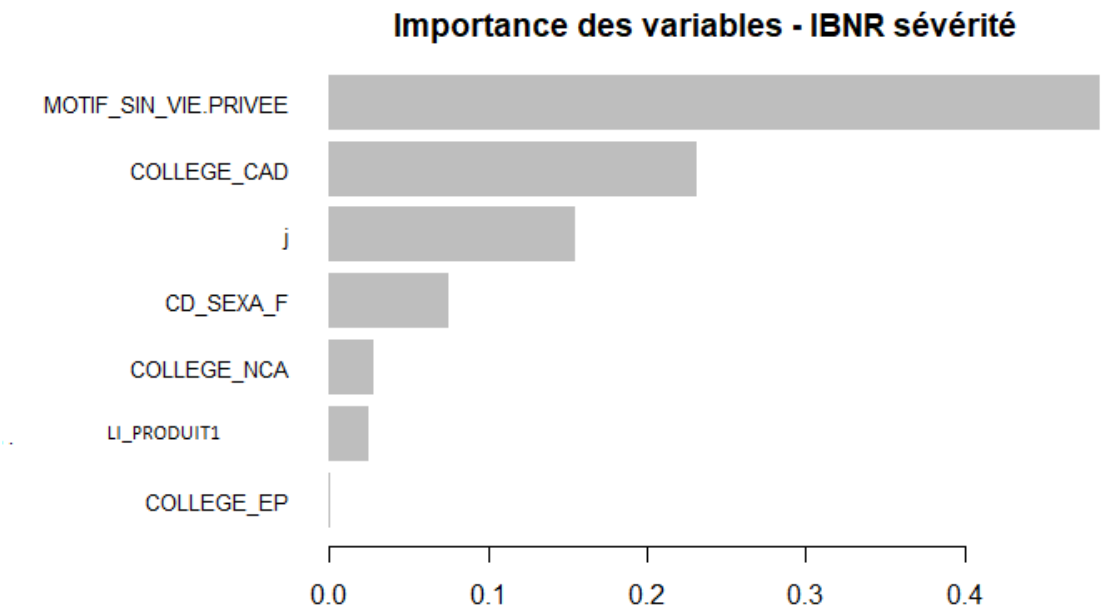


FIGURE 6.6 – Importances des variables dans l'algorithme

6.3 Comparaison des charges prédictes

Dans un premier temps, nous avons pu comparer visuellement l'ajustement réalisé par chaque modèle utilisé sur la même base de test. Dans cette partie, nous pourrons comparer quantitativement ces modèles. On peut synthétiser ces résultats avec le tableau suivant : Notre

Survenance	Provisions BE	CL	Schnieper	XGBOOST
2011	24 887	1 115	-206	0
2012	10 448	1 916	13	3437
2013	17 055	50 342	40 546	40 270
2014	947 005	1 228 861	1 002 577	617 121
2015	4 744 276	5 417 189	4 449 890	4 999 742
2016	13 138 527	14 894 023	12 624 545	9 674 298

TABLE 6.10 – Prédictions du coût moyen

modèle de machine learning sera situe à -19% en deçà des provisions best estimate. Cela s'explique par notre modèle approché pour les IBNR.
Les différentes méthodes n'ont pas cependant réussi à prédire nos sinistres très tardifs de 2011.

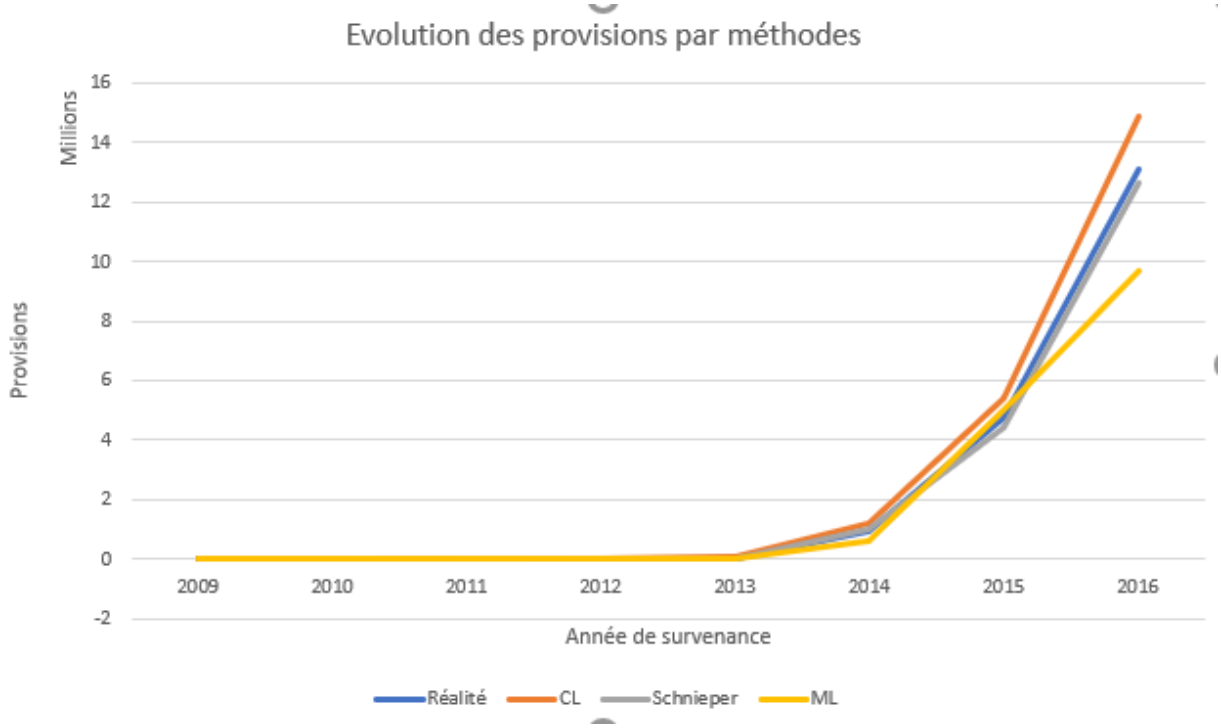


FIGURE 6.7 – Comparaison de l'estimation des provisions avec Chain-Ladder, Schnieper et XGBoost

Chapitre 7

Application sur le risque Invalidité

Contrairement à l'incapacité où nous avons une durée théorique maximale de 3 ans, sur le risque invalidité l'estimation de nos provisions de sinistres tardifs sera plus complexe.

Le principal problème dans cette partie était de distinguer les sinistres déjà dans notre portefeuille qui passent en invalidité et ceux qui récupérer à la suite de reprise de passifs lors d'une nouvelle recommandation.

Dans le cas d'une reprise de passif, nous nous sommes retrouvés avec des sinistres ayant une date du premier jour indemnisé ayant très lieu souvent bien après la date d'octroi de la rente qui correspond au moment où la rente invalidité a débuté.

Nous allons considérer ici la période avant vision qui correspond à la différence de la date du premier paiement et la date du premier jour indemnisé.

Nous utiliserons les méthodes définies dans les chapitres 2 et 3 pour déterminer le nombre de sinistres tardifs.

Pour ce risque long, nous avons décidé de considérer les sinistres survenus entre 01/01/2009 et le 31/12/2016. Nous nous intéresserons ici à l'ensemble des sinistres invalidité vu le faible de données.

Sur cette période, nous enregistrons 2 091 sinistres. Nous rappelons que notre objectif est de d'estimer les sinistres tardifs et le montant affecté à ceux-ci.

7.1 Résultats de nos méthodes classiques

7.1.1 Chain-Ladder

En appliquant la méthode de Chain Ladder, sur le triangle des sinistres nous obtenons :

Période de survenance	Nombre de sinistres	Observations des tardifs en vision 31/12/2021
2009	0	0
2010	0	0
2011	0	1
2012	1	1
2013	1	3
2014	6	12
2015	16	11
2016	262	374

TABLE 7.1 – Estimation du nombre de sinistres Chain-Ladder

Il nous apparaît que l'estimation des sinistres tardifs par la méthode de Chain-Ladder est très peu performante. Si cette méthode était retenue, on se retrouverait avec plus d'une centaine de sinistres tardifs qui n'auraient pas été estimé ce qui nous mettrait dans un cas de sous-provisionnement.

7.1.2 Modèle de Schnieper

Dans cette section, nous utiliserons la méthode de Schnieper appliquée pour déterminer le nombre de sinistres apparu toujours dans l'objectif de déterminer les tardifs.
Nous rappelons dans ce cas les notations :

- $C_{i,j}$ est le nombre cumulé de sinistres ;
- $N_{i,j}$ le nombre de nouveaux sinistres assimilés au IBNyR ;
- $D_{i,j}$ le nombre cumulé agrégé des variations des sinistres observées entres les années d'indemnisation $j-1$ et j pour chaque année de survenance i.

Nous interesserons ici à nos IBNyR qui sont représentés par la matrice N.

Le véritable problème dans cette modélisation a été de déterminer les estimations possibles des expositions sur l'ensemble d'étude. On a des écarts principalement sur la dernière année de survenance où nous avons une mauvaise prédiction. L'apport de l'exposition annuelle est véritablement un plus dans la précision de notre estimation contrairement à celle de Chain-Ladder. Egalement, on a une évolution de la population assurée en 2016 car c'est une année où nous avons plusieurs appels d'offres et nouvelles recommandations.

Survenance	$N_{i,1}$	$N_{i,2}$	$N_{i,3}$	$N_{i,4}$	$N_{i,5}$	$N_{i,6}$	$N_{i,7}$	$N_{i,8}$
2009	34	43	0	0	0	0	0	0
2010	102	45	1	1	0	0	0	
2011	115	69	13	1	0	2		
2012	131	94	10	3	1			
2013	158	98	11	2				
2014	199	95	6					
2015	231	110	6					
2016	275	6						

TABLE 7.2 – Triangles des sinistres nouveaux $N_{i,j}$

Période de survenance	Nombre de sinistres	Observations des tardifs en vision 31/12/2021
2009	0	0
2010	0	0
2011	0	1
2012	1	1
2013	1	3
2014	3	2
2015	12	11
2016	350	374

TABLE 7.3 – Estimation du nombre tardifs par la méthode de Schnieper

7.1.3 Méthode de provisionnement individuel

Nous nous intéressons à la fréquence de survenance des sinistres en invalidité.La fonction objective est un "count :poisson" comme dans le cas des fréquences IBNR.

Période de survenance	Nombre de sinistres	Observations des tardifs en vision 31/12/2021
2011	0	1
2012	0	1
2013	2	3
2014	3	2
2015	8	11
2016	313	374

TABLE 7.4 – Estimation du nombre tardifs par la méthode individuelle

L'erreur relative de prédictions est de -17% ce qui est en deçà de la réalité.

Conclusion générale

Dans ce mémoire, nous nous sommes intéressés au provisionnement de nos sinistres en particulier des sinistres tardifs des garanties Arrêt de Travail en utilisant des méthodes de provisionnement non-vie.
Les méthodes classiques d'évaluation de ces provisions, utilisent des triangles de liquidation par année de survenance et année de développement. Ces méthodes n'ont pas toujours une précision, les données étant agrégées, des informations non-prises en compte sur le sinistre et une absence de séparation des RBNS et des IBNR. Nous avons voulu explorer la piste d'une méthode de provisionnement individuel, en utilisant les travaux de Baudry & Robert [2] pour déterminerons provisions et cette séparation IBNR et RBNS.

Pour développer notre étude, nous sommes intéressés aux données à notre disposition. Nous avons écarté certaines années jugées atypiques mais aussi les sinistres atypiques par l'étude nos valeurs extrêmes. Nous avons veillé à posséder des variables fiables pour nos cas d'étude.

Pour le risque incapacité, nous avons estimé le montant des provisions que nous avons comparé aux provisions réelles. La méthode de Chain-Ladder ayant un écart de 14,4% avec nos provisions réelles en incapacité ; Cette méthode est inadaptée à ce genre de sinistre car la durée des sinistres ainsi que les montants réglés sont assez volatils selon le cas. Nous avons obtenu une incertitude à l'ultime du montant total des provisions de 23% pour le modèle de Mack.

Parmi ces méthodes agrégées, nous avons obtenus avec la méthode de Schnieper des résultats encourageants avec un écart de nos provisions de 5%. Avec le déploiement de la DSN cette méthode peut être très utile surtout qu'elle permet une séparation entre les provisions des sinistres connus et inconnus.

Notre modèle de provisionnement individuel sépare lui les RBNS et les IBNR, les IBNR déterminés par une méthode de fréquence/sévérité. Nous obtenons un écart de 11 % avec nos provisions réelles.

Pour le risque invalidité, nous avons voulu déterminer le nombre de sinistres tardifs. La méthode de Schnieper était la meilleure devant notre méthode de provisionnement individuel et la méthode de Chain-Ladder.

De notre étude, nous avons conclu que la méthode de Schnieper était la plus en adéquation avec la réalité. Elle profite de l'utilisation des données DSN pour être encore plus précises. Elle estla solution opérationnelle la plus fiable.

L'utilisation du modèle individuel peut être une alternative pour déterminer la fréquence ou le montant de nos sinistres inconnus. Cependant, en pratique l'utilisation de modèle individuel n'est pas chose aisée, à la fois chronophage et nécessitant des données déjà suffisamment traitées et la bonne compréhension des modèles.

Bibliographie

[1] R. Schnieper. Separating true ibnr and ibner claims. *ASTIN Bulletin : The Journal of the IAA*, page 111–127, 1991.

[2] Baudry and Robert. Non parametric individual claim reserving in insurance. 2017.

[3] Legifrance. http://www.legifrance.gouv.fr.

[4] Thomas Mack. Distribution-free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin : The Journal of the IAA*, pages 213–225, 1993.

[5] Astin working party (june 2016). http://www.actuaries.org/ASTIN/Documents/ASTIN_WP_NL_Reserving_Report1.0_2016-06-15.pdf.

[6] Elja Arjas. The claims reserving problem in non-life insurance : Some structural ideas. *ASTIN Bulletin*, pages 139–152, 1989.

[7] Ragnar Norberg. Prediction of outstanding liabilities in non-life insurance. *ASTIN Bulletin*, pages 95–115, 1993.

[8] Haastrup and Arjas. *Claims reserving in continuous time : a nonparametric bayesian approach.* 1996.

[9] Antonio and Plat. *Micro-level stochastic loss reserving for general insurance. Scan- dinavian Actuarial Journal.* 2014.

[10] Xian Zhou Xiao Bing Zhao and Jing Long Wang. Semiparametric model for prediction of individual claim loss reserving. *Insurance : Mathematics and Economics*, 2009.

[11] Xiao Bing Zhao and Xian Zhou. Applying copula models to individual claim loss reserving methods. *Insurance : Mathematics and Economics*, page 290–299, 2010.

[12] Wuthrich. Machine learning in individual claims reserving. 2016.

[13] Wuthrich. Neural networks applied to chain-ladder reserving. 2017.

[14] Pierre-E Thérond et al Olivier Lopez, Xavier Milhaud. Tree-based censored regression with applications in insurance. 2016.

[15] Chen and Guestrin. Xgboost : A scalable tree boosting system, eprint arxiv. 2016.

[16] Dares. hhttps://dares.travail-emploi.gouv.fr/.

[17] Marie KRATZ. Cours : Théorie des valeurs extremes. In *ISUP*, 2021.

Table des figures

1.1 Exemples de censures à droite et de troncature à gauche . . . 20
1.2 Schéma des étapes d'un sinistre . . . 21

2.1 Triangle de liquidation . . . 24

3.1 Vie sinistre . . . 35
3.2 Représentation graphiques des IBNR et RBNS . . . 36
3.3 Représentations graphiques des IBNR et RBNS . . . 38
3.4 Données de train X_{train} Y_{train} j=1 . . . 39
3.5 Données de test X_{test} et Y_{test} j=1 . . . 39

4.1 Méthodes ensemblistes . . . 42

5.1 Estimation des excès moyen - Incapacité . . . 55
5.2 Estimation des excès moyen - Invalidité . . . 56
5.3 Estimation de Hill - Incapacité . . . 57
5.4 Estimation de Hill - Invalidité . . . 57
5.5 Distribution de la sinistralité par année . . . 59
5.6 Durée moyenne de déclaration en jours par année de survenance . . . 60
5.7 Fréquence d'entrée en incapacité par âge . . . 61
5.8 Durée moyenne des sinistres par âge . . . 61
5.9 Nombre de passage en invalidité . . . 62
5.10 Répartition par âge en invalidité . . . 62
5.11 Durée en invalidité . . . 63
5.12 Matrice de V de Cramer . . . 64

6.1 Fréquence des coûts annuels RBNS . . . 70
6.2 Variables d'intérêt RBNS . . . 71
6.3 Fréquence des IBNR . . . 72
6.4 Fréquence des coûts annuels RBNS . . . 73
6.5 Fréquence des coûts IBNR . . . 74
6.6 Importances des variables dans l'algorithme . . . 74
6.7 Comparaison de l'estimation des provisions avec Chain-Ladder, Schnieper et XG-Boost . . . 75

Liste des tableaux

1 Ecart entre la provision réelle et l'estimation Chain-Ladder et Schnieper 5
2 Provisions individuelles 6
3 Estimation du nombre de sinisitres 7
4 Difference between the actual reserves and the Chain-Ladder and Schnieper estimates 9
5 Individual reserves 10
6 Estimation du nombre de sinisitres 10

1.1 Récap des indemnités journalières versées par la SS.- cas vie privée 17
1.2 Récap des indemnités journalières versées par la SS- cas vie pro. 17
1.3 Montant des pensions d'invalidité lié à la vie privée 19
1.4 Montant de l'indemnité en capital versée selon le taux d'incapacité permanente (depuis le 01/01/2022) 19

5.1 Récapitulatif des domaines d'attraction 54
5.2 Variations interquantiles- Incapacité 58
5.3 Variations interquantiles- Invalidité 58
5.4 Choix du seuil 58
5.5 Choix du seuil 59

6.1 Triangle des charges cumulés cas d'étude 67
6.2 Estimation des provisions Chain-Ladder 67
6.3 Triangles des sinistres nouveaux $N_{i,j}$ 68
6.4 Triangle des variations $D_{i,j}$ 68
6.5 Estimations des paramètres de Schnieper 68
6.6 Résultats provisions de Schnieper 69
6.7 Prédictions des RBNS 70
6.8 Prédictions de la fréquence 72
6.9 Prédictions de la sévérité 73
6.10 Prédictions du coût moyen 75

7.1 Estimation du nombre de sinistres Chain-Ladder 77
7.2 Triangles des sinistres nouveaux $N_{i,j}$ 78
7.3 Estimation du nombre tardifs par la méthode de Schnieper 78
7.4 Estimation du nombre tardifs par la méthode individuelle 78