Mémoire présenté devant l'ENSAE Paris
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires
le 15/03/2023

Par : **Tinhinane TALBI**

Titre : **Développement, extension et comparaison de modèles de provisionnement individuel : projections tenant compte des spécificités des sinistres graves de la branche RCC Automobile**

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

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

Membres présents du jury de la filière

Nom : Caroline HILLAIRET

Membres présents du jury de l'Institut des Actuaires

Entreprise : Milliman

Signature :

Directeur du mémoire en entreprise :

Eve Elisabeth TITON :
Signature :

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

Signature du responsable entreprise

Secrétariat :

Signature du candidat

Bibliothèque :

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

Résumé

L'assurance est une activité à cycle économique inversé : ainsi, les provisions pour sinistres sont d'une importance majeure pour l'assureur pour qu'il puisse honorer les engagements pris envers ses assurés. Afin d'évaluer les provisions techniques, les assureurs utilisent usuellement des méthodes de provisionnement dites agrégées, telles que Chain-Ladder. Ces méthodes, appréciées pour leur simplicité et la facilité de leur interprétation, présentent en revanche un certain nombre d'inconvénients, dont l'hypothèse d'indépendance forte entre les facteurs de développement et les années de survenance. Par ailleurs, ces méthodes sont relativement peu informatives sur les différents schémas de développement de sinistres.

Au cours des dernières décennies, un tournant s'est effectué dans la recherche actuarielle sur le sujet du provisionnement : différents travaux sur des méthodes individuelles ont été menés. Ces méthodes permettent d'explorer l'information relative à chaque sinistre individuellement ce qui garantit une meilleure connaissance des caractéristiques et des développements des paiements des sinistres. La contribution fondamentale des méthodes individuelles consiste en leur flexibilité et en la possibilité de séparer la réserve relative aux RBNS (Reported But Not Settled) de celle des IBNyR (Incurred But Not Yet Reported).

L'objectif de ce mémoire est de donner un aperçu global des méthodes de provisionnement ligne à ligne existantes à ce jour, de leurs avantages et inconvénients, et d'en comparer certaines, entre elles et avec les méthodes classiques Chain-Ladder et Mack, sur un portefeuille Responsabilité Civile Corporelle Automobile constitué de sinistres graves. Nous nous focalisons sur le provisionnement au titre des sinistres ouverts ayant déjà franchi un certain seuil de charge, mais non encore clôturés, notés RBNS.

Différents modèles individuels sont implémentés pour simuler la charge ultime au titre des RBNS : un modèle paramétrique (modèle stochastique à états) et deux modèles non-paramétriques (de type machine learning). Les modèles sont challengés et adaptés en fonction des spécificités du traitement des sinistres graves et des caractéristiques de la branche étudiée (recherche de variables prédictives de l'ultime). Une comparaison des réserves globales obtenues, y compris avec la méthode Chain-Ladder, ainsi qu'une étude des erreurs de chacun des modèles sont menées.

Mots clés : provisionnement individuel, RBNS, sinistres graves, méthodes stochastiques, machine learning, méthode agrégée, Chain-Ladder, RCC Auto, taux d'AIPP

Abstract

Insurance is a business with a reverse economic cycle : thus, claims reserves are of major importance for the insurer to be able to honour the commitments made to its policyholders. In order to evaluate the amount of reserves, insurers usually use so-called aggregate provisioning methods, such as Chain-Ladder. These methods are appreciated for their simplicity and ease of interpretation, but they have a number of drawbacks, including the assumption of independence between development factors and years of occurrence. In addition, these methods are relatively uninformative about the different patterns of claims development.

In recent decades, a turning point has been reached in actuarial research on the subject of reserving : various works on individual methods have been carried out. These methods make it possible to explore the information about each individual claim, which guarantees a better knowledge of the characteristics and developments of claim payments. The fundamental contribution of individual methods is their flexibility and the possibility to separate the reserve for RBNS (Reported But Not Settled) from IBNyR (Incurred But Not Yet Reported).

The objective of this paper is to provide an overview of the existing individual reserving methods, their advantages and disadvantages, and to compare some of them with each other and with the classical Chain-Ladder method on a Motor Bodily Injury Liability portfolio consisting of severe claims. We focus on reserving for open claims that have already crossed a certain threshold of charge but are not yet closed, called RBNS.

Different individual models are implemented to simulate the ultimate charge for RBNS : one parametric model (stochastic state model) and two non-parametric models (machine learning). The models are challenged and adapted according to the specificities of severe claims handling and the characteristics of the branch studied (search for predictors of the ultimate). A comparison of the overall reserves obtained, including with the Chain-Ladder method, as well as a study of the errors of each of the models, are carried out.

Keywords : individual reserving, RBNS, severe claims, stochastic methods, machine learning, aggregate methods, Chain-Ladder, Bodily Injury Liability, AIPP rate

---

Remerciements

Je tiens à saisir cette occasion afin d'adresser mes sincères remerciements à Milliman et l'ensemble de l'équipe R&D, plus particulièrement **Alexandre Boumezoued** directeur de l'équipe, pour m'avoir offert l'opportunité d'effectuer mon stage de fin d'études au sein de son équipe et pour la confiance accordée. Cette expérience était enrichissante tant au niveau professionnel que personnel et j'en suis très reconnaissante.

J'aimerais plus particulièrement remercier ma tutrice de stage **Eve Elisabeth Titon** de m'avoir fait confiance suite à notre entretien et pour ses nombreuses qualités professionnelles et humaines, pour son encadrement, sa disponibilité et ses précieux conseils. Elle m'a en effet transmis sa connaissance métier et son analyse perspicace.

Je remercie également **Kevin Lecomte** pour ses précieux conseils, sa bonne humeur et d'avoir pris le temps de relire mon mémoire d'actuariat.

J'adresse mes remerciements également à l'ENSAE pour la qualité de son enseignement, et à ma tutrice pédagogique **Caroline HILLAIRET** pour sa relecture.

Enfin, je souhaite remercier ma famille pour son soutien et ses encouragements durant tout mon parcours académique.

Note de Synthèse

Contexte et objectifs

L'activité assurantielle se caractérise par un cycle économique inversé, selon lequel l'assureur détermine le prix de vente du contrat (prime) avant d'en connaître le coût de production. Afin d'être en mesure d'honorer ses engagements futurs, il convient alors pour l'assureur d'estimer ses charges futures et de les provisionner : il s'agit des provisions techniques. La provision estimée doit être suffisante pour garantir une liquidité permettant de faire face aux engagements pris. En revanche, le montant de la provision ne doit pas être trop élevé, car une provision excessive implique une perte sur les gains qui auraient pu être générés par le placement de la somme excédentaire dans des actifs financiers.

En assurance non-vie, il existe différents types de provisions techniques. Dans ce mémoire, nous nous intéressons aux provisions pour sinistre à payer (PSAP) : dans ce mémoire, nous étudierons les provisions pour sinistres survenus, déclarés mais non encore complètement payés (RBNS) ; les provisions pour sinistres non encore déclarés (IBNyR) ne sont pas étudiées.

Dans le but d'évaluer les PSAP, les assureurs utilisent des méthodes de provisionnement standard, dites "agrégées" : les règlements sont agrégés par année de survenance du sinistre et par année de développement. Ces informations sont enregistrées dans un triangle de liquidation où seulement la partie supérieure du triangle est renseignée. L'objectif de ces méthodes est donc l'estimation de la partie inférieure du triangle de liquidation afin de constituer la charge ultime. Les PSAP sont ensuite calculées comme la différence entre la charge ultime et les paiements déjà effectués. Les différentes méthodes de provisionnement standard peuvent être regroupées au sein de deux catégories :

- Méthodes déterministes : connues pour leur simplicité et robustesse. Parmi ces méthodes, Chain-Ladder est la plus populaire car facile à comprendre, à interpréter et à mettre en oeuvre.
- Méthodes stochastiques : plus complexes, ces méthodes permettent de mesurer la variabilité des estimations. Parmi ces méthodes nous citons celle de Mack qui permet d'estimer les erreurs d'estimation.

L'avantage principal de ces méthodes réside dans leur simplicité d'implémentation, d'utilisation et d'interprétation des résultats. Les méthodes standard présentent toutefois des inconvénients, dont celui commun à la méthode de Chain-Ladder et celle de Mack qui est le fait que ces dernières reposent sur une hypothèse d'indépendance, forte, entre les facteurs de développement et les années de survenance. Les outils informatiques ayant aujourd'hui des capacités calculatoires et de traitement de bases de données volumineuses

importantes, de nouvelles méthodes de provisionnement ont pu voir le jour afin d'exploiter les informations à disposition des assureurs sur les sinistres et les assurés à la maille individuelle. Ceci, afin de constituer des provisions plus justes, à une maille plus fine. Autrement dit, plutôt que de se baser sur les données agrégées pour estimer la provision, utiliser les données individuelles.

L'objectif de ce mémoire est de donner un aperçu global des méthodes de provisionnement individuel existantes à ce jour : nous en appliquerons certaines, judicieusement choisies, à un portefeuille de sinistres graves de la branche RCC Automobile d'un grand assureur français. Nous comparerons ces méthodes entre elles d'une part, et avec les méthodes classiques Chain-Ladder et Mack d'autre part. Nous nous focalisons sur le provisionnement au titre des sinistres ouverts ayant déjà franchi un seuil de 500k € de charge mais non encore clôturés, notés RBNS. Enfin, une extension des modèles implémentés est effectuée afin d'étudier l'impact des caractéristiques de la branche étudiée sur l'estimation de la provision.

Modélisation

La visée de ce mémoire est d'estimer la charge ultime liée aux sinistres graves au titre des RBNS via différents modèles ligne à ligne afin de pouvoir comparer les résultats des modèles avec les résultats de la méthode standard Chain-Ladder et de confronter les avantages et les inconvénients de chacun des modèles.

Différents modèles ont été implémentés :

- Un modèle paramétrique (modèle à états) : le cadre d'étude de ce modèle est défini par Boumezoued et Devineau (2017). L'objectif est de modéliser le processus de développement des règlements et de clôture des sinistres, selon trois types d'évènements :
  1. Clôture sans paiement.
  2. Paiement sans clôture.
  3. Paiement avec clôture.

  Chaque type d'évènement (1, 2, ou 3) apparaît avec une intensité h_1, h_2 et h_3 respectivement.

  La modélisation du processus de règlements et de clôture est faite via un processus de Markov.

  Une fois cette modélisation faite, les intensités de transition sont estimées à l'aide de la méthode de maximum de vraisemblance. Sous l'hypothèse de constance (ou constance par morceaux) des fréquences associées à chacun des évènements, la maximisation de la fonction de vraisemblance conduit à des estimateurs classiques du type :

$$\hat{h_i} = \frac{\text{Nombre total d'évènements de type i}}{\text{Exposition au risque}}.$$

Ensuite, la distribution des paiements est à son tour calibrée via une loi Log-Normale qui caractérise la distribution des paiements observés.

$$P(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right) \tag{1}$$

Enfin, grâce aux intensités et aux paramètres de la loi Log-Normale calibrés, le paiement futur attendu peut être calculé à l'aide de la formule fermée spécifiée dans l'article comme suit :

$$\mu(s) = \mathbb{E}[X(u,\infty)|S(u) = 3] = \int_u^\infty \left\{ (h_2(v)y_2(v) + h_3(v)y_3(v)) \exp(-\int_s^u (h_1(v) + h_2(v))\mathrm{d}v) \right\} \tag{2}$$

Avec :

- $X(u, \infty)$: représente le paiement total dans $[u, \infty]$
- $S(u)$: processus décrivant la trajectoire des états des sinistres (après déclaration) et prenant ses valeurs dans $\mathbb{N}$
- y(ν) : moyenne des paiements réalisés

Une distribution des paiements futurs peut également être déterminée via un algorithme de simulation.

- Modèles non-paramétriques : les algorithmes XGBoost et Random Forest ont été implémentés en considérant deux types de données pour l'entraînement des modèles.

  1. Entraînement du modèle sur les sinistres clôturés seulement, ie : nous ne considérons pas les données censurées. La variable réponse à apprendre, l'ultime, n'est disponible que pour les sinistres clos.
  2. La deuxième approche consiste à développer les sinistres non clôturés à la date d'évaluation en utilisant les paramètres de l'approche classique de Mack Chain-Ladder et utiliser l'ensemble des observations disponibles : clos et RBNS complétés par des « pseudo-ultimes » pour l'entraînement des modèles.

     Les facteurs de développement sont calculés en considérant le triangle «Année de survenance de sinistre» x «Année de dépassement du seuil de 500k €».

Implémentation des modèles et résultats

Les modèles ainsi définis sont appliqués sur une base de sinistres grave en RCC Automobile. À la suite des traitements effectués sur la base, nous obtenons une base de données composée de 2460 dossiers victimes graves, dont 1365 sont clos. Afin de faciliter l'implémentation des modèles envisagés, la base de données est transformée en vision annuelle.

Calibrage du modèle à états

Le calibrage du modèle à états s'effectue en trois étapes : calibrage des fréquences des évènements de paiements et de clôture, calibrage des paiements et calcul de la provision avec la formule fermée 3.3 et par simulation.

Le montant de la provision calculée par formule fermée s'élève à 1.67 Mds €. Le montant des paiements à date est ajouté à cette provision afin de constituer la charge ultime au titre des RBNS vus au 31/12/2019 : celle-ci s'élève à 2.3 Mds €.

Le même résultat est obtenu avec l'algorithme de simulation, qui nous permet de représenter une distribution des paiements futurs :

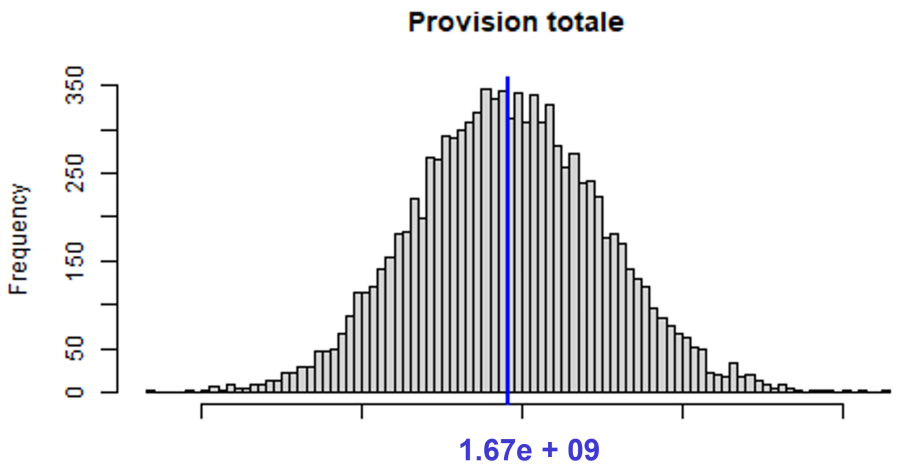


FIGURE 1 – Distribution de la provision totale

Le modèle à états tel qu'il a été défini nous permet de calculer une erreur quadratique moyenne des prédictions (MSEP) qui peut être scindée en deux : une erreur de process relative au caractère stochastique des trajectoires futures, plus une erreur d'estimation liée à l'incertitude sur la valeur des paramètres estimés.

Les résultats obtenus sont les suivants :

RMSE	Erreur de process	Erreur d'estimation
371 182 900	70 496 366	300 686 534

TABLE 1 – Erreurs de prédiction du modèle à états

Calibrage des modèles non paramétriques

La base de données initiale a été scindée aléatoirement en deux sous-bases indépendantes, une base d'apprentissage et une base de test. Selon l'approche considérée, les algorithmes Random Forest et XGBoost sont entraînés sur 80% des observations puis testés sur les visions des 20% restants.

Un exercice de backtesting est effectué afin de vérifier la robustesse des modèles. Autrement dit, les algorithmes sont appliqués sur la base de données dans des années N antérieurs à 2019, avec l'information sur les RBNS vus à fin N et clos entre l'année N et 2019.

Nous considérons les notations suivantes :

- Modèle A - global : modèle global entrainé sur les sinistres clos uniquement.
- Modèle A - différencié : modèle en deux parties entrainé sur les sinistres clos uniquement.
- Modèle B - global : modèle global entrainé sur les sinistres clos et RBNS.
- Modèle B - différencié : modèle en deux parties entrainé sur les sinistres clos et RBNS.

Avec :

— Modèle différencié : modèle en deux partie avec un modèle apprenant sur toutes les visions de sinistres qui auront un ultime inférieur à 500k €, et un autre apprenant toutes les visions de sinistres qui auront un ultime supérieur à ce seuil.

— Modèle global : un seul modèle qui apprend toutes les visions de sinistres sans distinction sur l'ultime.

Nous obtenons les résultats ci-dessous :

Gestionnaire sinistre

La provision globale estimée par le gestionnaire surestime l'ultime total payé d'au moins 20%, sur toutes les années de développement des sinistres.

Résultats pour les modèles entraînés sur les sinistres clos

Le modèle A-global-Random Forest sous-estime l'ultime sur certaines années de développement et surestime sur d'autres ; le modèle A-global-XGBoost sous-estime l'ultime quasiment sur toutes les années. Les modèles A-différencié-Random Forest et A-différencié-XGBoost surestiment l'ultime : ceci peut être dû au fait qu'un grand nombre de visions de sinistres qui auront un ultime inférieur à 500k € passent dans le modèle entraîné sur les sinistres ayant un ultime supérieur ou égal à 500k €, ce qui conduit à une surestimation de l'ultime pour ces sinistres.

Résultats pour les modèles entraînés sur les sinistres clos et RBNS

Le modèle B-différencié-XGBoost sous-estime l'ultime, tandis que le global le sous-estime sur l'année 2010, puis le surestime sur le reste des années de backtesting. Le modèle B-différencié-Random Forest surestime l'ultime plus que le modèle global.

D'après ces résultats et les prédictions des modèles au titre des RBNS vus au 31/12/2019 nous concluons que :

- Les modèles globaux sont privilégiés et un ultime « juste » serait plutôt égal à $2.0 - 2.1$ Mds €, en restant prudents.
- Les prédictions du modèle A-global-Random Forest, A-global-XGBoost et B-global-XGBoost donnent des estimations trop faibles. En revanche, le modèle B-global Random Forest fournit des estimations plus cohérentes avec l'ultime « juste » avec un ultime qui s'élève à 2.0 Mds €.

Une étude avec les méthodes classiques Chain-Ladder et Mack est également réalisée, les résultats obtenus sont présentés dans le tableau qui suit :

Modèle	Ultime	RMSE	Erreur de process	Erreur d'estimation	Borne inf	Borne sup
Modèle à états	2.3 Mds	371 182 900	70 496 366	300 686 534	2.2 Mds	2.4 Mds
B global RF	2.0 Mds			127 047 408	1.9 Mds	2.1 Mds
Chain-Ladder	2.1 Mds					
Mack	2.1 Mds					

TABLE 2 – Comparaison des résultats

Remarque : les intervalles de confiances sont à 90%.

L'ultime évalué avec les deux méthodes classiques s'élève à $2,1$ Mds €, cette estimation est cohérente avec l'estimation des deux modèles individuels et l'ultime « juste ».

Extension des modèles de provisionnement ligne à ligne implémentés

Pour prolonger cette étude, une analyse supplémentaire est effectuée sur le modèle à états et le modèle B - global Random Forest afin de mesurer l'impact des spécificités de la branche RCC Automobile étudiée sur les provisions estimées par ces modèles.

Une des spécificités des sinistres de notre base est l'évènement de dépassement de seuil de 500k€ de charge, qui définit le caractère "grave" des sinistres. Le modèle à états est calibré sur le portefeuille de données sans tenir compte de cet évènement, ainsi nous avons décidé d'adapter le modèle de base afin de l'intégrer. Deux nouveaux modèles sont implémentés : un modèle avec dépassement de seuil calibré selon la position du dernier paiement cumulé par rapport à ce seuil et un modèle avec dépassement de seuil calibré selon la position de la charge par rapport à ce seuil.

Une autre particularité que nous étudions est l'incidence de la présence d'une variable ayant un lien direct avec la sévérité des sinistres de cette branche. Pour ce faire, nous appliquons le modèle à états ainsi que le modèle B - global Random Forest aux dossiers pour lesquels nous disposons de l'information concernant le taux d'AIPP. Dans un premier temps, nous appliquons les modèles sur la base restreinte à ces dossiers, en présence de la variable taux d'AIPP. Dans un second temps, nous excluons cette variable de la base restreinte et nous appliquons les modèles à nouveau. Enfin, nous comparons les résultats afin de mesurer l'effet de cette variable sur les prédictions des modèles.

Résultats

Implémentation du modèle à états avec dépassement de seuil

Selon l'approche considérée nous obtenons des résultats différents :

- L'estimation du modèle à états avec dépassement de seuil calibré sur le dernier paiement cumulé, évaluée par formule fermée, s'élève à 1.00 Mds €.

- L'estimation du modèle à états avec dépassement de seuil calibré sur la charge, évaluée par formule fermée, s'élève à 1.60 Mds €, légèrement inférieur à la provision estimée par le modèle de base qui est évaluée à 1.67 Mds €.

Ces résultats ne donnent qu'une idée des tendances obtenues par les modèles. Pour avoir des valeurs plus fiables et pouvoir sélectionner le meilleur modèle, il faudrait comparer leurs variabilités et leurs erreurs de process et d'estimation.

Implémentation de modèles ligne à ligne sur la base restreinte

Modèle à états

Base	Provision	RMSE	Erreur de process	Erreur d'estimation
Y compris AIPP	2.21 Mds €	494 206 144	85 371 712	408 834 433
Sans AIPP	2.28 Mds €	633 113 859	88 954 958	544 158 902

TABLE 3 – Résultats obtenus sur les deux bases

Le montant des paiements à date est ajouté aux provisions calculées par formule fermée afin de constituer la charge ultime au titre des RBNS vus au 31/12/2019 : celle-ci s'élève à 2.77 Mds € sur la base AIPP et 2.84 € sur la base restreinte sans la variable AIPP.

À partir de ces résultats nous pouvons voir qu'en termes d'erreur le modèle calibré sur la base restreinte en tenant compte de la variable taux d'AIPP est légèrement meilleur. Ce résultat confirme notre hypothèse : la prise en compte de variables ayant un lien direct avec la sévérité à un impact positif sur la performance du modèle à états.

Modèle B - global Random Forest

Les prédictions du modèle en présence de la variable taux d'AIPP et en son absence sont globalement identiques.

	Moyenne	Coefficient de variation
Base restreinte y compris taux d'AIPP	2.26 Mds €	5.2%
Base restreinte sans taux d'AIPP	2.39 Mds €	5.0%

TABLE 4 – Performances des modèles

Ces résultats donnent une idée du comportement du modèle non paramétrique en présence de la variable taux d'AIPP et en son absence. Pour avoir des valeurs plus fiables en termes de robustesse de modèle et de variabilité, il nous faudrait une base de données plus grande.

À ce stade nous pouvons constater que la présence de la variable taux d'AIPP augmente la performance du modèle. Ainsi, nous espérons obtenir des résultats plus pertinents sur une base de données volumineuse.

Conclusion

Ce mémoire a donc permis de donner un aperçu global des modèles de provisionnement ligne à ligne existants, et d'en implémenter certains dans le cadre de l'estimation de la provision au titre des RBNS d'un portefeuille de RCC automobile et de prolonger l'étude en adaptant des modèles en fonction des spécificités de la branche étudiée. L'application des différents modèles nous donne des valeurs de provision assez diverses. Cependant des travaux restent à effectuer sur ces modèles de provisionnement. Parmi ceux-ci, une étude supplémentaire sur les IBNyR afin de pouvoir réaliser une comparaison complète entre les résultats des modèles ligne à ligne et ceux de la méthode classique Mack. Enfin, l'application des modèles sur une base de données plus conséquente incluant des variables en lien direct avec la sévérité des sinistres.

Mots clés : provisionnement individuel, RBNS, sinistres graves, méthodes stochastiques, machine learning, méthode agrégée, Chain-Ladder, RCC Auto, taux d'AIPP

Executive summary

Context and objectives

The insurance business is characterised by an inverted economic cycle, whereby the insurer determines the selling price of the contract (premium) before knowing the production cost. In order to be able to meet its future commitments, the insurer must then estimate its future expenses and make provisions for them : these are the technical provisions. The estimated provision must be sufficient to guarantee liquidity to meet the commitments made. On the other hand, the amount of the provision must not be too high, because an excessive provision implies a loss of the gains that could have been generated by investing the excess amount in financial assets.

In non-life insurance, there are different types of technical provisions. In this memory, we are interested in the reserves claims : in this survey, we will study the reserves for claims incurred, reported but not yet fully paid (RBNS) ; the reserves for claims not yet reported (IBNyR) are not studied.

For the purpose of assessing reserve claims, insurers use standard, so-called "aggregated" reserving methods : settlements are aggregated by year of occurrence and year of development. This information is recorded in a run-off triangle where only the upper part of the triangle is filled in. The objective of these methods is therefore the estimation of the lower part of the liquidation triangle in order to constitute the ultimate charge. The reserve claims are then calculated as the difference between the ultimate charge and the payments already made. The different standard provisioning methods can be grouped into two categories :

- Deterministic methods : known for their simplicity and robustness. Among these methods, Chain-Ladder is the most popular because it is easy to understand, interpret and implement.
- Stochastic methods : more complex, these methods allow the variability of estimates to be measured. Among these methods, we cite Mack's method, which allows to estimate the estimation errors.

The main advantage of these methods is their simplicity of implementation, use and interpretation of results. However, the standard methods have some disadvantages, including the one common to the Chain-Ladder and Mack methods, which is the fact that they are based on a strong assumption of independence between the development factors and the years of occurrence. As computer tools now have the capacity to calculate and process large databases, new reserving methods have been developed to make use of the information available to insurers on claims and policyholders at the individual level. This is done in order to establish more accurate reserves at a finer scale. In other words, rather than relying on aggregate data to estimate the reserve, use individual data.

The objective of this memory is to give a global overview of the individual reserving methods that exist to date : we will apply some of them, judiciously chosen, to a portfolio of serious claims from a great French insurer's Motor Bodily Injury Liability branch. We will compare these methods with each other on the one hand, and with the classic Chain-Ladder and Mack methods on the other. We focus on the provisioning of open claims that have already exceeded a threshold of 500k € in expenses but have not yet been closed, called RBNS. Finally, an extension of the implemented models is carried out in order to study the impact of the characteristics of the branch studied on the estimation of the reserve.

Modelling

The aim of this memory is to estimate the ultimate charge of severe RBNS claims using different individual models in order to compare the results of the models with the results of the standard Chain-Ladder method and to compare the advantages and disadvantages of each model.

Different models have been implemented :

- A parametric model (state model) : the framework of study of this model is defined by Boumezoued et Devineau (2017). The objective is to model the claims development and settlement process according to three types of events :
  1. Settlement without payment.
  2. Payment without settlement.
  3. Payment with settlement.

  Each type of event (1, 2, or 3) appears with an intensity h_1, h_2 and h_3 respectively.

  The modelling of the payment and settlement process is done via a Markov process.

  Once this modelling is done, the transition intensities are estimated using the maximum likelihood method. Under the assumption of constancy (or piecewise constancy) of the frequencies associated with each of the events, maximisation of the likelihood function leads to classical estimators of the following type type :

$$\hat{h_i} = \frac{\text{Total number of events of type i}}{\text{Risque exposure}}.$$

  Then, the distribution of payments is calibrated using a Log-Normal distribution that characterises the distribution of observed payments.

$$P(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)$$

  Finally, with the calibrated intensities and parameters of the Log-Normale law, the expected future payment can be calculated using the closed formula specified in the article as follows :

$$\mu(s) = \mathbb{E}[X(u,\infty)|S(u)=3] = \int_u^{\infty} \left\{ (h_2(v)y_2(v) + h_3(v)y_3(v)) \exp(-\int_s^u (h_1(v)+h_2(v))\mathrm{d}v) \right\} \tag{3}$$

With :

- X (u, ∞) : represent the total payment in [u, ∞]
- S(u) : process describing the trajectory of claims states (after reporting) and taking its values in $\mathbb{N}$
- y(ν) : average payments made

A distribution of future payments can also be determined by the mean of a simulation algorithm.

- Non-parametric models : the XGBoost and Random Forest algorithms have been implemented considering two types of data for training the models.
  1. Training the model on closed claims only, meaning that we will not consider censored data. The response variable to be learned, the ultimate, is only available for closed claims.
  2. The second approach consists of developing claims not closed at the valuation date using the parameters of the classical Mack Chain-Ladder approach and using all available observations : closed and RBNS supplemented by "pseudo-ultimates" to train the models.

     The development factors are calculated by considering the triangle "Year of the accident" x "Year of exceeding the threshold of 500k €".

Model implementation and results

The models so defined are applied to a database of serious motor third-party liability claims. After processing the database, we obtain a database of 2460 serious casualty files, of which 1365 are closed. In order to facilitate the implementation of the envisaged models, the database is transformed into an annual view.

State model calibration

The calibration of the state model is carried out in three steps : calibration of the frequencies of payment and closure events, calibration of payments and calculation of the provision with the closed formula 3.3 and by simulation.

The amount of the provision calculated by closed formula is e 1.67 billion €. The amount of payments to date is added to this provision in order to constitute the ultimate charge for the RBNS seen on 31/12/2019 : this amounts to 2.3 billion €.

The same result is obtained with the simulation algorithm, which allows us to represent a distribution of future payments :

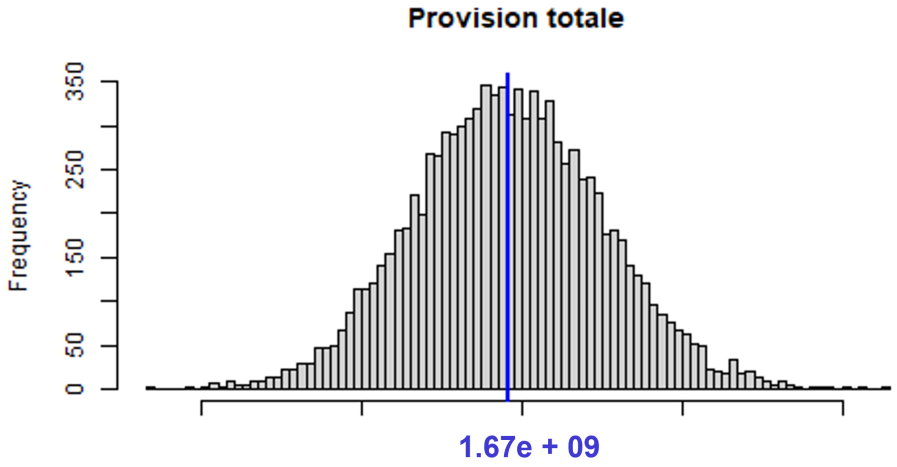


FIGURE 2 – Ultimate reserve distribution

The state model as defined allows us to calculate a mean square error of predictions (MSEP) which can be split into two : a process error related to the stochastic character of future trajectories, plus an estimation error related to the uncertainty on the value of the estimated parameters.

The following results were obtained :

RMSE	Process error	Estimation error
371 182 900	70 496 366	300 686 534

TABLE 5 – Prediction errors of the state model

Calibration of the non-parametric models

The initial database was randomly split into two independent sub-databases, a learning database and a test database. Depending on the approach considered, the Random Forest and XGBoost algorithms are trained on 80% of the observations and then tested on the visions of the remaining 20%. tested on the visions of the remaining 20%.

A backtesting exercise is performed to check the robustness of the models. In other words, the algorithms are applied on the basis of data in years N prior to 2019, with information on RBNS seen at the end of N and closed between year N and 2019. The following notations are considered :

- Model A - global : global model trained on closed claims only.
- Model A - differentiated : two-part model trained on closed claims only.
- Model B - global : global model trained on closed claims and RBNS.
- Model B - differentiated : two-part model trained on closed claims and RBNS.

With :

- Differentiated model : two-part model with one model learning on all visions of claims that will have an ultimate below 500k €, and another learning on all visions of claims that will have an ultimate above this threshold.
- Global model : a single model that learns all loss views without distinction distinction on the ultimate.

We get the following results : Claims manager
The overall provision estimated by the manager overestimates the total ultimate paid by at least 20%, over all years of claims development.

Results for trained models on closed claims
The A-global-Random Forest model underestimates the ultimate in some years of development and overestimates in others ; the A-global-XGBoost model underestimates the ultimate in almost all years. The A-differentiated-Random Forest and A-differentiated-XGBoost models overestimate the ultimate by a significant amount : this may be due to the fact that a large number of views of claims that will have an ultimate of less than 500k € are passed through the model trained on claims with an ultimate greater than or equal to 500k €, leading to an overestimate of the ultimate for these claims.

Results for trained models on closed claims and RBNS
The B-differentiated-XGBoost model underestimates the ultimate, while the global model underestimates it for the year 2010 and then overestimates it for the rest of the backtesting years. The B-differentiated-Random Forest model overestimates the ultimate more than the global model.

Based on these results and the predictions of the RBNS models as of 31/12/2019 we conclude that :

- Global models are preferred and a "fair" ultimate would be more like 2.0-2.1 billion €, with caution.
- The predictions of the A-global-Random Forest, A-global-XGBoost and B-global XGBoost models give too low estimates. In contrast, the B-global Random Forest model provides estimates that are more consistent with the "fair" ultimate with an ultimate that amounts to 2.0 billion €.

A study with the classical Chain-Ladder and Mack methods is also performed, The results obtained are presented in the following table :

Models	Ultimate	RMSE	Process error	Estimation error	lower bound	upper bound
State model	2.3 Mds	371 182 900	70 496 366	300 686 534	2.2 Mds	2.4 Mds
B global RF	2.0 Mds			127 047 408	1.9 Mds	2.1 Mds
Chain-Ladder	2.1 Mds					
Mack	2.1 Mds					

TABLE 6 – comparing results

Note : Confidence intervals are at 90%. The ultimate valued with the two classical methods amounts to e 2.1 billion €, which is consistent with the estimate of the two individual models and the "fair" ultimate.

Extension of the implemented individual provisioning models

To extend this study, an additional analysis is carried out on the state model and the B - global Random Forest model in order to measure the impact of the specificities of the RCC Automobile branch studied on the provisions estimated by these models.

One of the specificities of the claims in our database is the event of exceeding the 500k € load threshold, which defines the "serious" character of the claims. The state model is calibrated on the portfolio of data without taking into account this event, so we decided to adapt the basic model to integrate it. Two new models are implemented : a model with threshold overshoot calibrated according to the position of the last cumulative payment with respect to this threshold and a model with threshold overshoot calibrated according to the position of the load with respect to the threshold. according to the position of the load in relation to this threshold.

Another feature that we investigate is the impact of the presence of a variable that is directly related to the severity of claims in this branch. To do so, we apply the state model and the B - global Random Forest model to the files for which we have information on the AIPP rate. In a first step, we apply the models on the restricted basis to these files, in the presence of the AIPP rate variable. In a second step, we exclude this variable from the restricted base and apply the models again. Finally, we compare the results to measure the effect of this variable on the predictions of the models.

Results

Implementation of the state model with threshold exceedance

Depending on the approach considered, we obtain different results :

- The estimate of the state model with threshold overshoot calibrated on the last cumulative payment, evaluated by closed formula, amounts to e 1.00 billion €.
- The estimate of the state model with overshoot calibrated on the charge, evaluated by closed formula, amounts to e 1.60 billion €, slightly lower than the provision estimated by the basic model, which is valued at e 1.67 billion €.

These results only give an idea of the trends obtained by the models. In order to have more reliable values and to be able to select the best model, their variabilities and their process and estimation errors should be compared.

Implementation of individual models on the restricted database

State model

Database	Reserve	RMSE	Process error	Estimation error
With AIPP	2.21 Mds €	494 206 144	85 371 712	408 834 433
Without AIPP	2.28 Mds €	633 113 859	88 954 958	544 158 902

TABLE 7 – Results obtained on both databases

The amount of payments to date is added to the provisions calculated by closed formula in order to constitute the ultimate charge for the RBNS seen at 31/12/2019 : this amounts to 2.77 billion €with AIPP and 2.84 billion €without AIPP.

From these results we can see that in terms of error the model calibrated on the restricted basis taking into account the AIPP rate variable is slightly better. This result confirms our hypothesis : taking into account variables having a direct link with the severity has a positive impact on the performance of the state model.

Model B - Global Random Forest

The predictions of the model in the presence of the AIPP rate variable and in its absence are identical.

	Average	Coefficient of variation
Restricted datbase including AIPP rate	2.26 Mds €	5.2%
Restricted datbase without AIPP rate	2.39 Mds €	5.0%

TABLE 8 – Model performance

These results give an idea of the behaviour of the non-parametric model in the presence of the AIPP rate variable and in its absence. To have more reliable values in terms of model robustness and variability, we would need a larger database.

At this stage we can see that the presence of the AIPP rate variable increases the performance of the model. Thus, we hope to obtain more relevant results on a large database.

Conclusion

This paper has therefore provided an overview of existing individual reserving models and implemented some of them in the context of estimating the RBNS provision for a Motor Bodily Injury Liability portfolio and extended the study by adapting the models to the specificities of the branch studied. The application of the different models gives us quite different provision values. However, work remains to be done on these provisioning models. Amongst these, an additional study on IBNyR in order to be able to make a complete comparison between the results of the individual models and those of the classic Mack method. Finally, the application of the models on a larger database including variables directly related to the severity of claims.

Keywords : individual reserving, RBNS, severe claims, stochastic methods, machine learning, aggregate methods, Chain-Ladder, Bodily Injury Liability, AIPP rate

Table des matières

Introduction

Le travail de l'assureur se caractérise par la collecte des primes versées par les assurés afin de procéder au règlement des sinistres. En revanche, le montant des sinistres est inconnu lors de la détermination de la prime, ainsi pour être en mesure de faire face à ses engagements futurs, l'assureur estime le montant de ses charges futures et les provisionne.

L'estimation du montant des provisions est donc un enjeu important pour l'assureur. Les assureurs utilisent usuellement des méthodes de provisionnement classiques dites "agrégées". Ces dernières agrègent les charges liées à chaque sinistre par année de survenance du sinistre et par année de développement. Ces méthodes sont connues par la simplicité de leur mise en œuvre et la facilité de l'interprétation de leurs résultats. Cependant, elles présentent quelques inconvénients, notamment en termes de robustesse et de validité des résultats produits.

Suite à la révolution numérique que connait le secteur de l'assurance et l'importance accrue du Big Data, les modèles de provisionnement ligne à ligne attirent l'attention des assureurs. L'objectif principal de ces modèles est l'exploitation de toutes les informations liées à chaque sinistre individuellement, c'est à dire ne plus agréger, en vue de constituer une réserve adaptée à chaque sinistre.

L'objectif de ce mémoire est de présenter différents modèles de provisionnement ligne à ligne existants à ce jour, leurs avantages et inconvénients, et d'en comparer certains, entre eux et avec les méthodes classiques Chain-Ladder et Mack, dans le cadre d'une application sur un portefeuille de RCC Automobile constitué de sinistres graves. Enfin, nous testons la flexibilité des modèles individuels en adaptant certains des modèles ligne à ligne implémentés en fonction des spécificités de la branche étudiée.

Dans ce mémoire, nous nous intéressons à la partie RBNS. Pour ce faire, chaque dossier est considéré individuellement, il se caractérise donc par une date de survenance, une date de déclaration, un processus de règlement, un processus d'état à chaque date (clôturé ou en cours) et un évènement de dépassement de seuil (au-dessus ou en dessous du seuil).

Nous entamerons ce mémoire par un rappel des éléments clés du provisionnement en assurance non-vie. Ensuite, une présentation de l'état de l'art sur les méthodes de provisionnement ligne à ligne déjà existantes est effectuée. Puis, une introduction des modèles utilisés pour l'évaluation des provisions, aussi bien des modèles paramétriques que non-paramétriques est réalisée. En outre, une présentation du portefeuille RCC Automobile ainsi que les résultats issus de l'application des modèles ligne à ligne et classiques est faite. Enfin, nous procédons à une extension de certains des modèles de provisionnement ligne à ligne mis en place, les modèles sont adaptés selon les caractéristiques de la branche étudiée.

Remarque : Par soucis de confidentialité, les résultats présentés sur le montant des provisions/ultimes ont été obtenus après application d'un facteur multiplicatif.

Chapitre 1

Le provisionnement en assurance non-vie, le cas de la responsabilité civile

Dans ce chapitre nous nous intéressons à la responsabilité civile, nous rappelons les éléments clés du provisionnement en assurance non-vie et l'enjeu des provisions techniques. Ensuite, nous présentons les différentes méthodes classiques standards pour calculer ces provisions et leurs limites, ce qui nous amènera à présenter l'intérêt des modèles individuels. Enfin, nous définirons la problématique de ce mémoire.

1.1 La responsabilité civile

L'assurance est une opération par laquelle un assureur s'engage à fournir, dans le cadre d'un contrat, une prestation à l'assuré lors de la survenance d'un événement incertain et aléatoire en contrepartie d'une prime ou cotisation.

L'article R 321-1 du Code des Assurances cite 26 branches d'assurance, ces différentes branches peuvent être regroupées au sein de catégories plus vastes qui sont fondées sur deux critères différents :

1. Le mode de gestion de primes.
2. Le principe d'indemnisation des sinistres.

Le premier critère permet de distinguer entre "Assurance Non-Vie" et "Assurance Vie". Quant au deuxième il permet de différencier entre "Assurance Dommages" et "Assurance de Personnes".

Les primes collectées par les assureurs sont gérées ou bien par répartition ou bien par capitalisation et en fonction de la nature du risque assuré, l'assureur indemnisera ses assurés suivant le "principe indemnitaire" ou bien "le principe forfaitaire".

Dans la gestion par répartition, la totalité des prestations de l'exercice est assurée par les ressources de l'année (primes collectées au cours du même exercice), ces primes constituent ce que l'on appelle les "provisions techniques".

Quant au mode de gestion par capitalisation, les primes collectées sont capitalisées et les prestations servies sont prélevées sur des provisions constituées au fil des années.

Dans ce mémoire, nous nous intéréssons à la branche d'assurance Responsabilité Civile Corporelle Automobile qui fait partie des assurances dont les primes sont gérées par répartition et qui obéissent au principe indemnitaire.

1.1.1 L'assurance responsabilité civile

La responsabilité civile est engagée lorsqu'un acte volontaire ou non cause un dommage matériel ou corporel à autrui. La responsabilité civile peut être aussi engagée si l'acte est commis par l'un des dépendants de la personne assurée.

En assurance automobile, la garantie responsabilité civile est obligatoire et couvre les dommages causés aux tiers par l'assuré ou l'un de ses dépendants, ainsi que les dommages causés par le véhicule assuré sans l'intervention d'une personne.

Dans le cas de dommage corporel, la responsabilité civile ouvre droit à une indemnisation de la victime dès lors qu'il y a atteinte à l'intégrité physique ou morale de la personne. Pour que la victime puisse bénéficier d'une réparation, en plus de l'atteinte à l'intégrité physique ou morale de la personne, certaines conditions doivent être réunies :

- Existence d'une relation cause à effet entre l'accident ou l'agression et le dommage ;
- Le dommage doit être actuel ;
- La réparation concerne seulement la victime.

L'assurance de la responsabilité civile est fondée sur un principe indemnitaire. Ainsi, pour déterminer l'indemnisation nécessaire à la victime, une expertise médicale est nécessaire pour déterminer la présence du dommage et quantifier le préjudice.

1.1.2 Taux d'Atteinte à l'intégrité Physique et Psychique (AIPP)

Le taux d'AIPP (Atteinte à l'Intégrité Physique et Psychique), aussi nommé déficit fonctionnel permanent (DFP) correspond à l'invalidité que va conserver la victime à vie suite à un accident sur un plan physique ou sur un plan psychique. Elle a été introduite par la Confédération européenne d'experts en évaluation et réparation du dommage corporel (CEREDOC).

L'évaluation du taux d'AIPP se fait à partir d'un barème établi par un médecin expert. Le médecin expert de l'assureur va chiffrer une fourchette de taux d'AIPP dans le cas où la victime n'est pas encore consolidée c'est-à-dire que son état de santé n'est pas encore stable suite au dommage corporel, ce chiffrage permet à l'assureur d'évaluer son dossier et de savoir à priori combien il devra payer pour la victime. Après la consolidation de la victime, le taux AIPP est fixé définitivement.

Le taux d'AIPP est exprimé en pourcentage et dépend de plusieurs critères. L'Association Aide Indemnisation Victimes de France (AIVF) fourni plusieurs exemples de taux AIPP :[1]

- Perte d'un doigt (hors pouce) : 1 à 5% ;

1. `https://www.ornikar.com/assurance-auto/sinistre/assurance-accident/taux-aipp`

- Perte de toutes les dents ou de tous les orteils : 10 à 15% ;
- Perte d'un pied : 25 à 30% ;
- Perte d'une main : 30 à 50% (selon que la victime est droitière ou gauchère) ;
- Etc.

L'indemnisation financière de la victime repose sur trois critères : l'âge, le taux AIIP retenu et la région dont dépend la victime.

- **L'âge et le taux d'AIPP :** L'indemnisation est négativement corrélée avec l'âge de la victime de l'accident, car elle vivra plus longtemps avec des séquelles importantes et subira donc un plus grand préjudice ;
- **Le taux d'AIPP retenu :** L'indemnisation est positivement corrélée avec le taux AIPP, car plus le degré d'invalidité de la victime est élevé, plus cette dernière aura besoin d'aides pour compenser son invalidité ;
- **La région et le taux d'AIPP :** Les indemnisations diffèrent d'une région à une autre, car cela dépend des juges et de leurs décisions.

En assurance automobile, en cas d'accident entraînant une invalidité de la victime à vie, l'indemnisation va se déclencher à partir d'un certain seuil d'AIPP fixé par la compagnie d'assurance (chaque compagnie fixe un seuil d'AIPP). Si le taux d'AIPP chiffré par le médecin expert est inférieur au taux fixé par la compagnie d'assurance alors la victime de l'accident ne sera pas indemnisée. Le seuil d'AIPP fixé par les assureurs est souvent de 11%. [2]

1.1.3 La vie d'un sinistre corporel grave

Nous nous intéressons aux différentes étapes qui caractérisent la vie d'un sinistre corporel.

Un sinistre est caractérisé par une date de survenance, c'est-à-dire le moment où le dommage subi a eu lieu, puis l'accident va être déclaré à l'assureur et enclencher par la suite le processus d'indemnisation. La différence entre la date de déclaration et la date de survenance est appelée délai de déclaration.

Afin d'évaluer les dommages subis par chaque victime, une première expertise médicale déterminant la gravité des dommages corporels et le taux d'AIPP a lieu. À l'issue de cette expertise, une première évaluation du coût du sinistre est effectuée (plusieurs expertises peuvent être nécessaires). Puis, une deuxième expertise médicale définitive fixant la date de consolidation de la victime a lieu, suivie d'une réévaluation du sinistre, prenant en compte les conclusions médicales et les informations recueillies, notamment auprès des organismes sociaux. À l'issue de cette réévaluation, une offre définitive d'indemnisation est envoyée à la victime.

Enfin, un montant d'indemnisation est versé sous forme d'un capital ou de rentes temporaires ou viagères.

2. `https://www.index-assurance.fr/dictionnaire/seuil-aipp/`

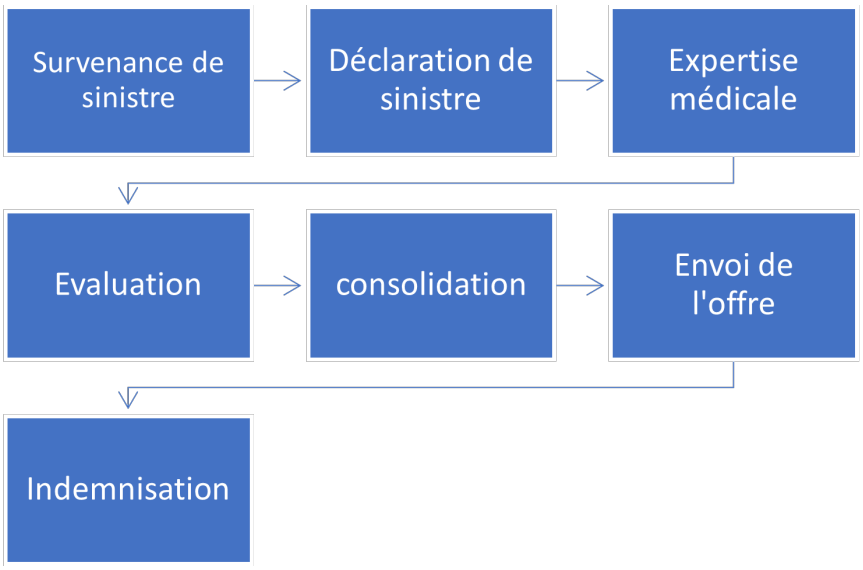


FIGURE 1.1 – La vie d'un sinistre corporel grave

1.2 Le provisionnement en assurance non-vie

Dans cette partie nous présentons quelques généralités sur l'assurance non-vie et nous définissons les provisions techniques.

1.2.1 La dynamique de la vie d'un sinistre

Pour pouvoir modéliser les paiements et provisions à constituer associés aux sinistres, il est intéressant d'aborder à présent la notion de la dynamique de la vie des sinistres.

La vie d'un sinistre se compose principalement de trois étapes cruciales. Selon un type de risque donnée (RC automobile, santé, marine, etc.), les sinistres sont constatés (avec un délai de constatation plus ou moins long), puis déclarés (là aussi avec un délai plus au moins long) et enfin payés avec un délai entre la déclaration et la date de paiement. Le processus de règlement d'un sinistre change selon le type de sinistre, sa gravité, etc. Pour les sinistre graves, l'évolution de la vie des sinistres va distinguer l'exercice de survenance, l'exercice de déclaration, l'exercice de dépassement de seuil et l'exercice de règlement de sinistre. Pour ce type de sinistres nous pouvons visualiser les différents aspects de leur vie sur la figure ci-dessous.

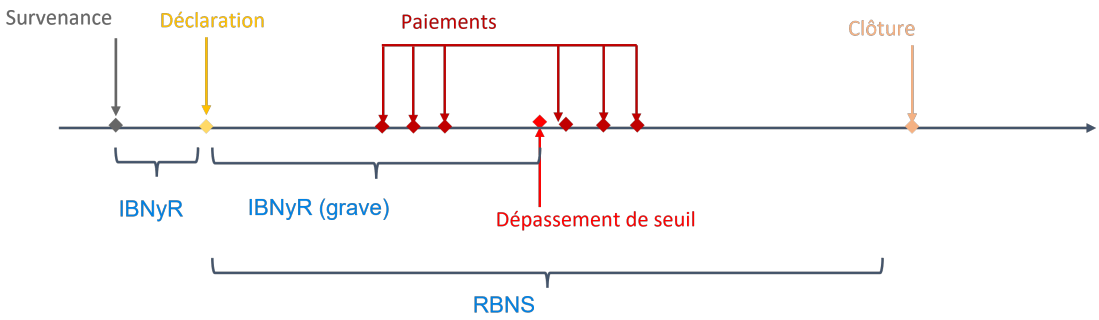


FIGURE 1.2 – Evolution d'un sinistre grave

Sur le graphique nous retrouvons les différentes étapes caractérisant la vie d'un sinistre grave depuis sa survenance jusqu'à sa clôture.

1.2.2 Les provisions techniques

La particularité de cycle de production inversé en assurance oblige l'assureur à fixer le montant de la prime avant de connaitre le coût réel des sinistres futurs. Cependant, l'assureur doit constituer des provisions techniques suffisantes pour pouvoir faire face à ses engagements futurs.

Les provisions techniques sont inscrites au passif du bilan d'une compagnie d'assurance et représentent le montant des engagements de l'assureur envers ses assurés.

Il existe différents types de provisions techniques, le Code des Assurances les définit dans l'article R331-6 :

- **Les provisions pour primes non acquises (PPNA) :** « provision, calculée selon les méthodes fixées par arrêté du ministre de l'économie, destinée à constater, pour l'ensemble des contrats en cours, la part des primes émises et des primes restant à émettre se rapportant à la période comprise entre la date de l'inventaire et la date de la prochaine échéance de prime ou, à défaut, du terme du contrat » ;
- **Les provisions pour risques en cours (PREC) :** « provision, calculée selon les méthodes fixées par arrêté du ministre de l'économie, destinée à couvrir, pour l'ensemble des contrats en cours, la charge des sinistres et des frais afférents aux contrats, pour la période s'écoulant entre la date de l'inventaire et la date de la première échéance de prime pouvant donner lieu à révision de la prime par l'assureur ou, à défaut, entre la date de l'inventaire et le terme du contrat, pour la part de ce coût qui n'est pas couverte par la provision pour primes non acquises » ;
- **Les provisions pour sinistres à payer (PSAP) :** « valeur estimative des dépenses en principal et en frais, tant internes qu'externes, nécessaires au règlement de tous les sinistres survenus et non payés, y compris les capitaux constitutifs des rentes non encore mises à la charge de l'entreprise ».

Pour un exercice comptable donné, nous pouvons définir la notion de charge ultime qui se compose des paiements déjà effectués pour les sinistres couverts pendant cet exercice, des provisions constituées pour couvrir les futurs paiements relatifs aux sinistres survenus

et déclarés et des provisions constituées pour couvrir les évolutions de ces provisions (IBNeR) ainsi que des provisions relatives aux sinistre non encore déclarés.

La figure ci-dessous présente les différentes composantes de la charge ultime (présente dans le cours de Dutang (2022)).

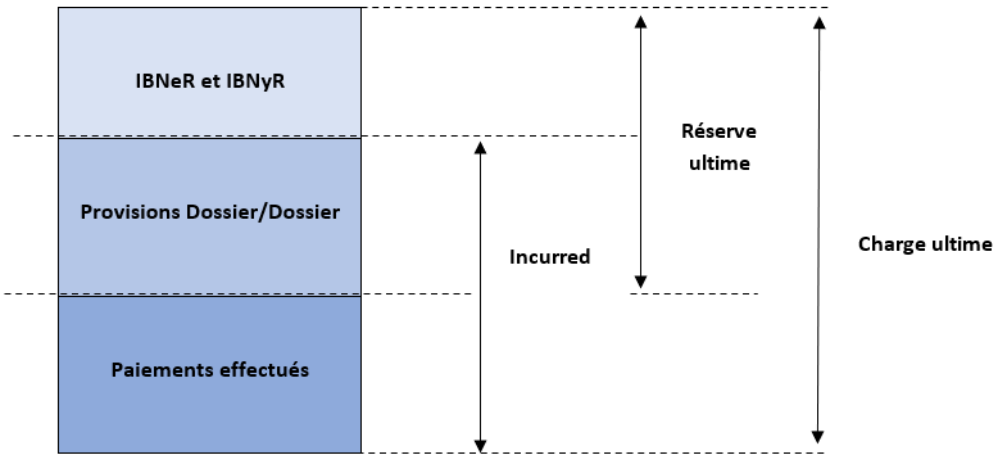


Figure 1.3 – Composantes de la charge ultime

Dans ce mémoire, nous nous intéressons uniquement au calcul de la charge ultime des RBNS.

1.2.3 Les triangles de liquidation

La dynamique des sinistres est reflétée par des triangles de liquidation sur lesquels les méthodes de provisionnement classiques sont basées. Les triangles dits run-off sont regroupés par année de survenance et par année de développement et permettent d'avoir une vision agrégée des sinistres à payer. Les méthodes standards de provisionnement, présentées dans la partie suivante, sont toutes basées sur des triangles de liquidation.

Nous définissons les notations suivantes :

- $\mathbf{i}$: correspond à la date de survenance des sinistres, avec $\mathbf{i=1,...,n}$;
- $\mathbf{j}$: correspond aux années de développement des sinistres, avec $\mathbf{j=1,...,n}$;
- $\mathbf{X_{i,j}}$: correspond aux montants des réglements réalisés l'année de développement $\mathbf{j}$ pour les sinistres survenus l'année $\mathbf{i}$;
- $\mathbf{C_{i,j}}$: correspond aux montants cumulé des réglements réalisés l'année de développement $\mathbf{j}$ pour les sinistres survenus l'année $\mathbf{i}$; $\mathbf{C_{i,j}} = \mathbf{X_{i,1}} + ... + \mathbf{X_{i,j}}$.

A partir de ces notations, nous pouvons représenter la sinistralité d'une branche par les triangles des paiements cumulés, ou incrémentaux comme ci-dessous :

$$\begin{bmatrix} C_{1,1} & C_{1,2} & \cdots & C_{1,j} & \cdots & C_{1,n-1} & C_{1,n} \\ C_{2,1} & C_{2,2} & \cdots & C_{2,j} & \cdots & C_{2,n-1} & \\ \vdots & \vdots & \cdots & \vdots & \iddots & & \\ C_{i,1} & C_{i,2} & \cdots & C_{i,j} & & & \\ \vdots & \vdots & \iddots & & & & \\ C_{n-1,1} & C_{n-1,2} & & & & & \\ C_{n,1} & & & & & & \end{bmatrix}$$

TABLE 1.1 – Triangle des paiements cumulés

ou

$$\begin{bmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,j} & \cdots & X_{1,n-1} & X_{1,n} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,j} & \cdots & X_{2,n-1} & \\ \vdots & \vdots & \cdots & \vdots & \iddots & & \\ X_{i,1} & X_{i,2} & \cdots & X_{i,j} & & & \\ \vdots & \vdots & \iddots & & & & \\ X_{n-1,1} & X_{n-1,2} & & & & & \\ X_{n,1} & & & & & & \end{bmatrix}$$

TABLE 1.2 – Triangle des paiements incrémentaux

Dans la suite de notre étude nous travaillons avec le triangle des paiements cumulés.

Une alternative possible est le calcul de triangle de charges (paiement cumulé + provision d/d à date), particulièrement pertinente lorsque nous travaillons sur des sinistres aux développements longs, qui mettent du temps à être réglés, et pour lesquels l'apport de l'information de la provision gestionnaire est cruciale.

Nous présentons maintenant les méthodes standards utilisées pour l'estimation des provisions.

1.3 Les méthodes classiques de provisionnement et leurs limites

Nous rappelons quelques méthodes classiques, la méthode déterministe de Chain-Ladder et la méthode stochastique de Mack. Ce rappel est nécessaire pour comprendre les avantages et les inconvénients relatifs à ces modèles ce qui nous amènera à présenter les avantages d'un modèle individuel.

1.3.1 La méthode déterministe de Chain-Ladder

La simplicité et la robustesse des méthodes déterministes expliquent leur utilisation sur le marché. Parmi ces méthodes, nous présentons la méthode de Chain-Ladder, une méthode très répandue car facile à comprendre et à mettre en œuvre.

La méthode de Chain-Ladder s'applique sur des triangles de règlement cumulés. Le principe de cette méthode repose sur l'estimation des facteurs de développement sur des données historiques (triangle des paiements cumulés) pour ensuite estimer les règlements futurs (le triangle inférieur).

Notations, hypothèses et estimation des provisions

Nous notons :

- $C_{i,j}$ **:** le montant cumulé à l'année de survenance i et l'année de développment j. i=1,...,n et j=1,...n.
- f_j **:** facteurs de développement. j= 1,...n.

La méthode de Chain-Ladder dans son cadre standard consiste à supposer que :

- **H1 :** Les $C_{i,j}$ sont indépendants pour les différentes années de survenance ;
- **H2 :** Il existe f_j tels que : $\mathbb{E}[C_{i,j}|C_{i,1},...,C_{i,j-1}] = \mathbb{E}[C_{i,j}|C_{i,j-1}] = f_{j-1} \times C_{i,j-1}$

L'estimation des facteurs de développement f_j se fait à partir des observations. Un estimateur naturel de ces facteurs est :

$$\hat{f}_j = \frac{\sum_{i=1}^{n-j} C_{i,j+1}}{\sum_{i=1}^{n-j} C_{i,j}}$$

A partir de ces estimateurs il est alors possible d'estimer les règlements futurs correspondant aux valeurs du triangle inférieur des paiements. L'estimation des paiements futurs est donnée pour i+j>n :

$$\hat{C_{i,j}} = \hat{f}_{n+1-i} \times ... \times \hat{f}_{j-1} \times \hat{C}_{i,n+1-i}$$

Une fois que les règlements futurs sont estimés, il est possible de connaitre l'estimation de la provision pour chaque année i, qui se calcul comme la différence entre ce que nous devrons payer et ce que nous avons déjà payé, c'est à dire :

$$\hat{R}_i = \hat{C}_{i,n+2-i} - C_{i,n+1-i}$$

Le montant de la provision global est donné par :

$$\hat{R} = \sum_{i=1}^{I} \hat{R}_i$$

$$\begin{bmatrix} C_{1,1} & C_{1,2} & \cdots & C_{1,j} & \cdots & C_{1,n-1} & C_{1,n} \\ C_{2,1} & C_{2,2} & \cdots & C_{2,j} & \cdots & C_{2,n-1} & \hat{C}_{2,n} \\ \vdots & \vdots & \cdots & \vdots & \ddots & \vdots & \vdots \\ C_{i,1} & C_{i,2} & \cdots & C_{i,j} & \cdots & \hat{C}_{i,n-1} & \hat{C}_{i,n} \\ \vdots & \vdots & \cdots & \vdots & \ddots & \vdots & \vdots \\ C_{n-1,1} & C_{n-1,2} & \cdots & \hat{C}_{n-1,j} & \cdots & \hat{C}_{n-1,n-1} & \hat{C}_{n-1,n} \\ C_{n,1} & \hat{C}_{n,2} & \cdots & \hat{C}_{n,j} & \cdots & \hat{C}_{n,n-1} & \hat{C}_{n,n} \end{bmatrix}$$

TABLE 1.3 – Triangle des paiements cumulés complété

1.3.2 La méthode stochastique de Mack

Le modèle de Mack est un modèle stochastique basé sur Chain-Ladder. C'est un modèle non paramétrique au sens où aucune hypothèse de distribution n'est faite sur les composantes du triangle. Cette méthode permet d'estimer les erreurs commises lors de l'évaluation des provisions.

Hypothèses et estimation des provisions

La méthode de Mack se base sur les hypothèses suivantes :

- **H1 :** Les années de survenances sont indépendantes entre elles, donc les $C_{i,j}$ sont indépendants pour les différentes années de survenance ;
- **H2 :** Il existe f_j tels que : $\mathbb{E}[C_{i,j}|C_{i,1},...,C_{i,j-1}] = \mathbb{E}[C_{i,j}|C_{i,j-1}] = f_{j-1} \times C_{i,j-1}$
- **H3 :** Il existe σ_j^2 tels que : $\mathbb{V}[C_{i,j}|C_{i,1},...,C_{i,j-1}] = \mathbb{V}[C_{i,j}|C_{i,j-1}] = \sigma_j^2 \times C_{i,j-1}$

A partir des hypothèses 1 et 2, Mack prouve, dans son article, que les estimateurs de Chain-Ladder $\hat{f}_j = \frac{\sum_{i=1}^{n-j} C_{i,j+1}}{\sum_{i=1}^{n-j} C_{i,j}}$ sont sans biais et non corrélés (MACK (1993)).

Autrement dit,

$$\mathbb{E}(\hat{f}_j) = f_j \text{ pour j = 1,...,n,}$$
$$Cov(\hat{f}_k, \hat{f}_k) = 0 \text{ pour j = 1,...,n, k = 1,...,n et j} \neq \text{k.}$$

A partir de ce résultat nous déduisons que l'estimateur des provisions de Chain-Ladder $\hat{R}_i = \hat{C_{i,n}} - \hat{C_{i,j}}$ est sans biais.

Maintenant, nous nous intéressons à l'erreur de prédiction associée à l'estimation des provisions. Pour ce faire, nous identifions un estimateur de la variance des coefficients de développement intervenant dans l'hypothèse 3. MACK (1993) propose l'estimateur suivant :

$$\begin{cases} \hat{\sigma_j}^2 = \frac{1}{n-j-1} \times \sum_{i=1}^{n-j} C_{i,j} \times \left(\frac{C_{i,j+1}}{C_{i,j}} - \hat{f}_j \right)^2 & \text{pour j = 1,...,n-2} \\ \hat{\sigma}_{n-1}^2 = \min \left\{ \frac{\hat{\sigma}_{n-2}^2}{\hat{\sigma}_{n-3}^2}, \min(\hat{\sigma}_{n-2}^2, \hat{\sigma}_{n-3}^2) \right\} & \text{pour j = n-1} \end{cases}$$

L'erreur de prévision

A présent, il est possible d'estimer l'erreur de prévision à partir de ces estimateurs. L'erreur de prévision peut être définie, pour chaque année de survenance i, comme la distance entre l'estimateur des provisions $\hat{R}_i$ et la valeur réelle R_i.

Une des distances utilisées pour mesurer cette erreur est l'erreur quadratique moyenne, ou MSEP (pour Mean Squared Error of Prediction). L'erreur quadratique moyenne du montant de provision pour l'année i est définie par :

$$MSEP(\hat{R}_i) = \mathbb{E}\left[(\hat{R}_i - R_i)^2\right]$$

Mack prouve qu'un estimateur de l'erreur quadratique de prédiction de la provision pour l'année de survenance i peut être obtenu par la formule suivante :

$$\hat{MSEP}(\hat{R}_i) = \hat{C}_{i,n}^2 \sum_{k=n-i+1}^{n-1} \frac{\hat{\sigma}_k^2}{\hat{f}_k^2} \times \left(\frac{1}{\hat{C}_{i,k}} + \frac{1}{\sum_{j=1}^{n-k} C_{j,k}}\right)$$

L'erreur quadratique moyenne de la provision totale est alors estimée par :

$$\hat{MSEP}(\hat{R}) = \sum_{i=2}^{n} \left(\hat{MSEP}(\hat{R}_i) + \hat{C}_{j,n} \times \left(\sum_{j=i+1}^{n-1} \hat{C}_{j,n}\right) \times \sum_{k=n-i+1}^{n-1} \frac{\frac{2\hat{\sigma}_k^2}{\hat{f}_k^2}}{\sum_{j=1}^{n-k} C_{j,k}} \right)$$

1.3.3 Limites des méthodes standards et avantages des modèles individuels

Malgré la simplicité de leur utilisation et la facilité de l'interprétation de leurs résultats, les méthodes standards présentent des inconvénients. L'un des inconvénients communs à la méthode déterministe de Chain-Ladder et la méthode stochastique de Mack est l'hypothèse d'indépendance entre les facteurs de développement et les années de survenance qui est une hypothèse forte car pour qu'elle soit vérifiée il faut avoir un historique de sinistres peu volatil et régulier.

De plus, en ce qui concerne la méthode de Chain-Ladder, sur les années de développement récentes pour lesquelles nous disposons de peu d'observations, l'incertitude autour des facteurs de développements estimés est élevée. De surcroît, cette méthode ne permet pas d'obtenir une mesure de l'erreur d'estimation.

Contrairement à Chain-Ladder, le modèle de Mack permet de mesurer la variabilité des estimations, et donc d'obtenir un intervalle de confiance. En revanche, elle ne permet pas d'obtenir la distribution des provisions sans faire appel à des hypothèses supplémentaires.

Enfin, les méthodes standards montrent leurs limites quand les règlements sont volatils ou quand les sinistres sont des sinistres longs. Dans ces cas de figure, les hypothèses sous-jacentes ne sont souvent plus vérifiées, ce qui remet en cause la validité des résultats obtenus avec ces méthodes.

Ainsi, peu à peu, des méthodes complémentaires ont vu le jour pour trouver une solution à ces limites. Parmi ces méthodes, le provisionnement ligne à ligne.

Un modèle individuel permet de considérer l'ensemble des caractéristiques importantes des sinistres, autrement dit, au contraire des méthodes standards qui utilisent les données agrégées, ces méthodes permettent de ne perdre aucune information. La contribution fondamentale des méthodes individuelles est leur flexibilité dans le champ d'application et l'intégration des informations détaillées qu'elles fournissent.

Enfin, ces modèles permettent la séparation de la provision IBNyR (Incurred But Not Yet Reported) de la provision RBNS (Reported But Not Settled).

1.4 Problématique

L'objet de ce mémoire consiste à donner un aperçu global des méthodes de provisionnement ligne à ligne existantes et d'en comparer certaines, entre elles et avec les méthodes classiques Chain-Ladder et Mack dans le cadre d'une application sur un portefeuille de Responsabilité Civile Corporelle Automobiles constitué de sinistres graves. Nous nous intéressons au provisionnement des sinistres ouverts ayant déjà franchis un certain seuil de charge mais non encore clôturés, notés RBNS. Les modèles sont adaptés en fonction des caractéristiques de la branche étudiée. La problématique principale de ce mémoire est d'évaluer les avantages et les inconvénients des modèles de provisionnement ligne à ligne, et de voir dans quelle mesure nous pouvons les utiliser en pratique, en remplacement ou en complément des méthodes agrégées.

Chapitre 2

État de l'art : Provisionnement ligne à ligne

Depuis plus d'un siècle, les actuaires, en assurance non-vie, utilisent des triangles de liquidation pour projeter les paiements futurs. Au cours des dernières décennies différents travaux sur des méthodes de provisionnement ligne à ligne ont été menés au sein de la communauté actuarielle et scientifique. Dans les méthodes ligne à ligne, chaque sinistre est considéré individuellement et toutes ses caractéristiques sont considérées pour prédire l'ultime. Plusieurs méthodes ont été proposées, chacune est spécifique pour un type de sinistres donné. Dans ce chapitre nous allons présenter l'état de l'art des modèles ligne à ligne existants.

Dans un premier temps, nous allons présenter la famille des modèles paramétriques en distinguant les modèles des IBNyR de ceux des RBNS. Ensuite, les modèles non-paramétriques seront présentés. Enfin, nous allons parler d'un modèle paramétrique qui considère une approche Payment-to-Payment.

2.1 Modèles paramétriques

Un cadre probabiliste approprié pour le provisionnement ligne à ligne a été introduit pour la première fois par Arjas (1989), Jewell (1989) puis suivi par d'autres études par Norberg (1993) et Hesselager (1994). Arjas (1989) développe un cadre mathématique utilisant la théorie des processus ponctuels et des martingales. Jewell (1989) formule un modèle bayésien en temps continu pour prédire le nombre total des IBNyR survenant dans un intervalle d'exposition donné, lorsque seul un nombre incomplet de ces revendications a été signalé à un moment donné. Norberg (1993) construit un processus stochastique continu décrivant la survenance et le développement des sinistres. Enfin, Hesselager (1994) donne une version continue du modèle de Hachemeister (1980) qui suggère de représenter l'information sur un sinistre non clôturé en modélisant son développement comme une réalisation d'une chaîne de Markov en temps discret.

2.1.1 Incurred But Not Yet Reported (IBNyR)

Lorsque nous nous intéressons aux IBNyR, le délai entre la date de survenance du sinistre et le moment de la déclaration est un élément crucial dans l'étude. Larsen (2007)

revisite le travail de Norberg (1993), Haastrup et Arjas (1996) en incluant des caractéristiques de sinistres pour spécifier les composantes du modèle. Il présente un ensemble de modèles stochastiques basés sur des hypothèses moins fortes comme l'indépendance entre les montants incrémentaux agrégés. Ces modèles sont capables de gérer la saisonnalité ainsi que les évolutions du portefeuille (type d'affaires) et de la sinistralité (type de sinistres, taille de sinistres). Antonio et Plat (2014), proposent d'appliquer le cadre de modélisation élaboré dans Norberg (1993) et Norberg (1999) à un portefeuille d'assurance responsabilité civile d'une compagnie d'assurance européenne et d'améliorer l'évaluation des réserves IBNR en utilisant un processus de Poisson marqué dépendant de la position. Dans leur étude empirique, et dans le contexte de cette étude de cas spécifique, il est démontré que le provisionnement ligne à ligne fournit une meilleure précision par rapport aux modèles agrégés sélectionnés. Badescu, Sheldon et Dameng (2016a) et Badescu, Sheldon et Dameng (2016b) étudient la modélisation des sinistres individuels au moyen d'un processus de Cox ; dans ce travail, l'intensité stochastique est cruciale pour tenir compte de la dépendance temporelle entre l'arrivée et le règlement des sinistres.

Parmi les auteurs qui ont considéré l'étude des délais de déclaration comme première étape dans le process de calcul d'une réserve IBNyR appropriée nous pouvons également nous référer à Guiahi (1986), Jewell (1989) et Zhao et Zhou (1986). Ces derniers, présentent un modèle ligne à ligne pour le développement des sinistres en utilisant des techniques (semi-paramétriques) issues de l'analyse de survie et des méthodes de copules pour étudier la dépendance entre le délai de déclaration et la date de survenance. Egalement, Verrall et Wüthrich (2016) fournissent des informations supplémentaires sur l'analyse et le calibrage de la distribution des temps d'arrivées et les délais de déclaration basées sur des données réelles, en effet nous observons un gain dans l'utilisation des méthodes ligne à ligne les dans des environnements caractérisés par une non-stationnarité. Plus récemment, Boumezoued et Devineau (2017) revisitent les formulations probabilistes originales de Norberg (1993) et Hesselager (1994) et élaborent un cadre de modélisation des survenances et des délais de déclaration des sinistres puis fournissent une présentation cohérente de la modélisation (avec simulation et formules fermées) des IBNyR.

2.1.2 Reported But Not Settled (RBNS)

Lorsque l'objet d'étude est les RBNS, la durée avant la clôture du sinistre et l'ultime sont étudiés séparément comme dans Ayuso et Santolino (2008) ou en utilisant un modèle multi-états pour modéliser le développement du sinistre comme dans Antonio, Godecharle et Oirbeek (2016) et dans Boumezoued et Devineau (2017). Ces derniers fournissent une présentation cohérente de la modélisation (avec simulation et formules fermées) des historiques de sinistres individuels ainsi que des quantités agrégées en tant que réserve globale pour les RBNS. Le modèle est construit sur une composante principale qui régit le cheminement des paiements de la déclaration jusqu'à la clôture. En revanche, un biais d'estimation peut être observé lorsqu'il s'agit de garanties avec un délai de réglement long. Cela est dû au phénomène de censure que nous pouvons observer dans certains cas (ex : absence de sinistres longs dans la base de données utilisée pour le calibrage de la distribution des montants de sinistres).

Afin de combler ces lacunes, Lopez (2018), étudie l'importance de développement des sinistres à travers le temps dans le cas des RBNS en construisant un modèle qui permet de comprendre la dépendance entre le temps avant la clôture du sinistre et son montant. L'un des éléments clés de la méthode est la modélisation de la dépendance entre le montant et la durée à l'aide d'une copule. Ainsi, plus le sinistre met du temps à être clôturé plus le montant est élevé. Le modèle est similaire à celui étudié dans Lopez, Milhaud et Thérond (2016), à la seule différence qu'ici la variable censure est considérée observable. Cette hypothèse s'explique par le fait que la seule raison de la présence de la censure est due à la fin de la période d'observation. Ceci permet d'obtenir des résultats asymptotiques plus précis de l'estimation. Les résultats du cas pratique sont très intéressants, en revanche aucune information sur des paiements partiels ou une réévaluation durant le développement n'a été considérée, ces derniers peuvent être pris en compte dans le modèle en les considérant comme des co-variables sans trop changer la structure de la procédure. Les sinistres considérés sont supposés être déclarés instantanément, ce qui dans la pratique n'est pas le cas : l'article propose d'incorporer une troncature à gauche dans le modèle pour corriger ce problème.

Pour évaluer l'incertitude liée aux réserves estimées par les méthodes standard et les modèles ligne à ligne Charpentier et Pigeon (2016) étudient les propriétés théoriques des modèles économétriques (Gaussien, Poisson et quasi-Poisson) sur des données individuelles et des données agrégées. Une application aux réserves de sinistres est présentée.

2.2 Modèles non-paramétriques

Les techniques de machine learning sont très flexibles pour le traitement des données structurées et non-structurées, ainsi ces techniques sont de plus en plus demandées en assurance.

Wüthrich (2018) propose pour la première fois une contribution pour illustrer comment les méthodes des arbres de régression peuvent être utilisées dans le cadre du provisionnement ligne à ligne. Il considère uniquement le nombre de paiements, et non pas les règlements. De plus, il suppose que la survenance des sinistres peut être décrite par un processus de Poisson Homogène Marqué dépendant de la position ; par conséquent le nombre des IBNyR peut être prédit par une méthode de Chain-Ladder.

Dans cet aspect, BAUDRY et ROBERT (2019) proposent un nouveau modèle non paramétrique pour estimer séparément les IBNyR et les RBNS. Cette approche peut :

1. Inclure les principales caractéristiques des sinistres afin de tenir compte de l'hétérogénéité des sinistres et de tirer parti d'importants ensembles de données ;
2. Saisir le schéma de développement spécifique des sinistres, y compris leur occurrence, et les caractéristiques de flux de trésorerie ;
3. Détecter les changements de tendance potentiels, en tenant compte des changements possibles dans le mix produit, le contexte légal ou le traitement des sinistres dans le temps, pour éviter les biais dans l'estimation et la prévision.

Un algorithme ExtraTrees a été utilisé dans le cas pratique, les résultats obtenus montrent que l'estimateur des réserves est sans biais avec un écart type petit comparé à la

méthode de Mack Chain-Ladder. De plus, les estimations du modèle de Machine Learning sont plus robustes et réagissent à tout changement dans les cadences de développement des sinistres, en comparaison aux estimations de la méthode standard.

Duval et Pigeon (2019) proposent des modèles de provisionnement en assurance non-vie, combinant l'approche traditionnelle comme celle de Mack ou un GLM et un algorithme de Boosting dans un cadre individuel. Les modèles sont entrainés sur deux types de données :

1. Entrainement du modèle sur les sinistres clôturés seulement, c'est-à-dire que les données censurées ne sont pas considérées. En revanche, ce choix va générer un biais de sélection car les sinistres clôturés juste avant la date d'évaluation vont avoir un développement plus court et donc un montant total payé petit. Par conséquent, le modèle va sous-estimer le montant global des paiements de certains sinistres. De plus, il va y avoir une perte d'information importante qui va aussi mener à une sous-estimation des montants ;
2. La deuxième approche consiste à développer les sinistres non-clôturés à la date d'évaluation en utilisant des approches classiques telles que le modèle de Mack ou les GLM.

Afin de corriger le biais causé par la prise en compte des sinistres clôturés seulement, une approche dite « inverse probability of censoring weighting » (IPCW) peut être considérée, elle est proposée dans l'article de Lopez, Milhaud et Thérond (2016).

Avec la croissance de la collecte des données des sinistres individuels, et l'amélioration des méthodes de stockage ainsi que la puissance de calcul, il devient intéressant d'envisager des formes sophistiquées de Machine Learning telles que les réseaux de neurones profonds (NN). Ces derniers nécessitent peu de restrictions et d'hypothèses, intègrent des tendances non linéaires complexes et ont des performances prédictives élevées.

Des NN avec diverses architectures ont été récemment appliqués à la provision des sinistres individuels. Wüthrich (2018) et Taylor (2019) étudient l'évolution récente des modèles ligne à ligne impliquant des NN. Gabrielli, Richman et Wüthrich (2020) proposent d'utiliser un NN avec un modèle de régression, tel qu'un Poisson sur-dispersé, pour mieux s'aligner sur la pratique actuarielle traditionnelle. Andrea (2021) présente un NN effectuant des tâches simultanées de régression et de classification pour la prédiction des paiements futurs. La synthèse de l'historique des sinistres sont les entrées de son réseau, tandis que Delong et Wüthrich (2020) utilisent les historiques entiers des sinistres dans leur NN pour prédire le développement conjoint de la survenance des sinistres et des paiements individuels.

En s'appuyant sur ces travaux, Delong, Lindholm et Wüthrich (2021) présentent plusieurs NN pour estimer le montant de la réserve des sinistres survenus et déclarés et ceux survenus mais non déclarés avec des architectures plus petites, réduisant ainsi le temps de calcul.

Une autre façon de considérer les historiques de sinistres passés est d'utiliser les réseaux de neurones récurrents (RNN), une classe très populaire de NN introduite par Hopfield (1982). Hochreiter et Schmidhuber (1997) ont introduit les réseaux LSTM (Long Short Term Memory), une classe de RNN, pour éviter l'explosion du gradient.

Kuo (2020) propose un réseau de densité de mélange bayésien multi-périodes basé sur les LSTMs. Malheureusement, ce modèle n'améliore pas la précision des prédictions par rapport à la méthode classique de Chain-Ladder.

Chaoubi et al. (2022) développent un nouveau modèle de provisionnement ligne à ligne applicable à tout ensemble de sinistre individuel avec un développement long et il se concentre sur l'étude des RBNS. Le modèle implique un réseau LSTMs qui effectue deux tâches afin de prédire les paiements futurs attendus : une classification pour déterminer la probabilité de paiements dans une période donnée, et une régression pour prédire le paiement incrémental. De plus, pour améliorer la prédiction des montants de paiements importants, ils conçoivent une approche de réserve qui combine la sortie LSTM et une distribution de Pareto généralisée.

2.3 Modèles payment-to-payment

Contrairement aux modèles paramétriques présentés au début du chapitre où l'approche considérée est par période de développement, Pigeon (2014) propose un modèle de provisionnement ligne à ligne dans un cadre paramétrique et en temps discret avec une approche Payment-to-Payment.

Dans un premier temps, il développe un modèle de paiement individuel (individual Paid (iP)) qui utilise uniquement les informations relatives aux montants réglés. Inspiré par Arjas (1989), Norberg (1993) et Antonio et Plat (2014), ils considèrent une structure micro-temporelle. Pour le développement de chaque sinistre, ils adaptent la structure multiplicative du modèle de Mack au modèle ligne à ligne. Contrairement au modèle de Mack où l'approche considérée est par période de développement (development-to-development), l'approche du modèle ligne à ligne est payment-to-payment c'est à dire qu'au lieu de considérer l'évenement de développement du sinistre il considère l'évènement "paiement partiel réalisé".

Dans un second temps, il généralise le modèle iP à un modèle iPIC (individual Paid and Incurred Chain model) qui inclut les montants de paiements et de charge au niveau micro. Pour lier ces deux sources d'informations, il considère un lien comme celui présenté dans Merz et Wüthrich (2010) et Happ et Wüthrich (2013).

Les modèles sont développés dans un cadre paramétrique. Il utilise la famille de distribution MSS (Multivariate Skew Symmetric).

Chapitre 3

Zoom sur les modèles ligne à ligne implémentés

Dans ce chapitre, nous présentons la théorie des modèles ligne à ligne implémentés et testés sur la base de données RCC Automobile. Pour pouvoir comparer deux grandes familles de modèles, nous avons testé différents modèles de provisionnement utilisant des données détaillées, aussi bien un modèle paramétrique, défini par Boumezoued et Devineau (2017), que des modèles non paramétriques.

Le choix de ces modèles s'est globalement basé sur la nature des sinistres et la simplicité des modèles. En effet, le modèle paramétrique est plus récent et adapté aux sinistres aux développements longs et les modèles non-paramétriques sont très simples, ce qui nous permet d'utiliser le même modèle selon deux approches différents : dans un premier temps les modèles sont entrainés sur les sinistres clos uniquement, puis dans un deuxième temps les mêmes modèles sont entraînés sur les sinistres clos et RBNS pour exploiter au maximum toute l'information de la base de données ; deux algorithmes différents (bagging et boosting) sont implémentés afin de les comparer et voir si plusieurs algorithmes peuvent être complémentaires.

3.1 Modèle stochastique à états

Nous reprenons le cadre d'étude défini par Boumezoued et Devineau (2017).

Nous considèrons que les sinistres surviennent à des dates $(T_n)_{n\geq 1}$ suivant un processus de Poisson d'intensité $\lambda(t)$ et sont déclarés avec des délais de déclaration $(U_n)_{n\geq 1}$ qui sont supposés suivre une loi $p_{u|Tn}$. Nous nous plaçons sur un horizon de temps d'observation $[0, \tau]$.

A la date d'observation τ, l'assureur observe les sinistres survenus tels que $T_n+U_n \leq \tau$, cela signifie qu'une partie des sinistres n'est pas encore observée : ce sont les IBNR qui ne serons pas étudiés dans le cadre de ce mémoire.

Dans la suite de notre étude nous allons nous intéresser au processus de développement des sinistres survenus et observés aux temps $(T_n^R)_{n\geq 1}$ définis par l'ensemble :

$$I^R(\tau) = \{(T_n, U_n) \text{ tels que } T_n + U_n \leq \tau\}$$

3.1.1 Modélisation du développement des sinistres

L'objectif est de décrire le processus de développement des réglements de sinistres depuis la déclaration jusqu'à la clôture.

Nous considérons que le processus de développement des sinistres est décrit par une variable aléatoire (V_k) ; en chaque temps aléatoire, une marque (E_k) est générée. (E_k) prend ses valeurs dans $\{1, 2, 3\}$ où :

- $E_k = 1$: indique que le sinistre est clôturé, en V_k, sans paiement.
- $E_k = 2$: indique que le sinistre est clôturé, en V_k, avec paiement.
- $E_k = 3$: indique qu'un paiement est enregistré en V_k, sans clôture du sinistre.

Les événements 1, 2, 3 se produisent, depuis la déclaration des sinistres, selon des fonctions d'intensité spécifiques : h_1, h_2 et h_3. Si un événement du type 2 ou 3 a lieu, alors un flux de paiement P_k est généré avec une distribution qui peut dépendre du temps V_k et de type d'évenement (sinistre clos ou non).

Modélisation avec un processus de Markov

Nous considérons un processus de Markov X_t qui prend ses valeurs dans l'espace $\mathbb{N}^*$. Si un saut est observé en V_k, alors :

- $X_{V_k} = 1$: indique que le sinistre est clôturé, en V_k, sans paiement,
- $X_{V_k} = 2$: indique que le sinistre est clôturé, en V_k, avec paiement,
- $X_{V_k} = \mathrm{j}$, pour $\mathrm{j} \geq 3$: indique qu'un paiement est enregistré en V_k, sans clôture du sinistre.

Nous considérons que $X_0 = 3$ et que les intensités de transition pour $\mathrm{j} \geq 3$ sont données par :

- $\lambda_{j,1}(t) = h_1(t)$,
- $\lambda_{j,2}(t) = h_2(t)$,
- $\lambda_{j,j+1}(t) = h_3(t)$.

Toutes les autres intensités de transition sont nulles.

3.1.2 Formules fermées pour le développement d'un sinistre

Hesselager (1994) décrit la trajectoire d'un sinistre comme étant un processus de Markov continu avec un espace d'états $\mathbb{N}$. En se basant sur les résultats de cet article, Boumezoued et Devineau (2017) dérivent des formules pour l'espérance et la variance dans le cadre d'un processus de Markov (non-homogène).

Notations

- $X^{(t)}(u,v)$: représente le paiement total, dans $[t+u, t+v]$, relatif à un sinistre survenu en t ;
- $S^t(u)$: processus décrivant la trajectoire des états des sinistres (après déclaration) et prenant ses valeurs dans $\mathbb{N}^*$, avec une probabilité de transition $p_{mn}^{(t)}$ et une intensité de transition associée $\lambda_{mn}^{(t)}(u)$;
- $Y_{mn}^{(t)}(u)$: paiement réalisé pour une transition $m \to n$ à la date u, avec une moyenne $y_{mn}^{(t)}(u)$ et un écart type $\sigma_{mn}^{(t)}(u)$

Paiement attendu le long d'une trajectoire d'un sinistre

Dans le cadre de ce modèle, il est clair que $X(u,\infty)$ est égal à zéro sachant que le sinistre est clôturé, c'est-à-dire sachant que $S(u) = 1$ ou $S(u) = 2$. Ainsi, il suffit d'étudier $X(u,\infty)$ sachant $S(u) = 3$. En se basant sur la Proposition 3 en annexe, le paiement attendu s'écrit comme suit :

$$\mathbb{E}[X(u,\infty)|S(u)=3] = \int_u^\infty \{h_2(v)y_2(v) + h_3(v)y_3(v)\} \left\{ \sum_{m\geq 3} p_{3m(u,v)} \right\} \mathrm{d}v \tag{3.1}$$

Notons que $\sum_{m\geq 3} p_{3m(u,v)}$ est la probabilité que le processus de Markov reste dans l'ensemble 3,4,..., dont les transitions vers les états 1 et 2 sont respectivement h_1 et h_2 ; ainsi :

$$\sum_{m\geq 3} p_{3m(u,v)} = \exp(-\int_s^u (h_1(v) + h_2(v))\mathrm{d}v) \tag{3.2}$$

Cela conduit au résultat suivant :

Proposition 1. *Le paiement futur espéré après un certain temps s pour une trajectoire d'un sinistre individuel peut s'écrire de la manière suivante :*

$$\mu(s) = \mathbb{E}[X(u,\infty)|S(u)=3] = \int_u^\infty \left\{ (h_2(v)y_2(v) + h_3(v)y_3(v)) \exp(-\int_s^u (h_1(v) + h_2(v))\mathrm{d}v) \right\} \tag{3.3}$$

Variance des paiements de sinistre

En se basant sur la Proposition 3 en annexe, la variance des paiements de sinistre s'écrit comme ci-dessous :

Proposition 2. *La variance d'un paiement total après un certain temps s pour une trajectoire d'un sinistre individuel peut s'écrit comme suit :*

$$\gamma(s) = Var[X(u,\infty)|S(u)=3] = \int_s^\infty H(u) \exp(-\int_s^u (h_1(v) + h_2(v))\mathrm{d}v)\mathrm{d}u \tag{3.4}$$

Avec : $H = h_1\mu^2 + h_2(\sigma_2^2 + (y_2 - \mu)^2) + h_3(\sigma_3^2 + y_3^2)$ et $\mu = \mu(s)$ (3.3).

Remarque 1. *Dans un cadre Markovien homogène, où $h_i(s) \equiv h_i$ et $y_i(s) \equiv y_i$, le paiement futur espéré ainsi que la variance associés au paiement total après un certain temps s s'écrivent :*

$$\mu(s) = \mu(0) = \frac{y_2 h_2 + y_3 h_3}{h_1 + h_2} \tag{3.5}$$

et

$$\gamma(s) = \gamma(0) = \frac{h_1\mu^2 + h_2(\sigma_2^2 + (y_2 - \mu)^2) + h_3(\sigma_3^2 + y_3^2)}{h_1 + h_2} \tag{3.6}$$

avec $\mu = \mu(s)$

3.1.3 Simulation de la distribution des paiement futurs

Grâce au modèle individuel et aux paramètres estimés, nous sommes en mesure de déterminer la distribution de la provision totale en considérant un algorithme de simulation.

Nous détaillons ci-dessous le détail des étapes de l'algorithme de simulation :

- **Étape 1 :** Pour chaque cluster, nous générons un vecteur des états $i \in \{1, 2, 3\}$ de taille égale au nombre des RBNS en effectuant un tirage avec remise, avec probabilité $p_i = \frac{h_i}{h_1+h_2+h_3}$.
- **Étape 2 :** A partir du vecteur généré, nous récupérons les indices des états associés à l'évènement de clôture, ie. l'indice des états 1 et 2. Cela nous donne un vecteur des indices de clôture des RBNS.
- **Étape 3 :** Ensuite, nous tronquons ce vecteur des indices de clôture pour avoir autant d'indices de clôture que de RBNS présents dans la base.
- **Étape 4 :** Dans cette étape, nous créons un nouveau vecteur des états en récupérant, depuis le vecteur généré à l'étape 1, les états allant de 1 jusqu'à la valeur récupérée à l'étape 3. L'ensemble de ces événements correspond à une simulation du nombre de paiements sans clôture, paiements avec clôture, et clôtures sans paiements, associés aux RBNS de la base.
- **Étape 5 :** Ensuite, à partir du nouveau vecteur des états, nous récupérons les indices des états associés aux événements de paiement, ie. les indices des états 2 et 3.
- **Étape 6 :** Enfin, nous simulons la distribution des paiements selon une loi calibrée sur les paiements observés de taille égale à la taille du vecteur créé à l'étape 5.

3.2 Modèles non-paramétriques

Contrairement aux modèles paramétriques, les modèles non-paramétriques permettent de prédire l'ultime payé en fonction des covariables, sans avoir à définir une forme structurelle pour les paiements. La variable réponse à apprendre est l'ultime qui n'est disponible que pour les sinistres clos.

Parmi les différents modèles envisagés pour la prédiction de l'ultime, nous avons choisi d'appliquer l'algorithme Extreme Gradient Boosting, aussi appelé XGBoost, et l'algorithme Random Forest en considérant deux types de données pour l'entraînement du modèle.

1. Entraînement du modèle sur les sinistres clôturés seulement, c'est-à-dire que nous n'allons pas considérer les données censurées.
2. La deuxième approche consiste à développer les sinistres non-clôturés à la date d'évaluation en utilisant les paramètres de l'approche classique de Mack Chain-Ladder .

Afin de prédire le montant de l'ultime, nous allons considérer toutes les visions de sinistre à chaque année de développement. La figure ci-dessous représente les visions des sinistres considérées pour les modèles :

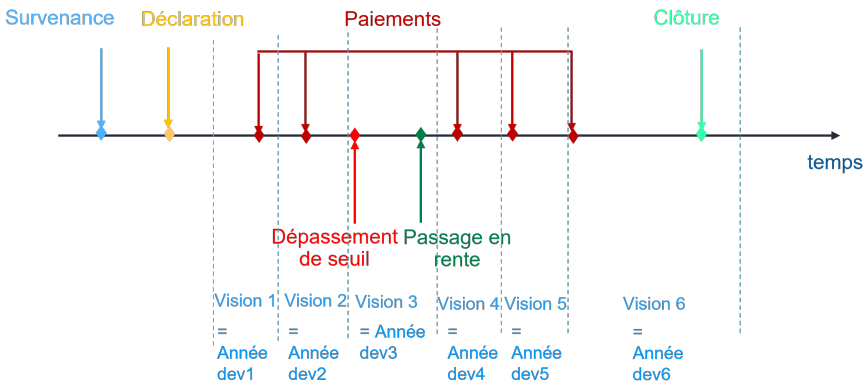


FIGURE 3.1 – Visions des sinistres

Ainsi, nous disposons des variables prédictives supplémentaires qui dépendent du temps :

- Dossier passé en rentes ou non
- Temps depuis le passage en rentes
- Nombre de paiements passés
- Montant cumulé payé
- Temps depuis l'ouverture du sinistre
- Temps depuis le dépassement du seuil de 500k
- Provision dossier/dossier

3.2.1 L'algorithme XGBoost : une extention du Gradient Boosting

XGBoost (ou contraction de Extreme Gradient Boosting) est une méthode de machine learning supervisée pour la classification et la régression. Le modèle est basé sur l'apprentissage d'ensemble séquentiel et les arbres de décision. Il repose sur le principe de gradient boosting qui est un modèle de boosting qui s'adapte et tente de s'autocorriger à chaque

ittération, il s'agit d'un ensemble d ' appreneurs faibles, créés les uns après les autres, formant ensuite un appreneur fort, chaque appreneur faible est entraîné pour corriger les erreurs des appreneurs faibles précédents. XGBoost est conçue pour être efficace, flexible et portable. Il a été introduit par Chen et Guestrin (2016).

L'algorithme XGBoost est un algorithme optimisé et régularisé. Afin de comprendre le fonctionnement de cet algorithme nous allons nous référer à l'article publié en 2016 sur la théorie mathématique de XGBoost Chen et Guestrin (2016). Nous considérons un ensemble de données $\mathcal{D} = \{(x_i, y_i)\}$ ($|\mathcal{D}| = n, x_i \in \mathbb{R}^m, y_i \in \mathbb{R}$).

La variable cible est prédite en utilisant K fonctions additives comme suit :

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F} \tag{3.7}$$

avec $\mathcal{F} = \{f(x) = \omega_{q(x)}\}$ ($q : \mathbb{R}^m \to T, \omega \in \mathbb{R}^T$), l'espace des arbres de régression (CART).

- q : représente la structure de chaque arbre.
- T : nombre de feuilles dans l'arbre.
- f_k : correspond à une structure d'arbre q.
- ω : poid associé aux feuilles.

Pour chaque feuille, à chacun des arbres de décision, est associé un score. Nous notons ω_i le score associé à la i-ème feuille.

Afin d'apprendre l'ensemble des fonctions utilisées dans le modèle, nous minimisons la fonction objectif suivante :

$$\mathcal{L}(\phi) = \underbrace{\sum_{i=1}^{n} l(y_i, \hat{y}_i)}_{\text{Perte sur l'entrainement}} + \underbrace{\sum_{k=1}^{K} \Omega(f_k)}_{\text{Régularisation}} \tag{3.8}$$

- Perte sur l'entrainement : mesure la qualité de prédiction du modèle sur la base d'entrainement.
- Régularisation : mesure la compléxité des arbres.

Un compromis entre la perte d'entrainement et la perte de régularisation a lieu c'est à dire lorsque nous cherchons à optimiser la perte d'entrainement nous allons avoir des prédictions plus précises sur nos données d'entrainement. Cependant, si nous optimisons la régularisation nous allons avoir un modèle plus simple qui s'adaptent aux données.

Afin de déterminer la complexité des arbres, l'algorithme XGBoost se base sur deux paramètres : le nombre de feuille et la norme L2 du poids de chaque feuille.

$$\Omega(f_t) = \underbrace{\gamma T}_{\text{nombre de feuilles}} + \underbrace{\frac{1}{2}\lambda \sum_{j=1}^{T} \omega_i^2}_{\text{norme L2 du poid de la feuille}} \tag{3.9}$$

où : γ et λ des hyperparamètres.

Optimisation de la fonction objectif

Maintenant que nous avons notre fonction objectif, nous allons chercher ω_j^* qui représente le poids optimal pour une feuille j. Pour ce faire, nous allons ajouter f_t dans l'expression de la fonction objectif.

Soit : ${\hat{y}_i}^{(t)}$ la prediction du i-ème instance à la t-ème itération. Ainsi, la fonction objectif à la t-ème ittération peut s'écrire comme suit :

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, {\hat{y}_i}^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{3.10}$$

Afin d'optimiser la fonction, nous allons utiliser l'approximation de Taylor au second ordre. La fonction objectif s'écrit alors comme suit :

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n} [l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \tag{3.11}$$

Avec : $g_i = \partial_{\hat{y}^(t-1)} l(y_i, \hat{y}^{(t-1)})$ et $h_i = \partial^2_{\hat{y}^(t-1)} l(y_i, \hat{y}^{(t-1)})$

Nous pouvons ne pas prendre en compte le terme constant. Ainsi, sur l'ensemble des feuilles $I_j = \{i | q(x_i) = j\}$, notre fonction objectif s'écrit :

$$\tilde{\mathcal{L}}^{(t)} = \sum_{j=1}^{T} [G_j \omega_j + \frac{1}{2}(H_j + \lambda)\omega_j^2] + \gamma T \tag{3.12}$$

Avec : $G_j = \sum_{i \in I_j} g_i$ et $H_j = \sum_{i \in I_j} h_i$

Ainsi pour une structure fixé $q(x)$, nous pouvons calculer le poid optimal ω_j^* associé à la j-ème feuille :

$$\omega_j^* = -\frac{G_j}{H_j + \lambda} \tag{3.13}$$

La valeur optimale correspondante est donnée par :

$$\tilde{\mathcal{L}}^{(t)} = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \tag{3.14}$$

Cette fonction permet d'évaluer la qualité d'une structure q. Plus la valeur retournée par la fonction objectif est faible, plus la structure q est bonne.

En pratique, il est impossible de tirer toutes les structures q pour construire l'arbre optimal. Un algorithme qui part d'une feuille unique et ajoute itérativement des branches à l'arbre est utilisé à la place. Chen et Guestrin proposent une mesure de réduction de perte induite par une division donnée. En notant I_L et I_R les noeuds candidats (respectivement gauche et droit) et $I = I_L \cup I_R$, la réduction de perte s'écrit :

$$\mathcal{L}_{split} = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma \tag{3.15}$$

Cette formule est utilisée en pratique pour sélectionner la meilleure division lors de la construction de chaque arbre.

3.2.2 L'algorithme Random Forest

Random Forest est un algorithme d'apprentissage automatique supervisé qui est largement utilisé dans les problèmes de régression et de classification. Dans le cadre de notre étude, les forêts aléatoires permettent de traiter de façon individuelle chaque sinistre, via une modélisation ligne à ligne.

Comme son nom l'indique, la forêt aléatoire se compose de nombreux arbres de décision. Plutôt que de dépendre d'un arbre, il prend la prédiction de chaque arbre et, sur la base des votes majoritaires des prédictions, prédit la sortie finale.

Random Forest combine plusieurs arbres de décision aléatoires et agrège leurs prédictions en faisant la moyenne. Il utilise à la fois le Bagging contraction de bootstrap aggregation, et la classification et les arbres de régression (CART).

Le bagging est un méta-algorithme d'apprentissage ensembliste conçu pour améliorer la stabilité et la précision des algorithmes d'apprentissage automatique utilisés en classification et régression. Cela réduit également la variance et aide à éviter le surapprentissage. Il génère des échantillons bootstrap à partir de l'ensemble de données d'origine et construit un prédicteur pour chaque échantillon. La prédiction finale prend en effet en considération l'ensemble des modèles entraînés pour réaliser sa prédiction finale.

L'algorithme de Random Forest est une technique qui consiste à construire N arbres à partir de N sous-ensembles du jeu de données initial. Les étapes suivantes résument le fonctionnement de l'algorithme Random Forest :

- **Étape 1 :** Tout d'abord, des sous-ensembles sont créés à l'aide d'un tirage aléatoire avec remise sur le jeu de données initial.
- **Étape 2 :** Ensuite, l'algorithme construira un arbre de décision pour chaque échantillon. Puis, il obtiendra le résultat de prédiction de chaque arbre de décision.
- **Étape 3 :** Dans cette étape, le vote sera effectué pour chaque résultat prédit.
- **Étape 4 :** Enfin, il sélectionne le résultat de prédiction le plus voté comme résultat de prédiction final.

Pour l'optimisation du Random Forest, les hyper-paramètres optimisés sont ceux du modèle CART et ceux qui sont propres au Random forest comme le nombre d'arbres créés.

3.2.3 Le stacking

Dans ce qui précède, nous avons vu deux algorithmes de machine learning qui reposent sur deux principes différents : le Bagging et le Boosting. Quand nous sommes amenés à travailler avec des modèles de types différents, les modèles peuvent s'avérer plus précis d'un cas à un autre. Ainsi, il serait utile de combiner les prévisions issues de différents modèles afin de tirer partie des forces de chacun des modèles, ceci est possible avec la méthode de Stacking. Le Stacking est une méthode qui consiste à appliquer un algorithme de machine learning à des classifieur générés par un autre algorithme de machine learning (Sill et al. (2009)). Autrement dit, il permet de combiner des modèles construits avec des algorithmes différents.

Chapitre 4

Exploration des données : sinistres graves en Responsabilité Civile Corporelle Automobile

Nous avons précédemment présenté les différents modèles que nous souhaitons mettre en oeuvre afin d'estimer le montant de la provision au titre des RBNS. La base sur laquelle nous implémentons nos modèles a été fournie par un grand assureur français et concerne la branche Responsabilité Civile Corporelle Automobile. Plus particulièrement, cette base est constituée des sinistres graves ayant dépassé un certain seuil de charge.

Dans ce chapitre, nous présentons dans un premier temps la base de données. Ensuite, nous effectuons des analyses de données puis, en nous restreignant à notre périmètre d'étude, nous appliquons des retraitements de données pour obtenir une base adaptée à l'implémentation des modèles choisis.

4.1 Données disponibles

Nous disposons de deux bases de données renseignant les sinistres corporels graves de la branche Auto.

La première base contient les développements de sinistres en RCC Automobile pour l'ensemble des victimes ayant dépassé le seuil de 500k€ de charge (désigne la charge sinistre qui se calcule comme la somme de la provision dossier/dossier calculée par le gestionnaire sinistre et le paiement cumulé). Nous disposons d'une ligne par mouvement : révision de charge, paiement, ou recours. Par conséquent, pour un même sinistre, nous pouvons avoir plusieurs lignes par année de développement.
La base est constituée de 2792 dossiers victimes distincts associés à 2693 sinistres distincts survenus entre le 23 Juin 1966 et le 21 Décembre 2019. Nous pouvons classer les variables disponibles par type de variables.

Le tableau ci-dessous présente les variables de la base :

Catégorie	Variables
Dates	Date de survenance de sinistre Date de déclaration de sinistre Date du mouvement généré (Date vision)
Véhicule	Âge du véhicule Usage du véhicule Puissance fiscale du véhicule Nombre de conducteurs designés État du véhicule au moment de l'accident (véhicule en stationnement ou non)
Conducteur et victime	Sexe du conducteur Sexe de la victime Catégorie socio-professionnelle Taux de responsabilité de l'assuré envers la victime
Sinistre	Statut du sinistre (clos ou ouvert) Circonscription dans laquelle l'accident s'est déroulé
Lieu	Code département Code commune Code INSEE du lieu du garage
Montants	Paiement décumulé Provision dossier/dossier (calculée par le gestionnaire sinistre) Paiement cumulé depuis l'origine Charge

TABLE 4.1 – Variables présentes dans la base de données

La deuxième base de données contient des informations supplémentaires en termes de variables et d'observations.

Concernant les variables, ont été ajoutées : la variable « Date de clôture du sous-dossier victime » et les variables de paiements suivantes :

- rente : le montant de la rente versé quand le dossier victime passe en rente ;
- règlement (Capital yc recours) : règlements effectués y compris les recours ;
- autres règlements : tous les autres règlements effectués.

La somme de ces trois variables de paiement vaut la variable Paiement (initialement présente dans la première base).

Concernant les observations, ont été ajoutés les sous-dossiers non graves associés à des sinistres identifiés comme graves et des sinistres ayant occasionné plusieurs victimes non-graves, mais qui, additionnés, forment un sinistre grave.

Les deux bases ont été fusionnées. *Pour toute la suite du mémoire, nous travaillons à la maille des sous-dossiers victime. Par souci de simplification, nous désignerons par la suite un sous-dossier victime par l'appellation "sinistre".*

4.2 Traitement des données

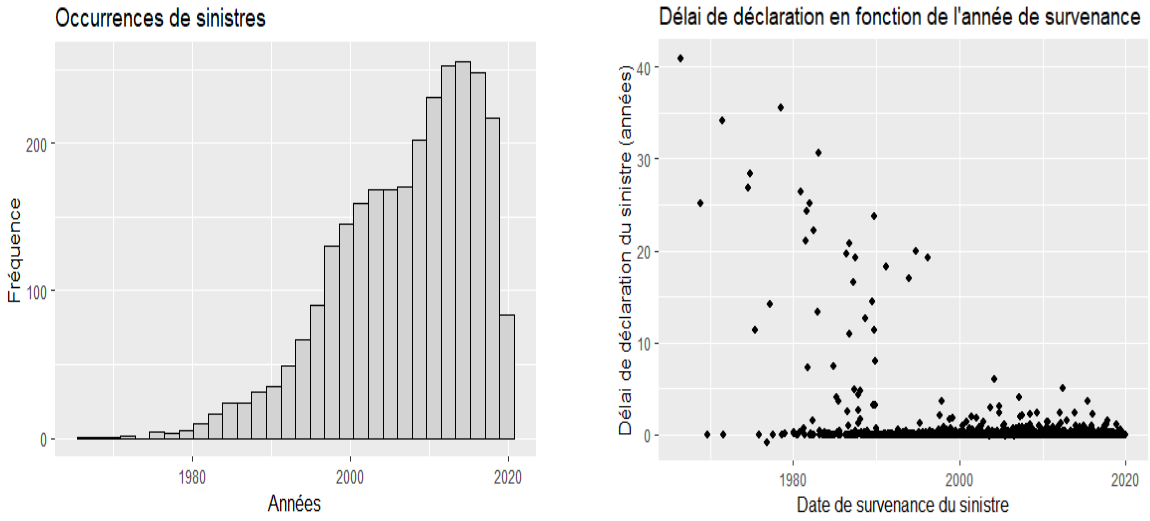
Avant même de manipuler les bases de données dans le but d'observer des statistiques descriptives ou encore d'appliquer les modèles présentés dans le Chapitre 3, il est nécessaire de définir notre périmètre d'étude et de nous assurer que nous travaillons avec des données exhaustives, exactes et appropriées.

4.2.1 Périmètre de l'étude

Afin d'éviter tout biais d'observation des victimes non graves, nous choisissons de nous restreindre aux sous-dossiers victimes graves : ainsi, les sous-dossiers victimes non graves rattachés à un sinistre grave ne sont pas considérés. Nous choisissons de travailler à la maille victime grave pour la modélisation de l'ultime.

Nous considérons, dans le cadre de notre étude, uniquement les sinistres survenus à partir du 01/01/1996 et dont le délai de déclaration est inférieur à 15 ans, pour deux raisons :

- La première raison étant l'observation de peu de sinistres survenus avant cette date. Par ailleurs, les sinistres survenus avant 1996 sont ceux avec des délais de déclaration anormalement longs, comme nous pouvons le remarquer sur le graphique de droite ci-dessous.

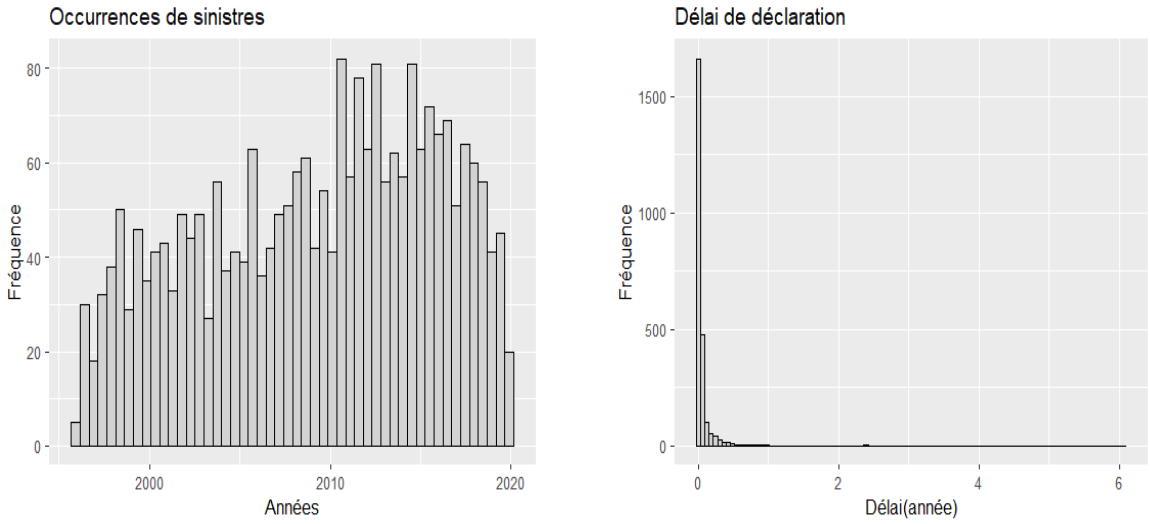


(a) Survenance des sinistres (b) Délai de déclaration

FIGURE 4.1 – Survenance et déclaration des sinistres de la base brute

- La deuxième raison repose sur le fait que cette restriction permet d'être en ligne avec le périmètre de calcul des triangles de projection calculés en interne chez l'assureur.

Ainsi, avec ces restrictions sur la date de survenance des sinistres et leur délai de déclaration, dans notre base nous observons les distributions de survenance et de délais de déclaration ci-dessous :



(a) Survenance des sinistres

(b) Délai de déclaration

FIGURE 4.2 – Survenance et déclaration des sinistres de la base retreinte au périmètre d'étude

4.2.2 Qualité des données

Après avoir défini notre périmètre d'étude et afin de pouvoir manipuler les données, il est nécessaire de s'assurer de la qualité des données. Nous voulons nous assurer qu'il n'existe pas de données manquantes pour des variables majeures (comme la date de survenance par exemple) et que les variables ne présentent pas d'incohérence. Pour ce faire, nous allons effectuer une série de contrôles sur des variables clés. Ces contrôles sont répertoriés ci-dessous.

Contrôle 1

Nous devons nous assurer qu'il n'y a pas de valeurs manquantes pour des variables majeures tels que : date de survenance, date de déclaration, date de règlement ainsi que les variables relatives aux flux enregistrés.

Contrôle 2

La date de survenance doit être antérieure à la date de déclaration. Nous vérifions si c'est bien le cas.

Contrôle 3

La date de clôture doit être postérieure à la date de dernier flux de règlement. Nous devons nous assurer qu'il n'y a pas de réouverture de dossier.

Contrôle 4

La charge ultime relative aux dossiers clos doit être positive. Nous vérifions ceci pour tous les dossiers clos.

Contrôle 5

Nous devons nous assurer que toutes les lignes de la base enregistrent un mouvement sur au moins une de ces variables : paiement, révision de charge ou passage en rentes. En effet, certaines lignes peuvent correspondre à des mouvements sur d'autres

garanties, mais seraient des doublons pour l'évolution des paiements/des charges au titre de la garantie RCC.

Contrôle 6

Nous vérifions le signe des montants de règlements et réserves en distinguant leurs natures : les recours doivent être négatifs tandis que les autres montants doivent être positifs.

A l'issue de l'application de ces contrôles, nous appliquons les retraitements suivants :

- Modification de la date de déclaration des sous-dossiers victimes pour lesquels nous observons une date de déclaration antérieure à la date de survenance. Nous imposons que la date de déclaration pour ces dossiers soit égale à la date de sinistre.
- Suppression des dossiers réouverts. Dans le cas où le mouvement enregistré après la clôture du dossier est un recours, nous avons décidé de retraiter les recours comme suit : si après la clôture du dossier nous observons des recours (1 à 3 recours) nous regroupons ces derniers dans le dernier mouvement observé en date de clôture. Sinon, nous supprimons le dossier réouvert : trois dossiers ont été supprimés à la suite de ce retraitement.
- Suppression des lignes sans mouvement.

4.2.3 Variables retenues pour catégoriser les sinistres

A la suite des contrôles effectués nous obtenons une base de données composée de 2460 dossiers victimes graves, dont 55% dossiers sont clos, répartis sur 2381 sinistres. La base restreinte au périmètre de modélisation contient 33 variables en tout, certaines de ces variables ne sont pas réellement utiles pour notre étude consistant en la modélisation de l'ultime, qui correspond au dernier paiement cumulé des sinistres clos. Nous pouvons ainsi supprimer ces variables.

Parmi les 33 variables nous avons conservé les variables qui sont indispensables pour le provisionnement, notamment les dates et les montants. Cependant, nous avons supprimé toutes les variables de lieu, car elles présentent trop de modalités.

Dans le but de prédire la charge ultime liées aux sinistres graves dans le cadre d'un modèle de provisionnement ligne à ligne, nous avons besoin, en plus des variables définies dans la section (4.1), de créer de nouvelles variables utiles pour notre études. Notamment, des variables caractérisant l'évènement de dépassement de seuil de 500k€ de charge que nous présenterons en détail dans la section (4.3).

Parmi les nouvelles variables créées, en voici quelques-unes :

- Dates : date de premier dépassement de seuil de 500k€ de charge, date de passage en rente.
- Délais : délai de déclaration, délai entre la déclaration et le passage en rente, délai entre déclaration du sinistre et premier dépassement de seuil.
- Variable caractérisant le développement de sinistre : nombre d'années de développement de sinistre depuis sa déclaration.

- Position : la position de dossier victime à chaque date de vision (charge en dessous ou au-dessus du seuil de 500k€) et la dernière position enregistrée à la date de clôture, pour les sinistres clos, ou à la dernière date de vision, pour les sinistres en cours.
- Passage de seuil : une variable indicatrice, affectant, pour chacun des dossiers de sinistre, la valeur 1 aux visions pour lesquelles la date de vision de sinistre est supérieure ou égale à la date de premier dépassement de seuil de 500k€ de charge et 0 sinon.

4.2.4 Transformation de la base en vision annuelle

Nous transformons à présent notre base de données en vision annuelle pour faciliter l'implémentation des modèles envisagés.

La vision de sinistre renseignée dans la base de données est mensuelle, en revanche nous n'avons pas exactement une ligne par mois, nous avons plutôt une ligne par mouvement.

Ainsi nous avons choisi de transformer les dossiers en vision annuelle en agrégeant, pour chaque dossier victime et pour chaque année de développement, les paiements intermédiaires enregistrés.

Nous obtenons les distributions de paiements et de recours non nuls ci-dessous :

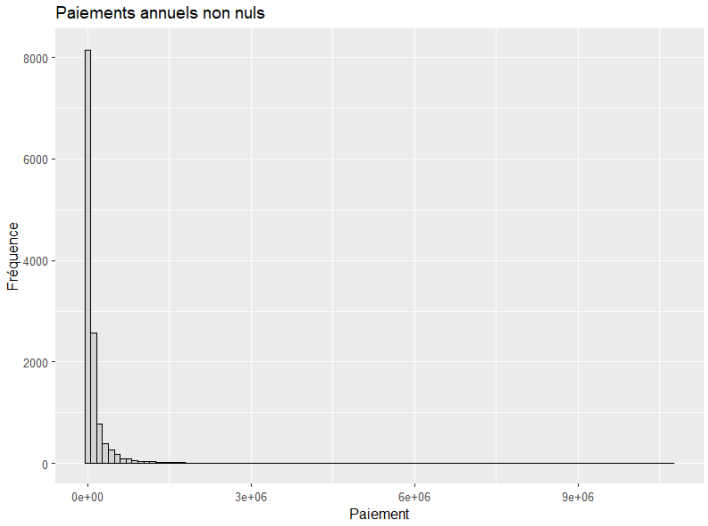


FIGURE 4.3 – Paiements annuels non nuls

Sur les 2460 dossiers nous en comptons 22 avec des recours, la distribution des montants de recours annuels non nuls est présentée dans la figure suivante : nous constatons que le montant des recours est très faibles par rapport aux paiements, nous sommons donc directement les recours et les paiements, que nous modélisons ensemble. Dans le cas où les recours auraient été plus conséquents, nous aurions opté pour une modélisation séparée des recours et des paiements.

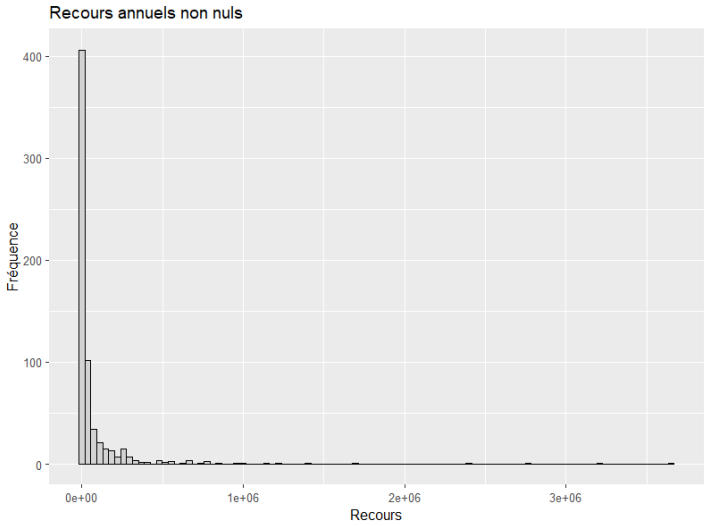
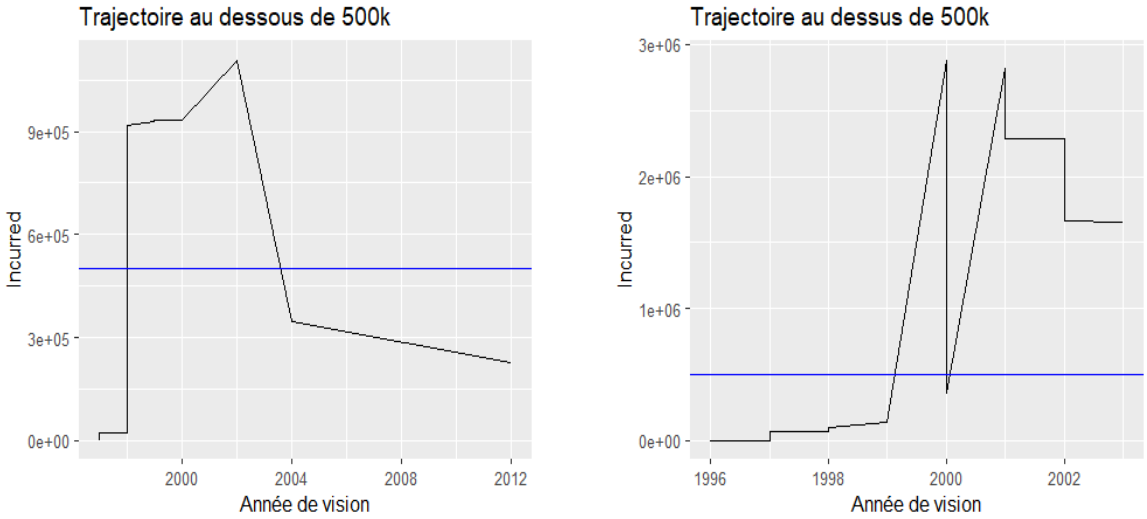


FIGURE 4.4 – Recours annuels non nuls

4.3 Dynamique de passage en seuil sur les sinistres clos

Au cours de l'exploration de la base de données, nous nous sommes rendu compte que 52% des sinistres graves clos ont un ultime inférieur à 500k €. Ceci nous amène à étudier les trajectoires des charges afin de comprendre les cadences de règlements et de révision de charge qui expliquent ces ultimes relativement faibles pour des sinistres graves.

L'étude sur les différentes trajectoires de charge montre qu'il existe deux principaux types de trajectoires : les sinistres qui restent graves jusqu'à leur clôture et ceux qui dépassent le seuil mais qui ensuite redescendent en dessous et y restent jusqu'à leur clôture.



(a) Trajectoire d'un sinistre cloturé au dessous du seuil de 500k € de charge

(b) Trajectoire d'un sinistre cloturé au dessus du seuil de 500k € de charge

FIGURE 4.5 – Exemple de deux trajectoires types de charge

Notons que, pour un même sinistre, il peut y avoir un seul ou plusieurs passages au dessus/en dessous du seuil (selon les révisions de charges/recours).

Cette dynamique de passage du seuil critique de 500k € fera l'objet d'une étude spécifique (voir Chapitre 6).

4.4 Clustering supervisé sur les sinistres clos

Nous cherchons à présent à décrire intelligemment la base de données, afin de comprendre le lien entre les co-variables et l'ultime. Pour ce faire, nous avons décidé de construire des groupes homogènes de dossiers victimes, en termes de sinistralité. Ainsi, nous construisons des clusters sur les dossiers victimes clos : nous avons choisi d'utiliser l'algorithme CART.

La construction de ces clusters nous sera très utile pour affiner les estimations des modèles, en particulier celles du modèle à états, comme nous le verrons dans le chapitre suivant. En effet, ces clusters nous permettront d'effectuer un calibrage du modèle pour chacun des groupes construits, aussi distinguer les paiements futurs attendus en fonction des clusters.

Pour réaliser cette segmentation, nous utilisons l'ensemble des co-variables disponibles pour chacun des sous-dossiers victimes clos ; la variable cible est la classe d'ultime payé, construite à partir de la variable dernier paiement cumulé (qui correspond à l'ultime pour les sinistres clos). Nous avons deux classes d'ultimes : «Ultime inférieur à 500k€» et «Ultime supérieur à 500k€». Ce choix de classe d'ultime est motivé par l'observation de trajectoires de règlements différents selon la valeur de l'ultime.

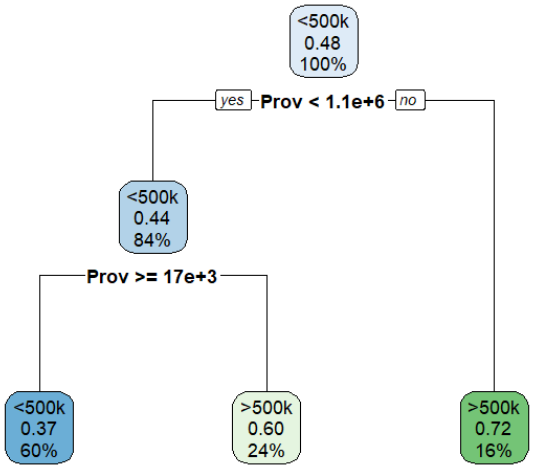


FIGURE 4.6 – Arbre CART

L'arbre de décision nous donne trois clusters qui se distinguent les uns des autres principalement par le montant de la provision dossier/dossier.

L'arbre simplifié construit en figure (4.6) se lit de la façon suivante : chaque noeud est représenté par une classe d'ultime «Ultime inférieur à 500k€» et «Ultime supérieur à 500k€». La racine est ainsi scindée en deux noeuds suivant que la provision est inférieure ou non à 1,1 M€. Puis le noeud de gauche obtenu est à son tour scindé suivant la valeur du montant de provision (inférieure ou non à 17 k€).

Ces trois groupes ont un délai de déclaration relativement homogène (figure 4.7) et un ultime significativement différent (figure 4.8).

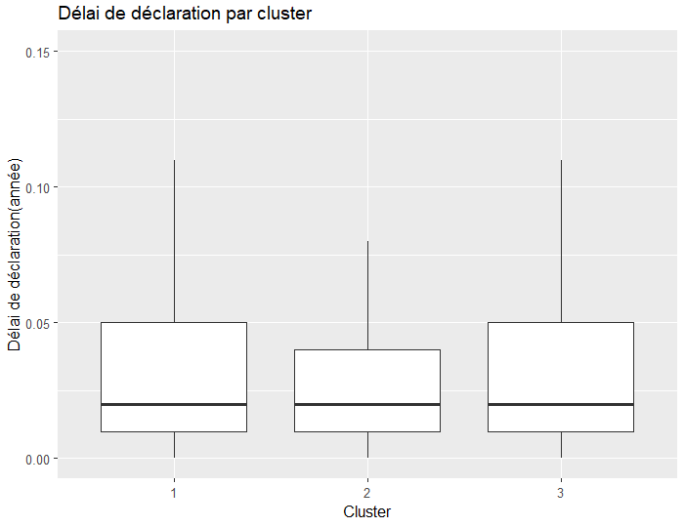


FIGURE 4.7 – Delai de déclaration en fonction des clusters

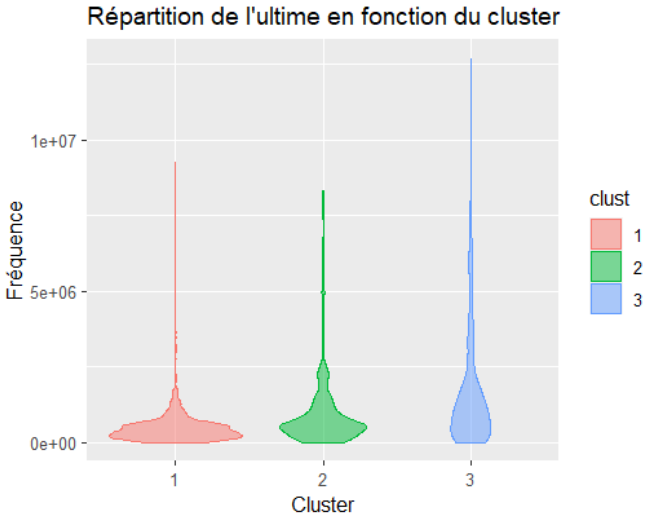


FIGURE 4.8 – Repartition de l'ultime en fonction des clusters

Nous constatons que le cluster 3 rassemble des sinistres qui ont en moyenne une sévérité plus élevée que dans le cluster 1 et 2. Cependant, nous avons une forte variabilité des ultimes par classe, et en particulier pour le cluster 3.

Cluster	Nombre de sinistres	Moyenne(Ultime)	Sd(Ultime)	CV(Ultime)
1	1485	481 073	678 726	1.4
2	479	967 249	1 351 442	1.4
3	496	1 528 954	2 245 832	1.5

TABLE 4.2 – Répartition des ultimes en fonction des clusters

Afin de mieux comprendre le schéma type des sinistres graves selon leur ultime (s'il en existe un), nous nous intéressons à la distribution du délai entre la déclaration du sinistre et le premier dépassement de seuil sur les sinistres clos en fonction des clusters.
Nous constatons que les sinistres du cluster 3, donc en moyenne les sinistres les plus graves, se caractérisent par un délai entre déclaration et premier dépassement de seuil très court (voire nul). Nous pouvons en déduire que les sinistres les plus graves sont très rapidement identifiés comme tels par les gestionnaires sinistres, qui leur affectent une charge supérieure à 500k€ dès la première année de développement pour la plupart. Le cluster 2 quant à lui regroupe les sinistres pour lesquels nous enregistrons un délai entre déclaration et premier dépassement de seuil long : ce sont des sinistres qui auront en moyenne un ultime supérieur à 500€, mais qui ont mis du temps à être identifiés comme graves. Enfin, les sinistres du cluster 1 (les sinistres les moins graves) mettent en moyenne un an avant d'être identifiés comme graves par les gestionnaires sinistres.

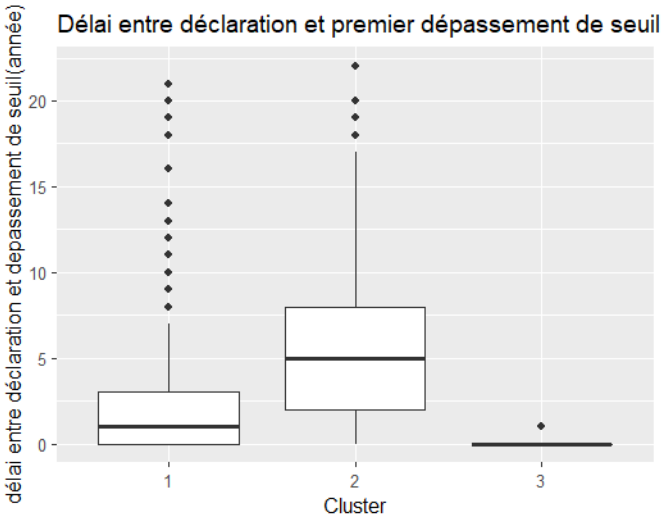


FIGURE 4.9 – Delai entre déclaration et dépassement de seuil en fonction des clusters

Enfin, regardons également la dernière position des dossiers clos, depuis le dernier passage en seuil jusqu'à la clôture, en fonction des clusters. C'était, nous le rappelons, la variable cible de l'arbre de décision calibré.
En cohérence avec la construction de l'arbre, environ 60% des sinistres sont clos avec un ultime inférieur à 500k€, contre seulement 35% et 25% environ pour les clusters 2 et 3.

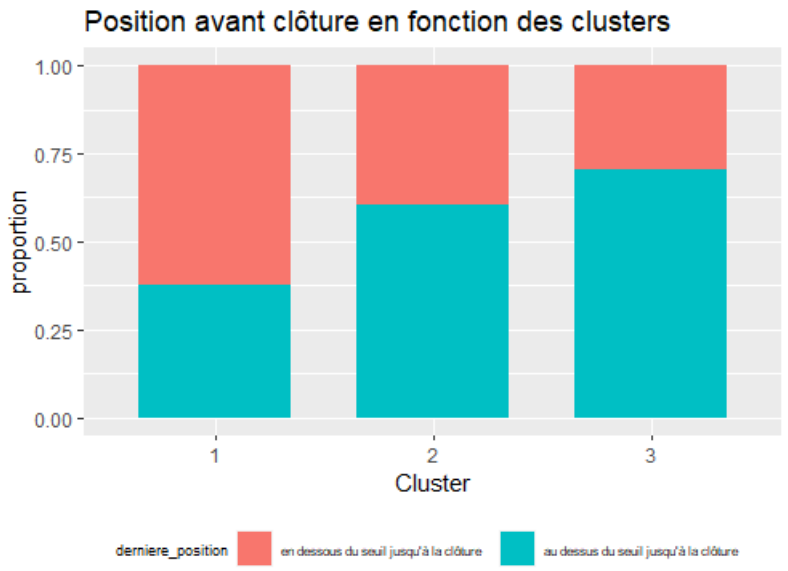


FIGURE 4.10 – Position en fonction des clusters

4.5 Analyse de la charge déterminée par le gestionnaire sinistre

Avant d'implémenter nos modèles sur la base de données nous souhaitons étudier la charge ultime déterminée par le gestionnaire de sinistres afin d'avoir une base de comparaison des charges ultimes qui seront prédites par les modèles sélectionnés.

Dans un premier temps nous nous intéressons aux erreurs d'estimation de la charge ultime commises par le gestionnaire de sinistres sur les sinistres clos par année de développement. La figure suivante présente les erreurs en question :

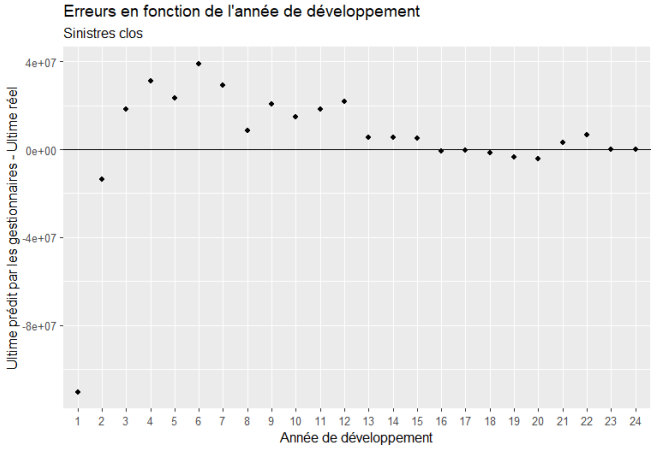
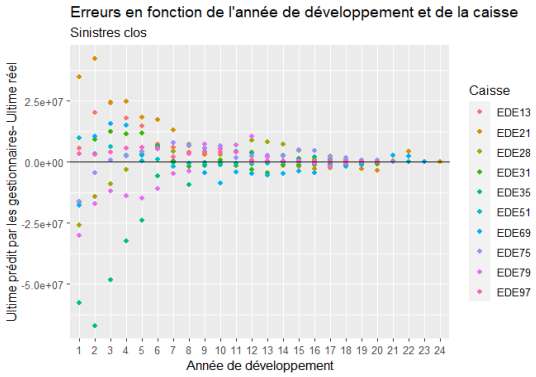


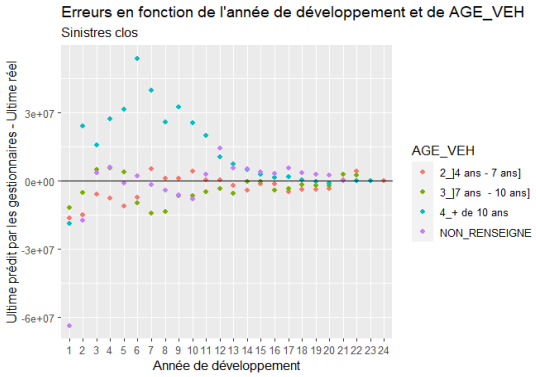
FIGURE 4.11 – Erreurs d'estimation de la charge ultime par le gestionnaire sinistre en fonction de l'année de développement

Comme nous pouvons le constater sur la figure, le gestionnaire de sinistres surestime la charge ultime sur la majorité des années de développement.

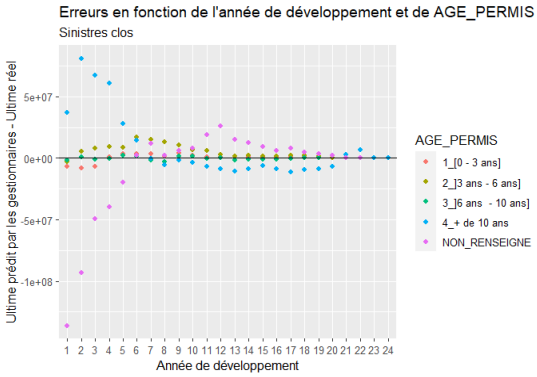
Afin d'affiner cette étude, nous réalisons une analyse descriptive afin de déterminer sur quels segments le gestionnaire de sinistres se trompe le plus. Les figures ci-dessous présentent les erreurs de prédiction de l'ultime en fonction des années de développement et en fonction de certaines variables catégorielles.



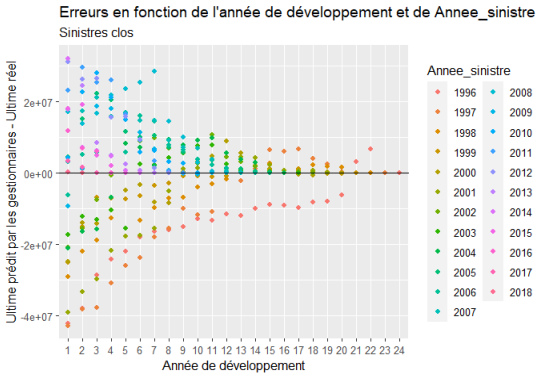
(a) Erreurs en fonction de l'année de développement et de la caisse



(b) Erreurs en fonction de l'année de développement et de l'âge du véhicule



(c) Erreurs en fonction de l'année de développement et de l'âge du permis



(d) Erreurs en fonction de l'année de développement et de l'année de survenance du sinistre

FIGURE 4.12 – Erreurs d'estimation de la charge ultime par le gestionnaire sinistre en fonction de l'année de développement et de certaines variables catégorielles

Nous constatons une surestimation de l'ultime pour les catégories suivantes, et pour les années de développement les plus faibles (inférieures à 8 ans) :

- Age du permis supérieur à 10 ans (38% des sinistres clos).
- Caisses EDE13 (14% des sinistres clos) et EDE21 (12% des sinistres clos).
- Age du véhicule supérieur à 10 ans (35% des sinistres clos).

Nous observons également une surestimation des provisions pour les premières années de développement des sinistres survenus à partir de 2008 : en particulier, pour les trois premières années pour les sinistres survenus entre 2013 et 2017, et pour les 7 premières années de développement des sinistres survenus entre 2008 et 2012.

Remarque : Cette analyse descriptive ne prend pas en compte les effets croisés des variables. Pour repérer les variables "responsables" des erreurs il faut faire une étude supplémentaire sur les erreurs, par exemple : un GLM[1] *sur les erreurs.*

1. Modèle linéaire généralisé

Chapitre 5

Implémentation et résultats

Ce chapitre est consacré à la présentation des résultats obtenus pour le calcul de la charge ultime des RBNS vus le 31/12/2019 via les modèles précédemment présentés et par les méthodes Chain-Ladder et Mack. Nous comparons également les résultats des méthodes ligne à ligne entre eux et avec les résultats des méthodes agrégées classiques.

Nous commençons par la présentation des résultats du modèle stochastique à états : les paramètres calibrés seront analysés, et les provisions au titre des RBNS vus à fin 2019 calculées. Le même travail sera ensuite effectué sur les modèles non-paramétriques (apprentissage et optimisation des hyperparamètres). Puis, nous effectuons une analyse classique avec les méthodes Chain-Ladder et Mack. Enfin, nous concluons avec une comparaison des résultats des différents modèles.

5.1 Calibrage du modèle à états

Le calibrage du modèle à états se déroule en deux étapes :

- La première consiste à calibrer les fréquences des évènements de paiement et de clôture.
- La deuxième étape est le calibrage de la distribution des paiements.

Le travail de clustering supervisé effectué précédemment nous ayant permis d'identifier trois groupes de sinistres distincts aux distributions d'ultimes différentes, nous calibrons le modèle stochastique à états sur chacun de ces clusters. Cette démarche permet ainsi de prendre en compte des co-variables - de façon indirecte - dans le calibrage d'un modèle paramétrique.

5.1.1 Calibrage des fréquences des évènements de paiement et de clôture

Afin de calculer le montant de la provision nous devons estimer les fréquences associées aux événements de paiement et de clôture.

Nous rappelons que les évènements sont les suivants :

1. Clôture sans paiement
2. Paiement avec clôture

3. Paiement sans clôture

L'estimation des fréquences repose sur la méthode du maximum de vraisemblance. La maximisation de la fonction de vraisemblance (A.1) conduit à des estimateurs classiques du type :

$$\hat{h_i} = \frac{\text{Nombre total d'évènements de type i}}{\text{Exposition au risque}},$$

avec :

Exposition au risque = Exposition au risque des RBNS + Exposition au risque des sinistres clos

L'exposition au risque des sinistres clos est calculée comme la somme des délais de clôture observés. Dans notre base nous observons 1365 sinistres clos, dont les délais de clôture vont de 0 à 24 ans, avec un pic de clôture entre 5 et 8 ans après la déclaration, selon le cluster.

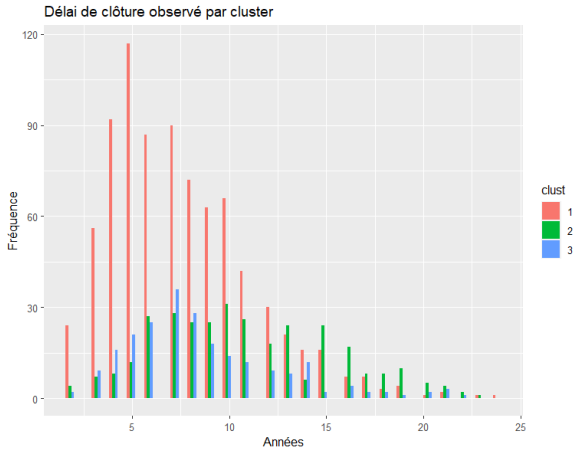


FIGURE 5.1 – Délais de clôture observés

L'exposition au risque des RBNS quant à elle, est calculée comme la somme des développements maximaux observés de ces derniers. Dans notre portefeuille nous comptons 1095 RBNS, la plupart observés entre leur première année de développement et 6 ans de développement. Certains RBNS sont cependant observés pendant des durées bien plus longues, allant jusqu'à 24 années.

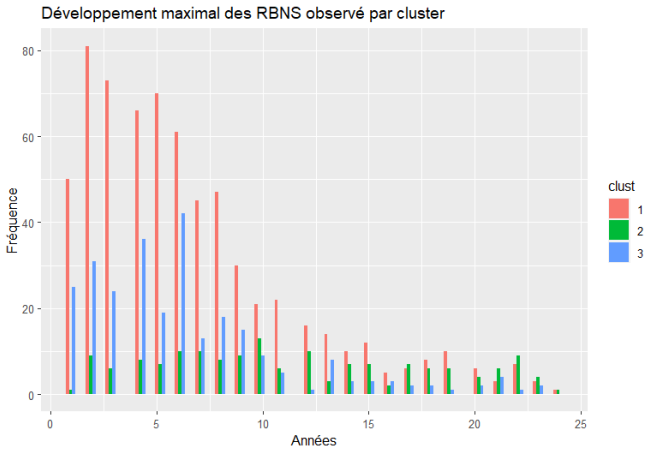
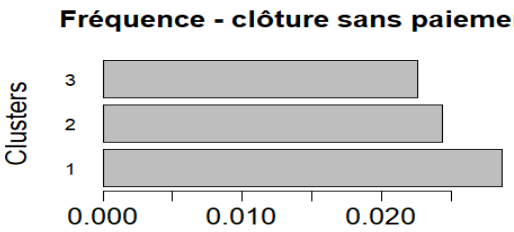


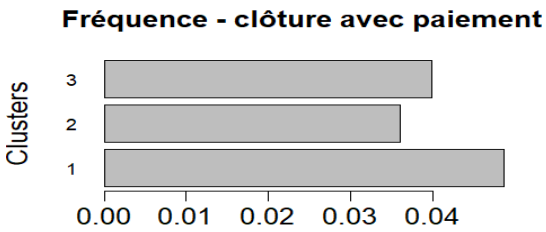
FIGURE 5.2 – Développement maximal observé des RBNS

La distribution des RBNS observés dans le cluster 2 est globalement homogène entre les différentes années de développement. Le cluster 3 contient principalement des RBNS à développement maximal observé entre année de développement 1 et 6.

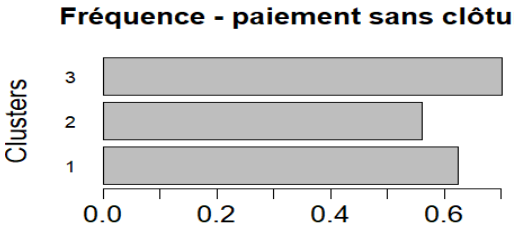
Dans un premier temps, nous supposons que les fréquences associées aux évènements de clôture et de paiement sont constantes et nous calibrons les fréquences h_1, h_2 et h_3 sous cette hypothèse. Les fréquences calibrées sont présentées ci-dessous :



(a) Fréquence de clôture sans paiement (h_1)



(b) Fréquence de clôture avec paiement (h_2)



(c) Fréquence de paiement sans clôture (h_3)

FIGURE 5.3 – Fréquences des évènements calibrées par cluster

À partir des résultats obtenus, nous pouvons facilement constater que les fréquences calibrées diffèrent d'un cluster à un autre.

Les sinistres du cluster 1 comportent plus de paiements et présentent des clôtures plus fréquentes, contrairement aux sinistres du cluster 2 qui présentent des clôtures moins fréquentes, combinées à des paiements moins fréquents. Les sinistres du cluster 3 quant à eux se caractérisent par des paiements fréquents et des clôtures moins fréquentes.

Les résultats de cette estimation seront ensuite utilisés dans le calcul de la provision au titre des RBNS vus au 31/12/2019.

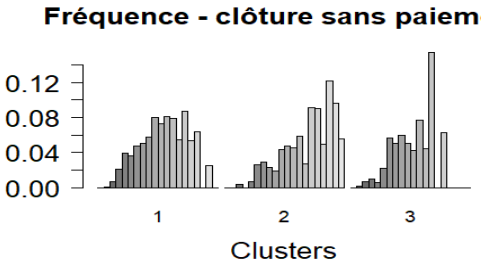
Afin d'affiner notre analyse, nous supposons maintenant que les fréquences des évènements sont constantes par morceaux : elles dépendent ainsi du temps écoulé depuis la déclaration des sinistres. Le pas de temps est annuel.

Les fréquences ainsi calibrées montrent que les développements des sinistres suivent globalement le même schéma, quel que soit le cluster, à quelques différences près.

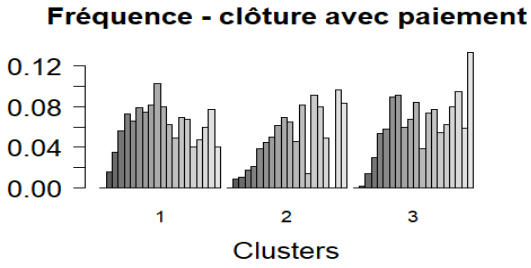
- Les sinistres du cluster 1 commencent à être payés dès la première année, très peu de sinistres sont clôturés au cours des premières années de développement. Dès la

3ème année de développement, nous observons plus de paiements sans clôture, et de paiements avec clôture. Plus les sinistres s'écoulent, plus nous avons de clôtures (avec/sans paiements) et de moins en moins de paiements sans clôture.

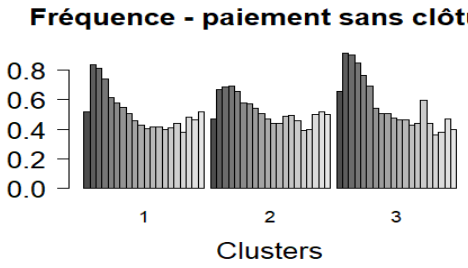
- Ce schéma de développement des sinistres est le même pour les clusters 2 et 3, avec quelques différences : les clôtures s'intensifient au bout d'un laps de temps plus long (9 ans pour le cluster 2, 6 ans pour le cluster 3).
- Par ailleurs, le cluster 2 comporte globalement plus de clôtures et moins de paiements sans clôture.
- Enfin, le cluster 3 comporte un pic de paiements avec clôture sur la dernière année de développement, ainsi qu'un pic de clôture sans paiements à 20 ans.



(a) Fréquence de clôture sans paiement (h_1)



(b) Fréquence de clôture avec paiement (h_2)



(c) Fréquence de paiement sans clôture (h_3)

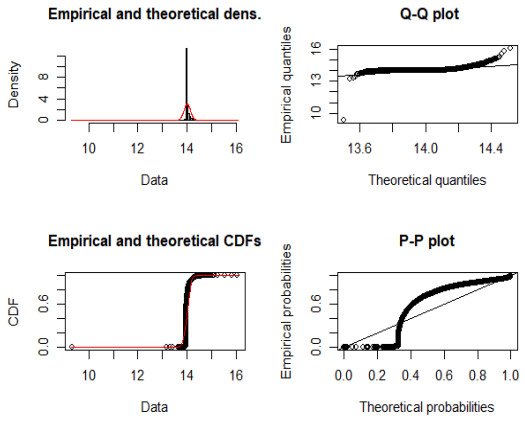
FIGURE 5.4 – Fréquences des évènements, calibrées par cluster, dans le cas où les fréquences sont dépendantes du temps écoulé depuis la déclaration des sinistres

5.1.2 Calibrage des lois de paiement

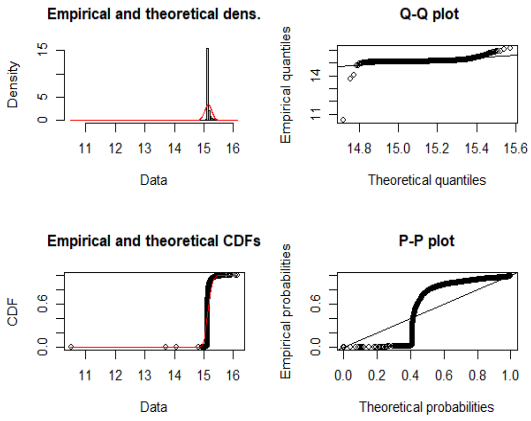
Nous calibrons ensuite la distribution des paiements. Les recours sont pris en compte dans le calibrage des paramètres caractérisant la loi des paiements. Ils sont directement intégrés dans la distribution des paiements et afin de les retraiter, nous décalons la distribution, ensuite nous estimons les paramètres sur la distribution des paiements décalée.

Nous avons testé l'adéquation de trois types de lois à la distribution empirique des paiements : la loi Log-Normale, la loi Log-Bêta, ainsi qu'une loi mélange Log-Normale et Exponentielle.
Afin de choisir la "meilleure" loi, nous examinons le graphique quantile-quantile, le graphique de densité, et nous comparons également les moyennes empiriques et théoriques. En effet, nous souhaitons obtenir une loi qui s'adapte bien aux données, et dont la moyenne s'approche de la moyenne empirique des paiements.

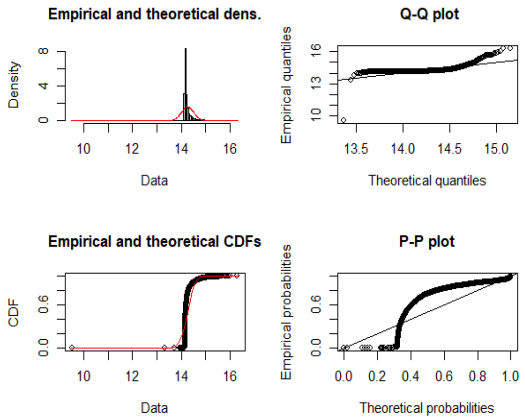
La loi Log-Normale s'adapte bien à la distribution des paiements (voir figure ci-dessous). En revanche, la moyenne de la loi calibrée est sensible au facteur de décalage de la distribution empirique qui a été appliqué. Ainsi, nous testons différentes valeurs du facteur de décalage afin de trouver la valeur qui permet que la moyenne théorique s'adapte à la moyenne empirique : un exemple de cette sensibilité est présenté dans le tableau 5.4.



(a) Cluster 1



(b) Cluster 2



(c) Cluster 3

FIGURE 5.5 – Loi Log-Normale calibrée sur les paiements observés par cluster

Nous avons ensuite décidé de tester une loi mélange, afin de modéliser séparément les paiements extrêmes : en effet, la loi Log-Normale précédemment calibrée ne semble pas capter correctement les queues de distribution. La loi mélange Log-Normale et Exponentielle ainsi calibrée ne s'adapte pourtant pas bien à la distribution des paiements. En particulier, cette loi est sensible à la proportion de paiements considérés comme valeurs extrêmes.

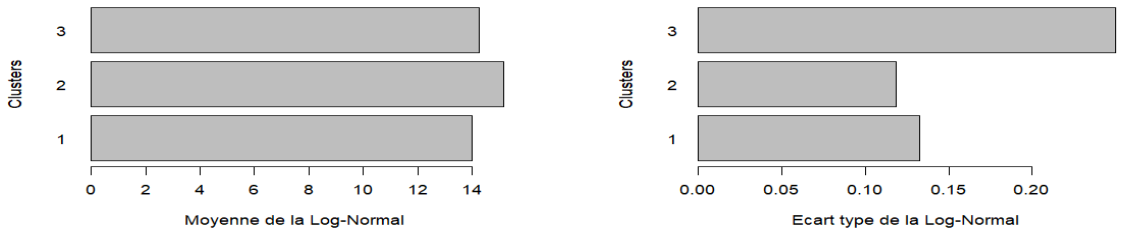
Les deux lois ainsi calibrées ne captent pas le pic des paiements (voir figures 5.5 et B.2), nous avons alors décidé de calibrer une loi Log-Bêta afin de capter ce pic. La loi Log-Bêta semble s'adapter mieux à la distribution des paiements (voir annexe B.1) et les résultats en termes de moyenne et de réserve sont quasiment les mêmes que ceux trouvés avec la loi Log-Normale. En revanche, la moyenne a été estimée par simulation. Ainsi, afin de rester dans le cadre de définition du modèle à états dans l'article BOUMEZOUED et DEVINEAU (2017) nous conservons la loi Log-Normale qui correspond à la distribution

des paiements, décalée à droite de 1.01 * la valeur minimale des recours et pour laquelle nous disposons d'une formule fermée de la moyenne théorique.

L'équation de la loi Log-Normale s'écrit comme suit :

$$P(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right) \tag{5.1}$$

Les valeurs des paramètres de la loi Log-Normale estimés par cluster sont présentés dans la figure suivante :



(a) Moyenne (μ) (b) Ecart type (σ)

FIGURE 5.6 – Paramètres de la loi des paiements estimés

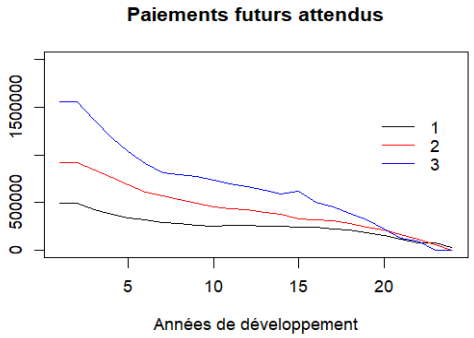
5.1.3 Calcul du montant de la provision par formule fermée et par simulation

Estimation de la provision par formule fermée

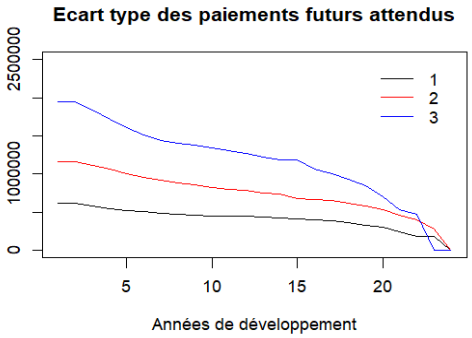
Maintenant que les fréquences et les paramètres associés aux paiements sont calibrés, grâce aux formules fermées (3.3) et (3.4), nous pouvons calculer le paiement attendu $\mu(s)$ et la variance associée $\sigma(s)$.

L'analyse des résultats présentés dans la figure ci-dessous, permet de tirer les conclusions suivantes :

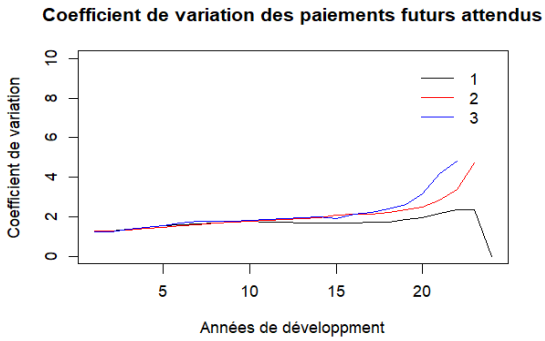
- Les paiements futurs attendus restent stables sur les deux premières années : cela est dû à la croissance de la fréquence de paiement et de la fréquence de clôture.
- Les paiements attendus baissent ensuite, à cause de la combinaison d'une décroissance de la fréquence de paiement et d'une croissance de fréquence de clôture.
- Les paiements futurs espérés, pour les sinistres du cluster 1 et 2, sont presque identiques à partir de la 15ème année de développement.
- L'analyse des coefficients de variation montre que les clusters 2 et 3 se caractérisent par un niveau d'incertitude légèrement élevé sur les dernières années de développement (à partir de la 18-ème année).



(a) Paiements futurs attendus



(b) Ecart type des paiement futurs



(c) Coefficient de variation des paiments futurs

FIGURE 5.7 – Espérance, écart type et coefficient de variation des paiements futurs en fonction des clusters

Le montant de la provision au titre des RBNS vus à fin 2019, que nous obtenons avec cette méthode, s'élève à 1.67 Mds €.

Simulation de la distribution des paiements futurs

Grâce aux paramètres estimés, nous sommes en mesure de déterminer la loi de la provision totale en considérant l'algorithme de simulation défini dans la partie 3.1.3 du chapitre 3.

Les résultats de la simulation sont présentés dans la figure suivante :

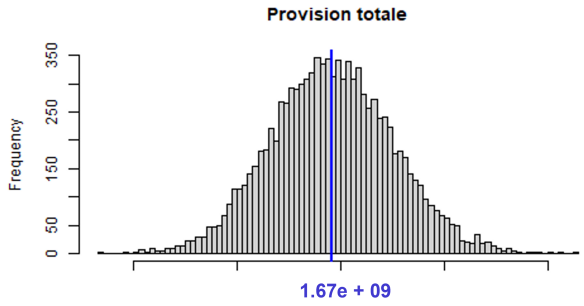


FIGURE 5.8 – Distribution de la provision totale

Comme attendu, le montant de la provision moyenne, représenté en bleu, s'élève à 1.67 Mds €.

Afin de constituer le montant de la charge ultime au titre des RBNS vus à fin 2019, les paiements à date sont ajoutés au montant de la provision estimé. Ainsi, nous retrouvons une charge ultime qui s'élève à 2.3 Mds €.

Dans le tableau ci-dessous nous détaillons la répartition de la provision et de la charge ultime en fonction des clusters :

Cluster	Nombre de RBNS	Provision	Charge ultime
1	667 (61% des RBNS de la base)	575 900 641 (35% de la provision totale)	803 091 611
2	159 (15% des RBNS de la base)	243 880 255 (15% de la provision totale)	351 109 301
3	269 (25% des RBNS de la base)	848 516 135 (50% de la provision totale)	1 148 064 643
Total	1095	1 668 297 031	2 302 265 555

TABLE 5.1 – Résultats du modèle à états par cluster

Les RBNS du cluster 1, soit 61% des RBNS de la base, représentent 35% de la provision globale, ce qui est cohérent avec ce que nous avons observé sur les sinistres clos de ce cluster : il est constitué des sinistres les moins graves, contrairement au cluster 3 qui contient moins de RBNS (25% des RBNS de la base) et qui représente à lui seul la moitié de la provision globale. Là aussi le résultat est adéquat avec le constat que le cluster 3 rassemble des sinistres ayant une sévérité élevée. Le cluster 2 quant à lui contient peu de RBNS (15% de nombre total) et représentent une proportion très faible de provision (15% de provision globale) en comparaison aux deux autres clusters.

5.1.4 Erreur de prédiction du modèle

Afin d'évaluer le modèle nous procédons au calcul de l'erreur quadratique moyenne des prédictions (MSEP).

Soit X, le montant de la réserve et $\hat{X}$, l'estimateur de la réserve, ainsi :

$$MSEP(\hat{X}) = \mathbb{E}[\hat{X} - X] \tag{5.2}$$

L'erreur de prédiction globale associée aux paiements futurs estimés MSEP intègre les deux composantes suivantes :

- **Erreur de process** : aléa pur dû au caractère stochastique des trajectoires futures.
- **Erreur d'estimation** : liée à l'incertitude sur la valeur des paramètres estimés.

Ces deux types d'erreurs peuvent être mesurées par simulations ou par formules fermées, voir BOUMEZOUED et DEVINEAU (2017).

Dans notre étude, nous mesurons les erreurs par simulation :

	Erreur de process	Erreur d'estimation
Simulation	Variance calculée sur les prédictions par l'algorithme de simulation (simulation à facteur fixé)	Simulation à partir de la matrice de variance-covariance des paramètres estimés

TABLE 5.2 – Erreur de process et erreur d'estimation

Les résultats obtenus sont les suivants :

RMSE	Erreur de process	Erreur d'estimation
371 182 900	70 496 366	300 686 534

TABLE 5.3 – Erreurs de prédiction du modèle à états

Nous pouvons constater que l'erreur d'estimation est élevée par rapport à l'erreur de process, elle représente 81% de l'erreur globale.

5.1.5 Etude de la sensibilité des résultats au décalage de la distribution des paiements

Comme nous l'avons évoqué dans la section précédente sur le calibrage de la loi des paiements, la loi Log-Normale est sensible au facteur de décalage (shift) de la distribution des paiements. Nous présentons dans le tableau suivant quelques résultats montrant cette sensibilité.

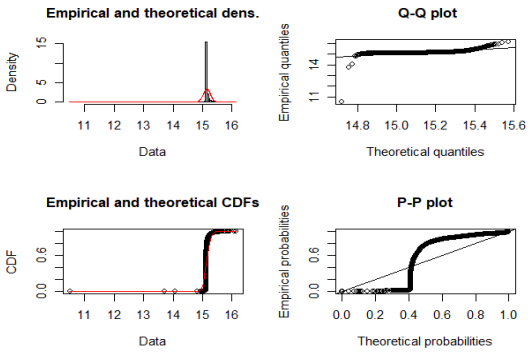
La loi mélange n'est pas très sensible à ce décalage ainsi, nous avons choisi de présenter les résultats avec un décalage de $1.00001*$ *la valeur minimale des recours afin d'avoir une idée sur la différence entre les résultats des deux lois.*

Loi	Shift	Provision (Mds €)	RMSE	Erreur d'estimation	Erreur de process
Log-Normale	1.01	1.67	371 182 900	70 496 366	300 686 534
	1.00001	1.95	510 992 228	89 716 123	421 276 106
Loi mélange	1.00001	1.70	329 842 755	75 232 633	254 610 122

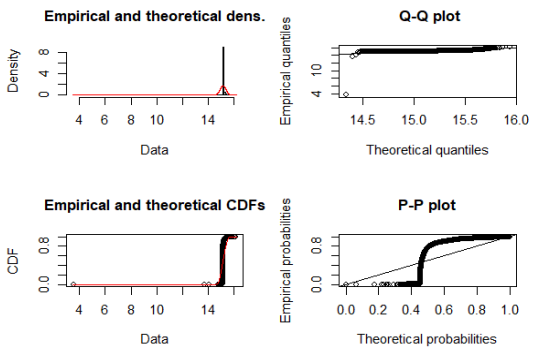
TABLE 5.4 – Sensibilité des résultats

Nous pouvons constater qu'en fonction du décalage considéré, la provision est différente et la performance du modèle également. Le montant de la provision est négativement corrélé au décalage de la distribution des paiements. En effet, un décalage important nous donne une provision plus petite que ce que nous obtenons avec un petit décalage. La performance du modèle quant à elle est meilleure avec un décalage important de la distribution des paiements.

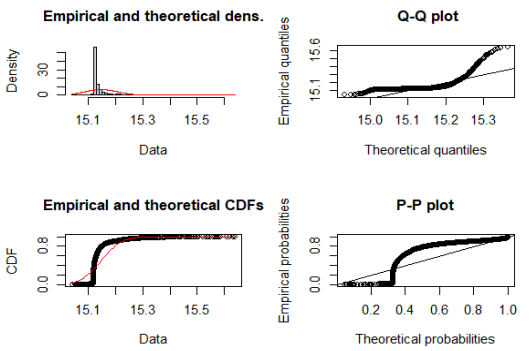
Les résultats des trois calibrages pour le cluster 2 sont présentés dans la figure suivante :



(a) Log-Normale(Shift=1.01)



(b) Log-Normale(Shift=1.00001)



(c) Loi mélange (paiements calibrés par la loi Log-Normale)

FIGURE 5.9 – Calibrage des différentes lois sur les paiements observés (Cluster2)

Ces résultats justifient notre choix de considérer une loi Log-Normale décalée à droite de $1.01*$ la valeur minimale des recours.

5.2 Modèles non-paramétriques

Dans cette section nous considérons deux algorithmes de machine learning (Random Forest et XGBoost) que nous appliquons chacun sur deux ensembles de données. Notre choix initial s'était porté sur l'algorithme Random Forest, pour son nombre limité d'hyperparamètres à optimiser, ainsi que sa relative simplicité d'interprétation. Nous avons ensuite choisi de compléter notre analyse en calibrant également un algorithme de boosting, afin d'étudier la possibilité d'exploiter les avantages de deux algorithmes via un modèle de stacking.

Tout d'abord, nous avons entraîné les modèles sur l'ensemble des sinistres clos uniquement. Nous détaillerons les résultats de calibrage et de validation des modèles. Cependant, cette approche comportant plusieurs biais, nous incluons ensuite les RBNS dans la base de données d'entraînement, en complétant leur trajectoire grâce aux coefficients de développement de Chain-Ladder calibrés sur le triangle correspondant aux données. De la même façon, nous explicitons les résultats de calibrage et de validation des modèles.

Puis, nous vérifions la conformité des résultats des modèles calibrés en effectuant un exercice de backtesting.

Enfin, nous comparons les résultats de prédictions des modèles au titre des RBNS estimés au 31/12/2019 et nous concluons avec une étude de variabilité du modèle qui semble fournir des estimations plus cohérentes.

La mise en application des algorithmes de machine learning a été effectuée sur R.

Partitionnement des données

La base de données initiale a été scindée aléatoirement en deux sous-bases indépendantes, une base d'apprentissage et une base de test, de façon à conserver une proportion identique de sinistres dont le dernier paiement cumulé est supérieur ou égale à 500k € dans les deux bases : nous avons utilisé un partitionnement stratifié.

Les algorithmes sont entraînés sur 80% des sous-dossiers victime, puis testés sur les visions des 20% restants.

Pour bien tenir compte de l'évènement de dépassement de seuil de 500k € de charge qui caractérise la base de données, les modèles sont calibrés sur les visions de sinistres pour lesquelles la date de vision est supérieure ou égale à la date de premier dépassement de seuil.

Dans la suite de l'étude, nous considérons l'ensemble des co-variables, car nous travaillons avec des populations différentes et des années différentes, et il se peut qu'une «nouvelle» variable ressorte.

5.2.1 Entraînement sur les sinistres clos

Nous entraînons tout d'abord les algorithmes sur les sinistres clos uniquement. En effet, la variable réponse à apprendre est l'ultime ; or, cette information est disponible uniquement pour les sinistres clos.
Avec cette approche, plusieurs éléments peuvent biaiser les résultats :

- Un biais de sélection apparait car les sinistres déjà clos à la date de provisionnement peuvent avoir des développements plus courts.
- Les sinistres dont le développement est plus court peuvent avoir une tendance à avoir des montants totaux payés plus faibles.

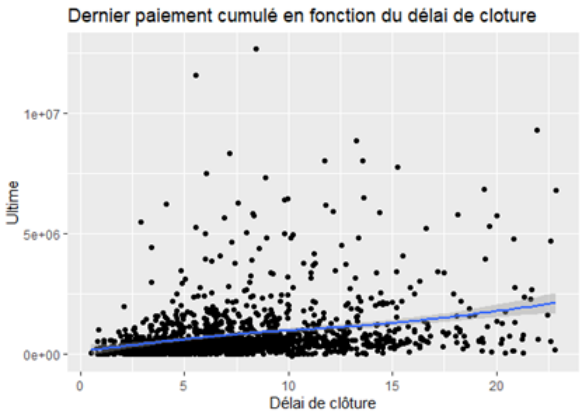


FIGURE 5.10 – Dernier paiement cumulé en fonction du délai de clôture

- De plus, une proportion importante de sinistres est retirée de l'analyse (40%), ce qui entraine une perte d'information.

Dans un premier temps, nous calibrons le modèle Random Forest et présentons les résultats de validation du modèle sur la base test. Ensuite nous passons au calibrage du modèle XGBoost.

Calibrage de l'algorithme Random Forest

Lors du calibrage d'un modèle de machine learning, il convient, afin de maximiser les performances du modèle et de réduire le sur-apprentissage, d'optimiser les hyperparamètres principaux du Random Forest : ntree (nombre d'arbres) et mtry (nombre de variables testées à chaque split).

Pour ce faire, nous définissons un objet ctrl, qui permet de contrôler la manière dont se fait l'entraînement du modèle en effectuant une validation croisée (5 folds). Ensuite, nous définissons une grille de paramètres du modèle Random Forest appelée rf_tune_grid.
La grille est la suivante :

Hyperparamètres	Valeurs testées
ntree	{2, 40, 60, 80, 100}
mtry	{1, 3, 5, 7, 10}

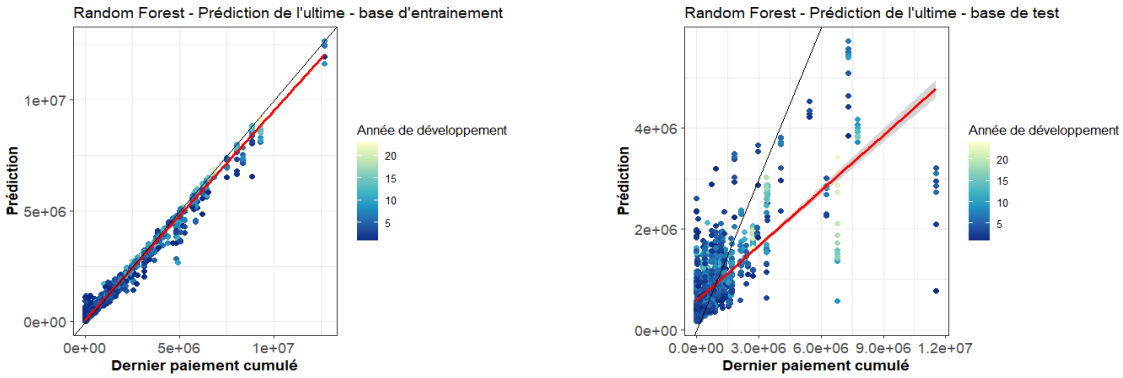
TABLE 5.5 – Grille des hyper-paramètres - Random Forest

Ensuite, en nous basant sur le critère de RMSE, nous récupérons les paramètres optimaux associés au meilleur modèle.

- Nous constatons que les résultats sont stables à partir de ntree=60, nous retenons donc ce nombre d'arbres.
- Pour mtry, les résultats sont stables à partir de mtry=5, nous retenons donc ce paramètre.

Validation du modèle avec la base test

Maintenant que le modèle est calibré sur la base d'entraînement, nous pouvons réaliser des prédictions sur la base test puis analyser les résidus et les erreurs de prédiction avec ce modèle. La figure suivante présente les prédictions de l'ultime en fonction des ultimes observés sur la base d'entraînement à gauche et la base test à droite.



(a) Prédiction de l'ultime sur la base train (b) Prédiction de l'ultime sur la base test

FIGURE 5.11 – Prédictions de l'ultime par le modèle global entraîné sur les clos-Random Forest

Le modèle entraîné est très performant sur la base d'apprentissage (95.11% de variance expliquée). Cependant, malgré l'optimisation des hyperparamètres, nous constatons que le modèle surapprend la base d'apprentissage : l'erreur, très faible en apprentissage, est relativement élevée sur la base de test.

%Variance expliquée	RMSE base d'apprentissage	RMSE base de test	Erreur globale sur la base test (prédiction-ultime réel)
95.11%	139 710	1 034 101	-89 681 503

TABLE 5.6 – Performances du modèle global entraîné sur les clos- Random Forest

Ensuite, nous étudions les erreurs de prédiction du modèle : pour ce faire, nous distinguons les erreurs par classe d'ultime. Deux classes d'ultimes ont été définies : « Ultime inférieur à 500k€» et « Ultime supérieur ou égal à 500k€».

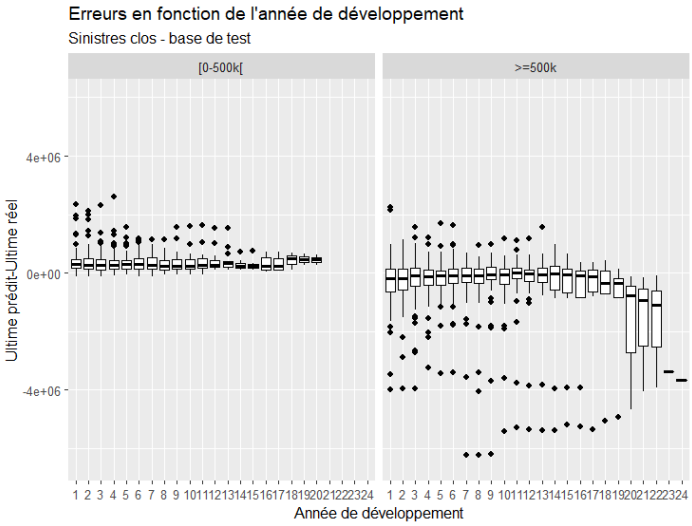


FIGURE 5.12 – Erreurs de prédiction du modèle global entraîné sur les clos sur la base test en fonction des années de développement-Random Forest

Nous constatons que les sinistres pour lesquels l'ultime est inférieur à 500k€, sont

surestimés par le modèle, et que, au contraire, les sinistres ayant un ultime supérieur à 500k€ sont sous-estimés.

Afin d'affiner un peu la prédiction, nous décidons alors de construire un modèle en deux parties, un modèle apprenant sur toutes les visions de sinistres qui auront un ultime inférieur à 500k €, et un autre apprenant toutes les visions de sinistres qui auront un ultime supérieur à ce seuil.
Avant de passer à l'entraînement du modèle en deux parties, nous souhaitons savoir si le bon modèle est appliqué à la bonne proportion de sinistres. Nous voulons savoir si le modèle apprenant sur les visions de sinistres ayant un ultime inférieur à 500k € est appliqué aux sinistres ayant une charge inférieur à 500k € et celui apprenant sur les visions de sinistres ayant un ultime supérieur à ce seuil est appliqué aux sinistres ayant une charge supérieure à ce seuil.
La figure ci-dessous présente la proportion de visions de sinistres pour lesquelles la classe de charge ne correspond pas à la classe d'ultime : pour ces observations nous n'appliquons pas le bon modèle, c'est-à-dire que le modèle en deux parties risque de se tromper sur ces visions.

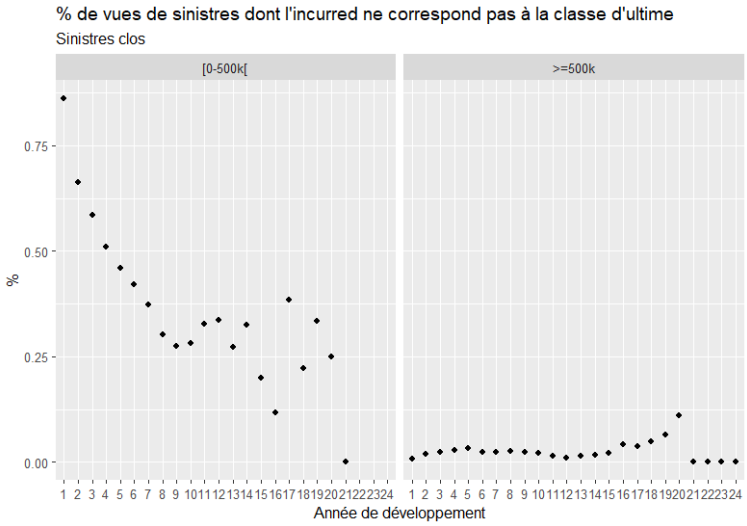


FIGURE 5.13 – Proportion de sinistres auxquels le bon modèle n'est pas appliqué

Nous pouvons constater que pour la classe d'ultime inférieur à 500k€, sur presque toutes les visions de sinistre, la charge ne correspond pas à cette classe d'ultime. Près de 80% des sinistres dont l'ultime sera inférieur à 500k€, comportent une charge supérieure à 500k € sur leur première année de développement. En fait, la provision d/d est fortement surestimée sur les première années de développement, cette surestimation a tendance à baisser jusqu'à la 9-ème année de développement, puis augmente et baisse selon l'année de développement.

Ainsi, pour la classe d'ultime inférieur à 500k€, comme la provision d/d est globalement surestimée, les modèles auront tendance à surestimer l'ultime car il sera prédit par le modèle entraîné sur les sinistres ayant un ultime supérieur ou égal à 500k€.

En ce qui concerne la classe d'ultime supérieur ou égal à 500k €, la proportion de sinistres dont la charge ne correspond pas à la classe d'ultime est relativement basse.

Nous observons une hausse de la proportion de visions de sinistres qui passeront dans le mauvais modèle, entre 16 et 20 ans de développement. Ainsi, en termes d'erreurs, l'ultime sera sous-estimé, sur ces visions de sinistres, car il sera prédit par le modèle entraîné sur les sinistres ayant un ultime inférieur à 500k €.

Les résultats obtenus avec le modèle en deux parties, en termes de RMSE et d'erreur globale sur la base de test, sont présentés dans le tableau suivant :

%Variance expliquée	RMSE base d'apprentissage	RMSE base de test	Erreur globale sur la base test (prédiction-ultime réel)
RF-sup-500k€ : 96% RF-inf-500k€ : 87%	486 093	1 061 494	324 836 336

TABLE 5.7 – Performances du modèle en deux parties entraîné sur les clos - Random Forest

Les résultats des deux modèles sont assez similaires, le modèle en deux parties ne semble pas améliorer la RMSE par rapport au modèle global. En revanche, nous pouvons constater que le modèle en deux parties surestime l'ultime réel, par construction, tandis que le modèle global le sous-estime.

Algorithme XGBoost

Transformation des variables qualitatives en variables binaires

Contrairement à l'algorithme Random Forest, l'application de l'algorithme XGBoost nécessite la transformation des variables qualitatives en variables binaires. Autrement dit, il faut décomposer les modalités associées à une variables unique en variables individuelles, ensuite chacune des nouvelles variables vérifie si la modalité était présente pour une observation et lui associée 1 en cas de présence et 0 sinon. Pour ce faire, nous utilisons la fonction model.matrix de R.

Calibrage de l'algorithme

Tout comme pour le Random Forest, les hyperparamètres du modèle ont été optimisés en définissant un objet trainControl, qui permet de contrôler la manière dont se fait l'entraînement du modèle en effectuant une validation croisée (5 folds).
Nous définissons ensuite une grille de paramètres du modèle XGBoost.

Hyperparamètres	Valeurs testées
n_rounds	[10 :1000], pas = 10
max_depth	{3, 6, 8}
eta	{0.01, 0.05, 0.1, 0.3}
gamma	{0, 1000, 10000, 100000}

TABLE 5.8 – Grille des hypersparamètres - XGBoost

Avec :

- nrounds : nombre d'itérations de boosting à effectuer

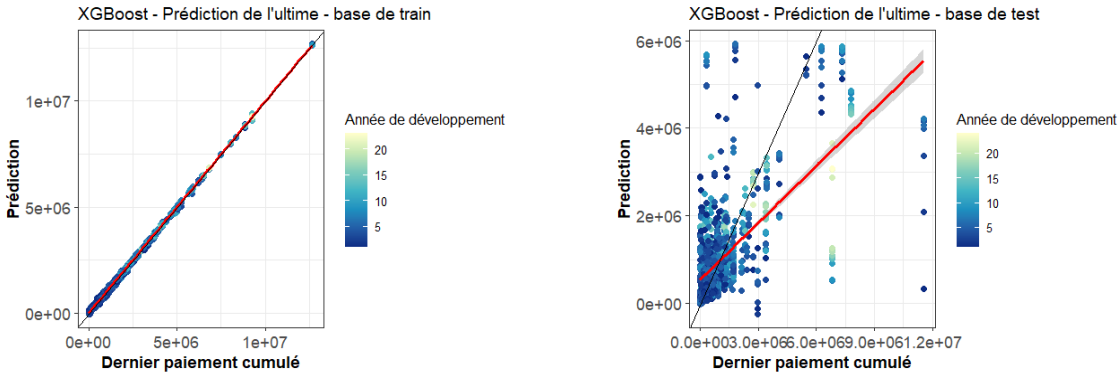
- max_depth : profondeur d'arbre maximale
- eta (ou learning rate) : ce paramètre contrôle la vitesse à laquelle nous convergeons lors de la descente du gradient fonctionnel
- gamma : diminution minimale de la valeur de la fonction objectif pour prendre la décision de partitionner une feuille

Ensuite, en utilisant la fonction train() de R, nous pouvons récupérer le modèle optimal et les paramètres associés. Ainsi, pour l'entrainement du modèle, nous retenons les paramètres suivants :

- nrounds = 1000
- max_depth = 6
- eta = 0.1
- gamma = 100000

Validation du modèle avec la base test

Nous passons maintenant à la phase de validation du modèle avec le jeu d'hyperparamètres retenu. La figure suivante présente les prédictions de l'ultime en fonction des ultimes observés sur la base d'entraînement à gauche et la base test à droite.



(a) Prédiction de l'ultime sur la base train (b) Prédiction de l'ultime sur la base test

FIGURE 5.14 – Prédictions de l'ultime par le modèle global entraîné sur les clos -XGBoost

Nous observons également le phénomène de surapprentissage avec l'algorithme XGBoost.

RMSE base d'apprentissage	RMSE base de test	Erreur globale sur la base test (prédiction-ultime réel)
35 914	1 112 581	-27 893 472

TABLE 5.9 – Performances du modèle globale entraîné sur les sinistres clos - XGBoost

Les résultats de ce modèle sont légèrement meilleurs en termes d'erreur globale par rapport au modèle global Random Forest, en revanche en terme de RMSE le modèle global Random Forest est meilleur. Ainsi, à ce stade, nous ne pouvons pas choisir entre les deux.

Nous analysons maintenant les erreurs de prédiction par classe d'ultime.

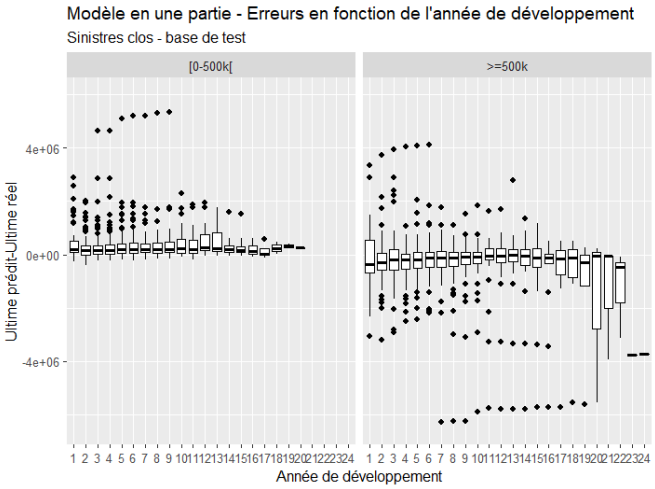


FIGURE 5.15 – Erreurs de prédiction du modèle global entraîné sur les clos sur la base test en fonction des années de développement - XGBoost

Nous constatons que les sinistres pour lesquels l'ultime est inférieur à 500k€ sont surestimés par le modèle, et que, au contraire, les sinistres ayant un ultime supérieur à 500k€ sont sous-estimés.

Comme pour le modèle Random Forest, nous construisons un modèle en deux parties, un modèle apprenant sur les sinistres ayant un ultime inférieur à 500k €, et un autre apprenant sur ceux ayant un ultime inférieur à ce seuil.

RMSE base d'apprentissage	RMSE base de test	Erreur globale sur la base test (prédiction-ultime réel)
614 051	1 216 108	533 427 420

TABLE 5.10 – Performances du modèle en deux parties entraîné sur les clos-XGBoost

Comme pour les résultats du modèle Random Forest, le modèle en deux parties ne semble pas améliorer la RMSE par rapport au modèle global. En revanche, nous pouvons constater que le modèle en deux parties surestime l'ultime réel, toujours par construction (cf. analyse de la proportion de sinistres pour lesquels "nous nous trompons" de modèle à appliquer"), tandis que le modèle global le sous-estime.

Nous avons l'impression que les modèles entraînés présentent du surapprentissage. Pourtant, les hyperparamètres ont bien été optimisés. Ce surapprentissage peut s'expliquer par le fait que la base contient peu d'observations pour construire du machine learning et qu'il est normal d'avoir une baisse de performance entre la base d'entraînement et la base de test.

5.2.2 Entraînement sur les sinistres clos et RBNS complétés

Une deuxième approche consiste à utiliser l'ensemble des observations disponibles : les sinistres clos, ainsi que les RBNS, pour lesquels nous ne disposons pas de l'ultime, mais que nous nous proposons de compléter par des "pseudo-ultimes".

Ainsi, avant d'entraîner les algorithmes de machine learning sur l'ensemble des sinistres (clos et RBNS), nous développons les sinistres encore ouverts en utilisant les facteurs de développement de la méthode classique Mack Chain-Ladder.

Pour ce faire, nous considérons le triangle « Année de survenance de sinistre » x « Année de dépassement du seuil de 500k €», en vue d'intégrer la dynamique de dépassement de seuil de 500k €, et de coller au mieux à la dynamique de développement des sinistres observés dans la base des graves.

Les étapes de calcul des « pseudo-ultimes » sont les suivantes :

1. Construction du triangle de paiements nets de recours correspondant à une année N de provisionnement.
2. Calcul des facteurs de développement.
3. Application des facteurs de développement calculés sur un triangle agrégé, de façon individuelle sur toutes les visions des sinistres non-clos avant le 31/12/N : nous calculons ainsi les « pseudo-ultimes ».

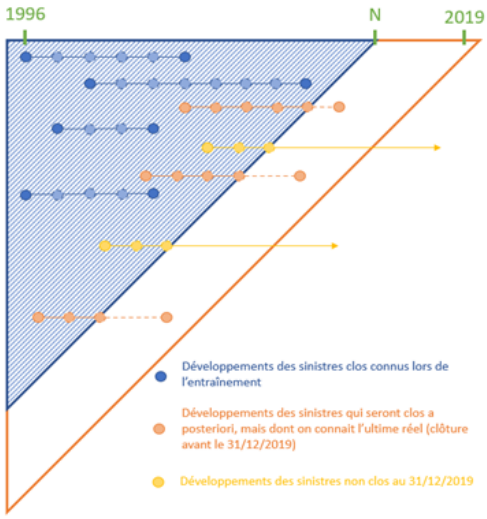


FIGURE 5.16 – Visions utilisées pour l'exercice de comparaison des pseudo-ultime

Comparaison des ultimes réels et des pseudo-ultimes pour les années de back-testing

Avant d'entrainer les modèles de machine learning sur la base des sinistres clos et RBNS complétés par des « pseudo-ultimes », nous nous intéressons aux écarts entre les ultimes réels et les « pseudo-ultimes » calculés par Mack Chain-Ladder.

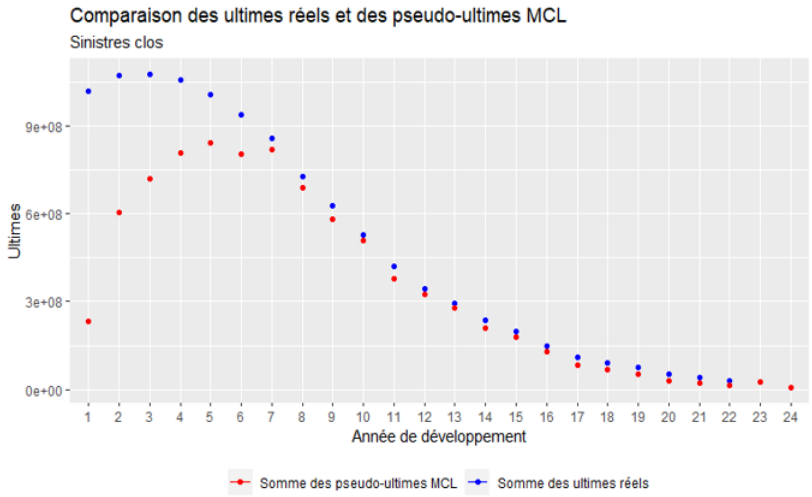


FIGURE 5.17 – Comparaison des ultimes réels et pseudo-ultimes par année de développement

Les « pseudo-ultimes » calculés sous-estiment l'ultime réel sur les années de développement 1 à 8, mais ces écarts diminuent sur les développements plus longs.
Ces résultats ne sont pas surprenants : en effet, les sinistres de la branche étudiée ont des développements longs, les paiements observés les premières années ne reflètent pas l'ultime réel, et cela se ressent donc dans les facteurs de développement calibrés.

Nous avons également effectué cette étude en calant les facteurs de développement non pas sur le triangle de paiements nets de recours, mais sur le triangle de charges. Or, les provisions dossier/dossier attribuées aux sinistres lors de leurs premières années de développement étant bien supérieures aux ultimes observés, nous observons les résultats inverses : les pseudo-ultimes surestiment l'ultime réel les premières années de développement, puis les écarts diminuent sur les développements plus longs.

Nous comparons aussi, pour chaque année de backtesting, la valeur des « pseudo-ultimes » pour les RBNS qui sont ensuite clos avant le 31/12/2019, avec leur valeur réelle.

Année N	Ultime réel	Pseudo-ultime	Erreur
2010	627 418 259	377 905 349	-40%
2011	681 314 138	421 598 084	-38%
2012	688 372 794	449 711 732	-35%
2013	662 947 427	456 239 324	-31%
2014	617 699 408	463 556 268	-25%
2015	553 574 836	433 298 149	-22%
2016	422 190 776	392 421 999	-7%
2017	276 170 582	253 374 296	-8%
2018	136 971 059	160 268 444	17%

TABLE 5.11 – Données sur les RBNS vus à fin N, clos entre le 31/12/N et le 31/12/2019

Pour analyser les résultats de cet exercice de backtesting, nous devons prendre en compte la composition des bases de backtesting observées à fin N. En effet, pour les années N les plus anciennes, les sinistres sont encore "jeunes", nous les observons au début de

leur trajectoire de paiements, et donc le pseudo-ultime que nous allons leur attribuer sera, selon l'étude précédente effectuée sur l'écart entre les pseudo-ultime et l'ultime réel, plus faible que l'ultime réel, d'où l'erreur fortement négative.
Pour les années les plus récentes, c'est le contraire qui se produit. Par ailleurs, pour les années de backtesting les plus récentes, nous avons peu de RBNS qui sont clos avant fin 2019, ce qui rend l'interprétation plus délicate, car l'erreur concerne assez peu de sinistres.

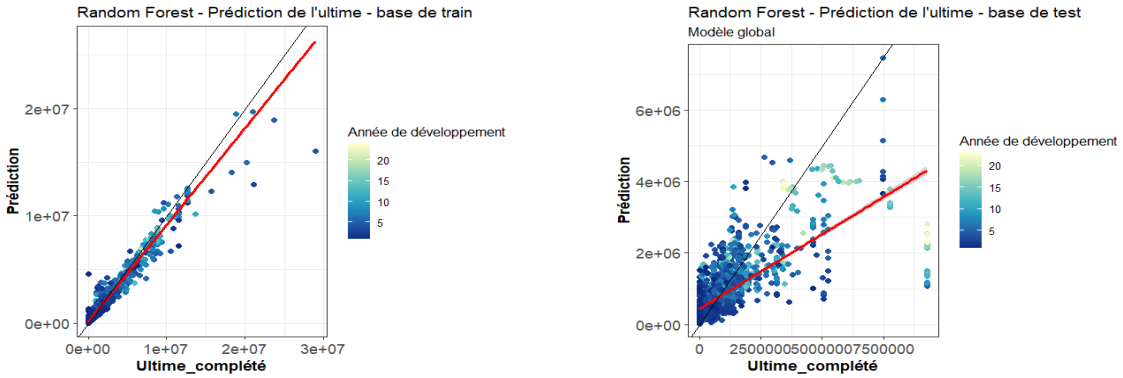
Calibrage du modèle Random Forest

Pour cette approche, la variable à prédire est l'ultime réel pour les sinistres clos à date, et le pseudo-ultime calculé par Mack Chain-Ladder individuel pour les autres sinistres.

Les paramètres optimaux retenu pour l'entrainement du modèle Random Forest sont : ntree = 60 et mtry = 5.

Validation du modèle avec la base test

À partir des prédictions des « pseudo-ultimes » présentées dans la figure ci-dessous et le tableau 5.12, nous pouvons facilement constater que l'algorithme Random Forest, s'il est très bon sur la base d'apprentissage, sous-estime largement l'ultime réel sur la base de test. Nous pouvons, ici aussi, penser à un phénomène de sur-apprentissage, malgré l'optimisation des hyperparamètres de l'algorithme.



(a) Prédiction de l'ultime sur la base train (b) Prédiction de l'ultime sur la base test

FIGURE 5.18 – Prédictions de l'ultime par le modèle global entraîné sur les clos et RBNS- Random Forest

%Variance expliquée	RMSE base d'apprentissage	RMSE base de test	Erreur globale sur la base test (prédiction-ultime complété)
89.23%	257 869	861 337	-77 166 899

TABLE 5.12 – Performances du modèle global entraîné sur les clos et RBNS- Random Forest

Comme pour le modèle entrainé sur les sinistres clos uniquement, nous distinguons les erreurs selon la classe d'ultime. Le constat est le même : le modèle surestime les sinistres inférieurs à 500k €, et sous-estime les sinistres supérieurs à 500k €. La figure suivante présente les résultats de cette étude.

A noter que l'erreur globale étudiée sur les sinistres clos uniquement est principalement portée par l'erreur due aux pseudo-ultimes, et non pas à l'erreur de prédiction du modèle.

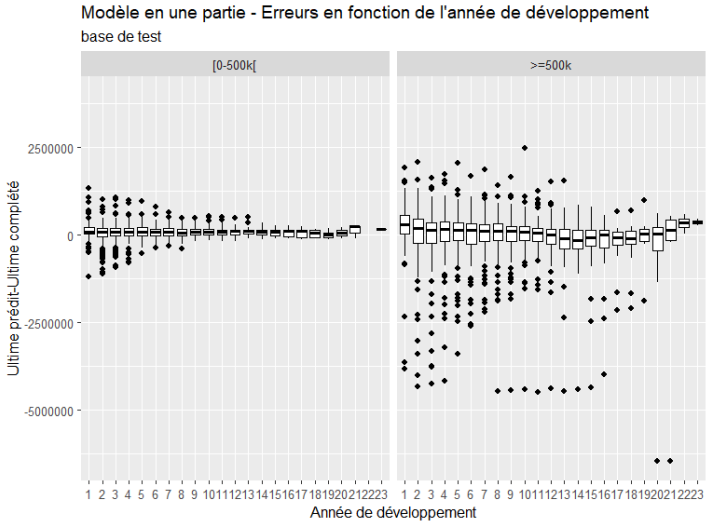


FIGURE 5.19 – Erreurs de prédiction du modèle global entraîné sur les clos et RBNS sur la base test en fonction des années de développement - Random Forest

Nous construisons alors un modèle différencié en deux parties, un modèle apprenant sur les sinistres ayant un ultime inférieur à 500k €, et un autre apprenant sur ceux ayant un ultime inférieur à ce seuil.

%Variance expliquée	RMSE base d'apprentissage	RMSE base de test	Erreur globale sur la base test (prédiction-ultime réel)
RF-sup-500k€ : 89% RF-inf-500k€ : 80%	583 421	981 487	673 315 314

TABLE 5.13 – Performances du modèle en deux parties entraîné sur les clos et RBNS - Random Forest

Le constat est le même : le modèle en deux parties ne semble pas améliorer la RMSE par rapport au modèle global. En revanche, nous pouvons constater que le modèle en deux parties surestime l'ultime réel, par construction, tandis que le modèle global le sous-estime.

Calibrage du modèle XGBoost

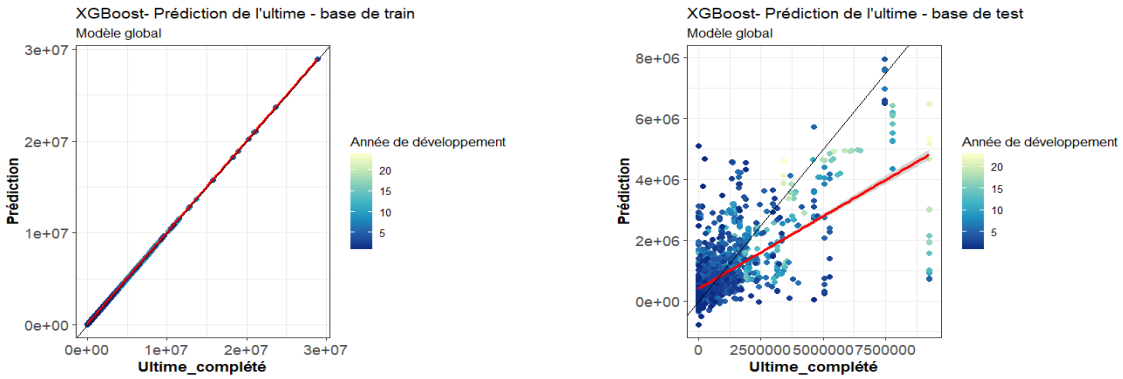
De la même façon que l'algorithme précédent, un algorithme XGBoot est entrainé sur la base des sinistres clos et RBNS complétés.

Pour l'entrainement du modèle, nous retenons les paramètres suivant :

- nrounds = 10000
- max_depth = 6
- eta = 0.1
- gamma = 0

Validation du modèle avec la base test

Comme nous pouvons le constater sur la figure suivante, l'algorithme XGBoost sous-estime les prédictions des « pseudo-ultimes » sur la base test.



(a) Prédiction de l'ultime sur la base train (b) Prédiction de l'ultime sur la base test

FIGURE 5.20 – Prédictions de l'ultime par le modèle entraîné sur les clos et RBNS-XGBoost

RMSE base d'apprentissage	RMSE base de test	Erreur globale sur la base test (prédiction-ultime complété)
29 605	1 343 896	-102 799 459

TABLE 5.14 – Performances du modèle global entraîné sur les clos et RBNS - XGBoost

La distinction des erreurs selon la classe d'ultime nous donne ce qui suit :

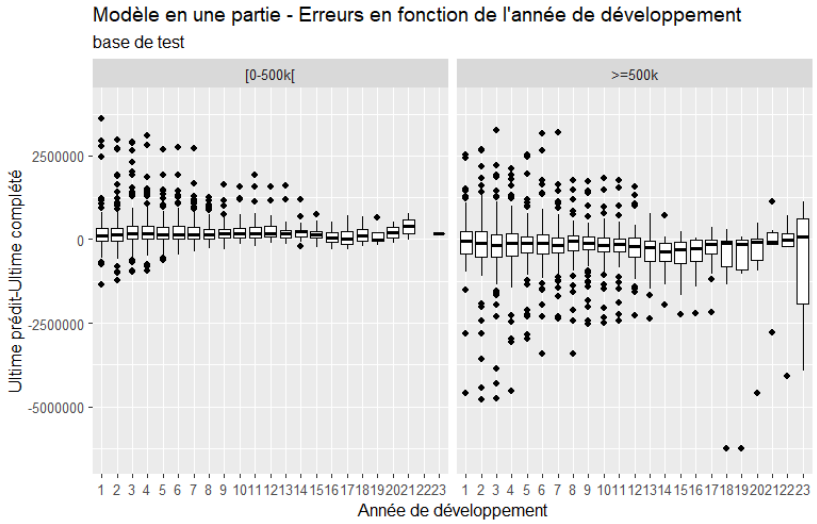


FIGURE 5.21 – Erreurs de prédiction du modèle entraîné sur les clos et RBNS sur la base test en fonction des années de développement - XGBoost

Le modèle surestime légèrement les sinistres inférieurs à 500k€, et sous-estime les sinistres supérieurs à 500k€.
Comme nous pouvons le constater sur les tables 5.12 et 5.14, les performances du modèle

global Random Forest semblent meilleures en termes de RMSE et d'erreur globale.
Les résultats du modèles en deux parties sont les suivants :

RMSE base d'apprentissage	RMSE base de test	Erreur globale sur la base test (prédiction-ultime complété)
568 704	1 437 219	475 052 440

TABLE 5.15 – Performances du modèle en deux parties entraîné sur les clos et RBNS-XGBoost

Le modèle en deux parties surestime les « pseudo-ultimes », contrairement au modèle global qui les sous-estime, et il n'améliore pas la RMSE.

5.2.3 Comparaison des prédictions des modèles sur les sinistres clos – exercice de backtesting

Le calibrage de l'ensemble des modèles précédents - 8 en tout : modèles globaux ou différenciés en deux classes d'ultime, modèles sur clos uniquement ou sur clos et RBNS complétés par pseudo-ultimes, et choix de l'algorithme Random Forest ou XGBoost - et les résultats obtenus sur base de test, nous donnent une idée de la performance de tous ces modèles.

En général, lors de la construction d'un modèle de machine learning, nous divisons la base disponible en trois sous-base : apprentissage, test, et backtest. La base de backtesting consiste d'ordinaire en la dernière année de la base disponible, sur laquelle nous appliquons le modèle entrainé.
Une première baisse de performance est attendue entre apprentissage et test - nous avons vu que c'était effectivement le cas - puis une seconde, moindre, entre test et backtest.
Au vu de la taille restreinte de notre base de données, nous avons cherché un moyen d'effectuer un exercice de backtesting approprié, pour juger de la performance des modèles calibrés dans une situation "réelle".

Ainsi, pour une année N de backtesting (N allant de 2010 à 2018), nous entraînons les algorithmes sur l'ensemble des observations correspondant aux dates de vision des sinistres avant le 31/12/N. Nous appliquons ensuite l'algorithme ainsi entraîné sur les RBNS vus à la fin de l'année N : ce sont les sinistres déclarés avant le 31/12/N et qui ont également dépassé le seuil de 500k € avant cette date, et clôturés ou non après cette date.
Parmi les RBNS de la base de backtesting N il existe des sinistres qui ne sont toujours pas clôturés en 2020. En revanche, certains de ces RBNS sont clos entre le 31/12/N et 2020, ainsi nous disposons de l'information de leurs ultimes. Grâce à cet exercice de backtesting nous pouvons donc comparer les prédictions des modèles avec l'ultime réel sur les années N de backtesting.

Dans la suite de cette étude, nous considérons les appellations suivantes pour les modèles :

- Modèle A - global : modèle global entrainé sur les sinistres clos uniquement.
- Modèle A - différencié : modèle en deux parties entrainé sur les sinistres clos uniquement.

- Modèle B - global : modèle global entrainé sur les sinistres clos et RBNS.
- Modèle B - différencié : modèle en deux parties entrainé sur les sinistres clos et RBNS.

Le tableau suivant récapitule les résultats de prédiction de chacun des modèles ainsi que les informations nécessaires à la comparaison des résultats.

	Erreurs de prédiction entre l'ultime réel et les différentes prédictions								
		Modèles Random Forest				Modèles XGBoost			
Année N	Gestionnaire	Modèle A global	Modèle A différencié	Modèle B global	Modèle B différencié	Modèle A global	Modèle A différencié	Modèle B global	Modèle B différencié
2010	25%	-9%	3%	2%	16%	-13%	15%	-40%	-4%
2011	34%	4%	18%	4%	19%	-5%	19%	-38%	3%
2012	41%	6%	23%	15%	25%	-9%	27%	-35%	11%
2013	38%	10%	19%	14%	27%	-2%	25%	-31%	10%
2014	44%	11%	22%	13%	25%	1%	28%	-25%	16%
2015	29%	-2%	7%	9%	22%	-14%	15%	-22%	13%
2016	32%	-3%	6%	29%	32%	-13%	22%	-7%	24%
2017	28%	17%	19%	31%	34%	1%	34%	-8%	26%
2018	21%	14%	18%	37%	30%	-6%	19%	17%	44%

TABLE 5.16 – Erreurs de prédiction entre l'ultime réel et les différentes prédictions

Cet exercice de backtesting doit s'interpréter de façon prudente. En effet, pour les années de backtesting N les plus anciennes, la plupart des sinistres vus à fin N sont des RBNS qui seront clos avant fin 2019 : la base d'apprentissage est restreinte (peu d'historique de sinistres), mais nous pouvons comparer la valeur prédite sur un grand nombre de sinistres. Par contre, pour les années de backtesting les plus récentes, si la base d'apprentissage est plus conséquente, nous avons finalement assez peu de RBNS qui seront clos avant fin 2019, donc peu de sinistres pour lesquels comparer le prédit au réel.

Nous constatons tout d'abord que la provision globale estimée par le gestionnaire sinistres surestime l'ultime total payé, d'au moins 20%, sur toutes les années de backtesting.

Les modèles A-globaux (Random Forest et XGBoost) sous-estiment l'ultime : ceci était attendu, compte tenu du fait des biais évoqués (sinistres clos observés aux développement plus courts, notamment).
Les modèles A-différenciés, quant à eux, surestiment de beaucoup l'ultime réel, par construction : pour rappel, un grand nombre de visions de sinistres qui auront un ultime inférieur à 500k€ passent dans le modèle entraîné sur les sinistres ayant un ultime supérieur ou égal à 500k€, ce qui conduit à une surestimation de l'ultime pour ces sinistres.

Les modèles B ont un comportement un peu différent : compte tenu du fait que les pseudo-ultimes sous-estiment l'ultime réel sur les 8 premières années de développement, nous nous attendons à une sous-estimation par rapport aux résultats des modèles A. C'est effectivement ce que nous observons sur le Modèle B -Global - XGBoost (forte sous-estimation par rapport au Modèle 1 - Global - XGBoost), ainsi que sur le Modèle B -

différencié - XGBoost (la surestimation de l'ultime due à la construction en deux sous-modèles qui ne s'appliquent pas forcément aux "bons" sinistres est contrebalancée par la sous-estimation par pseudo-ultime).
Cette dernière explication peut par ailleurs également s'appliquer au Modèle B - différencié - RF.
Reste enfin le Modèle B - global - Random Forest, qui ne semble pas être affecté par la sous-estimation des pseudo-ultimes : il surestime l'ultime de façon raisonnable.

Pour la suite de l'étude, lorsque nous souhaiterons affiner les prédictions (notamment au Chapitre 6), c'est ce modèle que nous retiendrons. En effet, les modèles A sont biaisés de par la nature des données auxquelles ils s'appliquent ; les modèles différenciés sont quant à eux biaisés par construction. Le Modèle B - global - Random Forest semble être le meilleur modèle non-paramétrique calibré sur les données.

Etude de corrélation entre les erreurs des modèles

Nous avons implémenté deux types d'algorithmes différents dans le but de comparer les erreurs commises par les modèles, et, si elles s'avèrent complémentaires, construire un modèle de stacking pour tirer avantage des forces de chacun des deux algorithmes. Pour cela, il faut que les modèles XGBoost et Random Forest se trompent sur des observations différentes (auquel cas le stacking prendrait l'une ou l'autre des prédictions selon les co-variables), ou sur les mêmes observations, mais dans un sens différent (dans ce cas le stacking serait une combinaison des deux prédictions).

Une étude sur les erreurs de prédiction entre l'ultime réel et les prédictions des modèles a été réalisée dans ce but. L'étude consiste à comparer les erreurs de prédiction du modèle Random Forest et celles du modèle XGBoost.

Pour ce faire, nous calculons le coefficient de corrélation de Pearson sur les erreurs de prédiction des modèles.

Nous rappelons que le coefficient de corrélation de Pearson entre deux variables aléatoires X et Y est le suivant :

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

Avec :

- cov(X,Y) : la covariance entre X et Y.
- σ_X : écart-type de la variable X.
- σ_Y : écart-type de la variable Y.

Le tableau suivant résume les coefficients de corrélation entre les différents modèles entrainés sur les sinistres clos uniquement :

Modèle	A-global-RF	A-différencié-RF
A-global-XGBoost	0.93	0.90
A-différencié-XGBoost	0.92	0.89

TABLE 5.17 – Corrélation entre les erreurs des modèles A

Le tableau suivant résume les coefficients de corrélation entre les différents modèles entrainés sur les sinistres clos et RBNS :

Modèle	B-global-RF	B-différencié-RF
B-global-XGBoost	0.89	0.81
B-différencié-XGBoost	0.90	0.87

TABLE 5.18 – Corrélation entre les erreurs des modèles B

À partir des résultats ci-dessus, nous constatons que les modèles se trompent dans le même sens et ceci quelque soit l'approche considérée. Ainsi, l'application d'un modèle de stacking ne semble pas pertinent de prime abord.

Nous pourrions explorer la piste du stacking en construisant un modèle apprenant une combinaison linéaire des prédictions Random Forest et XGBoost, dont les coefficients varieraient selon les covariables. Ceci ne sera pas effectué dans ce mémoire mais pourra faire l'objet de travaux futurs.

5.2.4 Comparaison des ultimes des RBNS estimés au 31/12/2019

Dans cette section, nous nous intéressons aux prédictions des ultimes des RBNS des différents modèles au 31/12/2019.

Le tableau ci-dessous résume les prédictions des différents modèles implémentés avec l'algorithme Random Forest :

	Prédictions des différents modèles			
	Modèles Random Forest			
Année N	Modèle A global	Modèle A différencié	Modèle B global	Modèle B différencié
2019	1 660 579 614	1 917 411 576	2 026 624 429	2 083 595 693

TABLE 5.19 – Prédictions du modèle Random Forest

Le tableau suivant présente les résultats de prédicition des différents modèles implémentés avec l'algorithme XGBoost :

	Prédictions des différents modèles			
	Modèles XGBoost			
Année N	Modèle A global	Modèle A différencié	Modèle B global	Modèle B différencié
2019	1 561 868 840	1 967 686 860	1 018 816 484	1 425 869 582

TABLE 5.20 – Prédictions du modèle XGBoost

Nous rappellons que le gestionnaire sinistre surestime les ultimes d'au moins 20% : un ultime « juste » serait plutôt égal à $2-2.1$ Mds €, en restant prudents. Ainsi, à partir des tableaux 5.19 et 5.20 nous pouvons conclure que les modèle A-global-Random Forest et

A-global-XGBoost donnent des estimations trop faibles, ce qui est en ligne avec l'exercice de backtesting qui avait été effectué plus haut. En revanche, le modèle B-global-Random Forest, qui est le modèle retenu, fournit des estimations plus cohérent avec le « juste » ultime.

Variabilité du modèle global B Random Forest

Une étude sur la variabilité du modèle Random Forest entraîné sur les sinistres clos et RBNS complétés a été effectuée. En effet, ce modèle fournit des résultats plus cohérents ainsi nous voulons étudier la variation relative à l'ultime prédit par ce modèle pour nous assurer de la fiabilité du modèle.

Dans un premier temps, nous calculons un coefficient de variation relatif à l'ultime prédit en itérant 500 fois l'apprentissage du modèle B. Le résultat de ce calcul est présenté dans le tableau suivant :

	Moyenne	Ecart type	Coefficient de variation
Modèle B global	2 056 143 444	127 047 408	6.2%

TABLE 5.21 – Incertitude sur le modèle B

Nous pouvons constater que le coefficient de variation est relativement faible : les estimations sont peu volatiles.

Afin de réaliser une étude plus approfondie sur l'incertitude du modèle nous construisons un intervalle de confiance à partir des 500 itérations réalisées sur l'apprentissage du modèle.

Remarque : Le choix du nombre d'itérations est limité par le temps de calcul. En effet, un nombre plus important nécessiterait un temps de calcul relativement long.

La méthodologie suivie pour la construction de l'intervalle de confiance est la suivante :

1. Nous itérons n fois l'apprentissage du modèle B.
2. Nous récupérons le vecteur des prédictions de taille n.
3. Nous supposons que la distribution des prédictions suit une loi N (μ, σ^2).
4. Nous supposons que la variance est inconnue et nous estimons cette dernière avec la variance empirique.
   Estimation de la variance par la variance empirique : $S_n^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(X_i - \bar{X_n}\right)^2$
   Ainsi nous avons : $S_n^2 \sim \frac{\sigma^2}{(n-1)} \chi^2_{(n-1)}$
5. Pour estimer μ nous utilisons la moyenne empirique $\bar{X_n} = \frac{1}{n} \sum_{i=1}^{n} X_i$ qui a pour loi N$(\mu, \frac{S_n^2}{n})$. Ainsi, $\sqrt{n}(\frac{\left(\bar{X_n} - \mu\right)}{S_n})$ suit une loi de Student à $(n-1)$ degrés de liberté.
6. L'intervalle de confiance de la moyenne s'écrit donc comme suit :
   $IC_{(1-\alpha)}(\mu) = \left[\bar{x}_n - t_{(1-\alpha/2)} \frac{S_n}{\sqrt{(n}}, \bar{x}_n + t_{(1-\alpha/2)} \frac{S_n}{\sqrt{(n)}}\right]$, où $\bar{x}_n$ et S_n^2 sont les estimateurs ponctuels respectives de la moyenne μ et de la variance σ^2

Pour $n = 500$ nous construisons un intervalle de confiance à 90% de la moyenne et nous obtenons les résultats suivants :

$$IC_{(90\%)}(\mu) = [1.9Mds, 2.1Mds]€, \text{ avec } \mu = 2Mds\ €$$

La figure suivante présente la distribution des ultimes prédits avec le modèle B global Random Forest avec la valeur moyenne représentée en rouge et les bornes de l'intervalle de confiance en bleu :

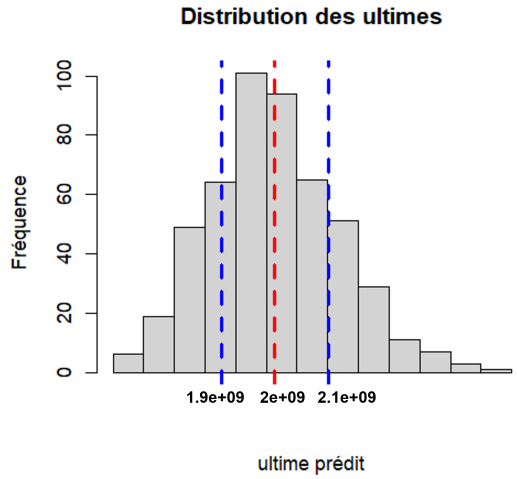


FIGURE 5.22 – Intervalle de confiance à 90% de la moyenne des prédictions du modèle B global Random Forest

Les résultats obtenus sont très satisfaisants, malgré la petite taille de l'échantillon, dans 90% des cas la valeur estimative de la provision est comprise dans l'intervalle $[1.9Mds, 2.1Mds]€$.

5.3 Méthodes agrégées

Dans cette partie, nous allons appliquer les approches classiques décrites précédemment à notre base de sinistres graves. Ainsi nous agrégeons les données de charge en triangle de charge : «Année de survenance de sinistre» x «Année de développement».

Nous rappelons que les méthodes classiques ne nous permettent pas de séparer la réserve des IBNyR de celle des RBNS. Ainsi, la principale difficulté avec cette étude réside dans l'estimation des provisions en distinguant une provision pour RBNS et une provision pour IBNyR.

Afin de constituer la réserve RBNS nous retraitons la réserve globale de nombre de tardifs « graves ». Pour ce faire, nous projetons le triangle de nombres relatifs à l'observation des sinistres graves : nous construisons le triangle « Année de survenance de sinistre » x « Année de dépassement du seuil de 500k €». Nous projetons ainsi un nombre de tardifs « graves » : 590. Ensuite, le nombre des tardifs est apporté au nombre total des dossiers encore ouverts en 2019 (1685 dossiers), nous obtenons ainsi une part de tardifs « graves » de 35%.
Les provisions effectuées concernent ainsi les 1095 RBNS observés au 31/12/2019, ainsi que 590 tardifs : les RBNS représentent donc 65% des sinistres à provisionner.

Deux méthodes ont été implémentées. Nous avons d'abord calibré un Chain-Ladder déterministe classique, de deux manières : le premier calibrage, que nous notons "Chain-Ladder standard", consiste en la considération de l'ensemble des années à l'exception de l'année de survenance 1996 qui semble atypique dans le triangle de liquidation ; quant au second calibrage, que nous noterons "Chain-Ladder alternatif", intègre quelques autres retraitements sur les années jugées atypiques dans le triangle de liquidation des charges. La deuxième méthode est le modèle stochastique de Mack.

5.3.1 Méthode Chain-Ladder

Facteurs de développement

À partir du triangle de charges cumulées, nous calculons ensuite les facteurs Chain-Ladder sur l'ensemble des années de survenance, à l'exclusion de 1996 qui semble atypique. Les résultats obtenus sont les suivants :

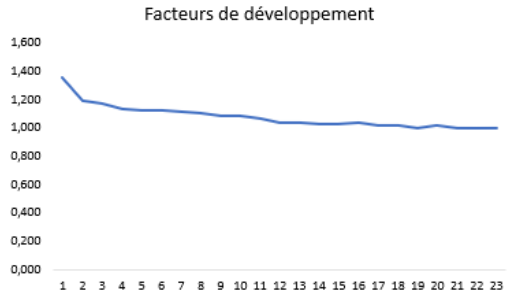


FIGURE 5.23 – Facteurs de développement Chain-Ladder

Afin de calculer les facteurs de développement de la deuxième méthode nous appliquons des retraitements sur certaines années comme suit :

- Facteur de développement 5 : nous excluons l'année de survenance 2003.
- Facteur de développement 11 : nous excluons l'année de survenance 1997.
- Facteur de développement 19 : nous excluons l'année de survenance 2000.

La figure suivante présente les facteurs de développement et les facteurs retraités sont entourés en bleu :

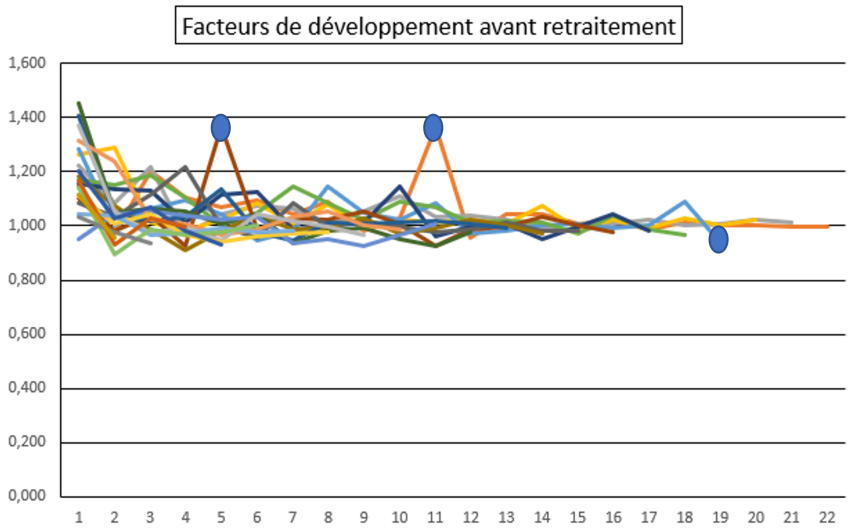


FIGURE 5.24 – Facteurs de développement avant retraitement

Les retraitements ont été décidés en concertation avec un actuaire expert en provisionnement.

Résultats

À partir du triangle de charges et les facteurs de développement nous constituons une charge ultime pour chaque année de survenance à partir du quelle nous pouvons déduire le montant de la réserve globale au titre des RBNS et IBNyR.

Le retraitement de cette réserve des 35% de tardifs « graves », nous donne une provision au titre des 1095 RBNS au 31/12/2019. Les résultats obtenus avec les deux calibrages Chain-Ladder standard et Chain-Ladder Alternatif sont présentés dans le tableau suivant :

	Chain-Ladder	Chain-Ladder Alternatif	Ecart
Provision RBNS	1 505 248 424	1 494 984 394	-0.69%
Paiements à date	633 968 524	633 968 524	0.00%
Charge Ultime	2 139 216 948	2 128 952 917	-0.48%

TABLE 5.22 – Résultats de la méthode Chain-Ladder

5.3.2 Méthode Mack

Le modèle de Mack est un modèle stochastique basé sur Chain-Ladder. Le modèle de Mack permet d'obtenir une erreur d'estimation.

Nous obtenons les résultats suivants avec l'implémentation sous R de cette méthode, à l'aide de la fonction MCL du package Chain-Ladder :

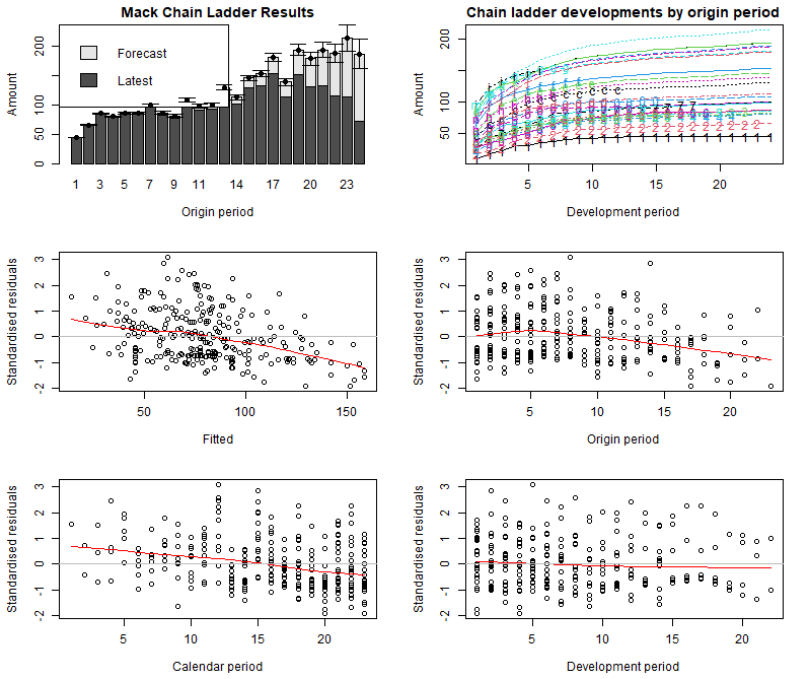


FIGURE 5.25 – Résultats du modèle de Mack

L'incertitude est importante pour les générations dont nous disposons de peu de données par rapport aux années de développement (notamment celle de 2019), la prévision est supérieure aux règlements effectués. L'analyse des résidus nous montre que le modèle s'adapte aux données.

Le tableau suivant présente les résultats obtenus avec le modèle de Mack :

Ultime	Réserve globale	RMSE
4 285 664 400	2 349 601 266	495 777 553

TABLE 5.23 – Résultats de la méthode Mack

Afin d'obtenir le montant de la provision au titre des RBNS au 31/12/2019, nous appliquons le même traitement appliqué sur la méthode Chain-Ladder : en effet, les résultats présentés ci-dessus représentent l'ultime estimé pour les RBNS+IBNyR. Ainsi, nous retrouvons une provision au titre des RBNS de 1.52 Mds €et une charge ultime de 2.16 Mds €.

À ce stade, la RMSE calculée avec le modèle de Mack est obtenue sur la réserve globale, ceci nous donne une vision globale sur l'erreur d'estimation. Afin de séparer l'erreur commise sur l'estimation des RBNS de celle des IBNyR, une étude sur les IBNyR est nécessaire.

5.4 Comparaison des résultats

En comparant les résultats du modèle à états aux résultats du modèle global Random Forest entraîné sur les sinistres clos et RBNS complétés, nous constatons que le modèle à états donne une charge ultime supérieure de 0.3 Mds € à celle estimée par le modèle B -

global Random Forest. L'estimation du modèle B global Random Forest est plus proche de l'ultime « juste ».

Les résultats obtenus avec la méthode Chain-Ladder et Mack quant à eux sont quasiment identiques, l'ultime calculé avec ces méthodes s'élève à 2.1 Mds €. Cette estimation est cohérente avec l'estimation des deux modèles individuels et l'ultime « juste » de 2.0 Mds € (obtenu par retraitement de 20% de la provision constituée par les gestionnaires sinistres).
En conclusion, à ce stade, nous ne pouvons pas dire qu'un modèle est meilleur qu'un autre : l'implémentation de ces différents modèles individuels permet essentiellement de tirer un maximum d'informations sur le développement des sinistres selon leurs caractéristiques, et peuvent permettre d'obtenir des estimations plus fines par sous-populations.

Enfin, il nous semble essentiel de dire un mot sur la comparaison des modèles en termes d'incertitude : il est en effet primordial, lorsque nous comparons des modèles entre eux, de comparer non seulement l'estimation en moyenne, mais également les mesures d'incertitude autour de cette valeur. Les modèles implémentés ne peuvent tous être comparés sur la base des mêmes mesures, nous avons cependant récapitulé l'ensemble des mesures disponibles sur chacun des modèles.

Le tableau suivant récapitule les estimations d'ultime ainsi que les intervalles de confiance à 90% construits sur les prédictions lorsque c'est possible :

Modèle	Ultime	RMSE	Erreur de process	Erreur d'estimation	Borne inf	Borne sup
Modèle à états	2.3 Mds	371 182 900	70 496 366	300 686 534	2.2 Mds	2.4 Mds
B global RF	2.0 Mds			127 047 408	1.9 Mds	2.1 Mds
Chain-Ladder	2.1 Mds					
Mack	2.1 Mds					

TABLE 5.24 – Comparaison des résultats

Nous rappelons qu'avec le modèle de Mack il est possible de calculer la RMSE et la séparer en erreur de process plus une erreur d'estimation, en revanche une étude sur les IBNyR est nécessaire afin de les séparer des RBNS.

Ce tableau nous donne une vue d'ensemble sur les moyens qui nous permettent de comparer les modèles de provisionnement ligne à ligne avec les méthodes agrégées et sur les travaux qu'il reste à effectuer afin de réaliser une comparaison complète.

Chapitre 6

Extension des modèles de provisionnement ligne à ligne implémentés

L'implémentation des différents modèles présentés dans les chapitres précédents a montré que les modèles individuels pouvaient donner des résultats tout à fait comparables à ceux obtenus par des méthodes plus classiques. Par ailleurs, nous avons appliqué des modèles qui certes traitent les données individuelles et donc apportent plus de sens et plus d'explications sur la dynamique de paiements des sinistres, mais qui pourraient, dans leur spécification, se prêter au provisionnement de n'importe quel type de sinistres, du moment que nous disposons d'assez de covariables pour les modèles de machine learning, et d'assez de paiements pour le calibrage du modèle stochastique à états. Nous souhaitons dans ce chapitre aller plus loin que le cadre strict des spécifications des modèles tels que présentés dans la littérature : le but est de tenir compte des spécificités du portefeuille sur lequel nous travaillons. Ceci, afin d'exploiter la flexibilité des modèles individuels, et d'évaluer dans quelle mesure cette flexibilité peut améliorer les résultats.

Ainsi, nous allons explorer deux spécificités de notre base de données :

- La première tient au fait que nous modélisons l'ultime de sinistres graves, ces derniers étant définis par le dépassement d'un certain seuil de charge. En particulier, le modèle stochastique à états est calibré sans tenir compte de cet événement de dépassement de seuil, alors qu'intuitivement, les distributions de paiements pourraient être différentes suivant la position du paiement cumulé ou de la charge du sinistre.
- La seconde met en avant le fait que les modèles individuels (paramétriques et non-paramétriques) permettraient de meilleures prédictions si des covariables plus prédictives de l'ultime étaient disponibles. Nous pensons en particulier à des variables telles que la gravité des blessures ou le taux d'AIPP. Dans le cas des modèles paramétriques, ces variables permettent de construire des clusters plus fins en termes de sévérité des sinistres, en amont de l'application du modèle en lui-même. Dans le cas des modèles non-paramétriques, ces variables pourraient améliorer leur précision.

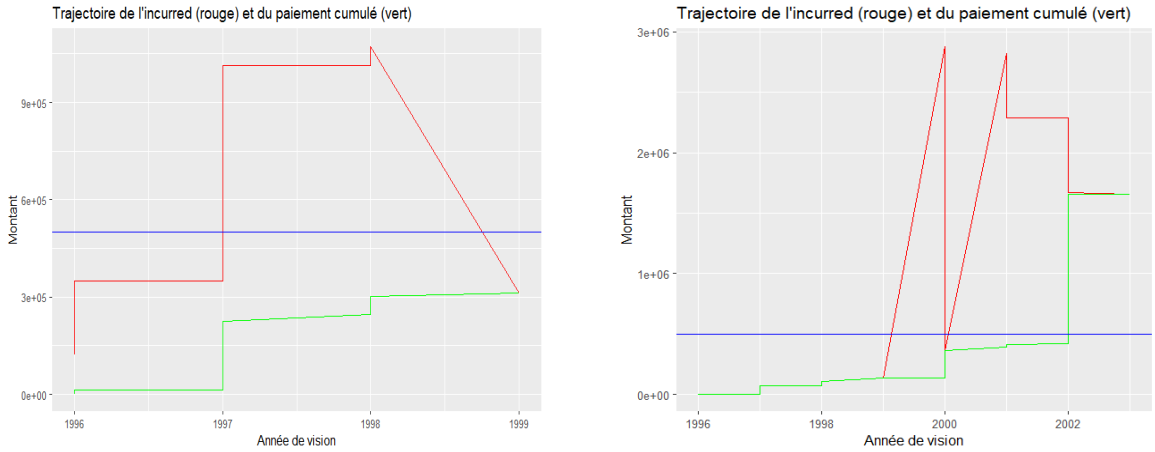
Dans cette optique, nous allons dans un premier temps implémenter le modèle stochastique à états sur nos données en tenant compte de l'évènement de dépassement de seuil de 500k € de charge qui définit nos sinistres graves. Deux calibrages sont effectués : un calibrage du modèle selon la position du dernier paiement cumulé et un calibrage selon

la position de la charge par rapport à ce seuil. Les résultats des deux calibrages sont ensuite comparés aux résultats du modèle à états "de base" (celui des chapitres précédents). Notons que la prise en compte de ce dépassement de seuil n'est pas pertinente pour les modèles de machine learning, qui tiennent compte implicitement de cet événement via les variables "paiement cumulé à date" et "charge à date".

Puis, afin de traiter le second "raffinement" de nos modèles, nous avons récupéré le taux d'AIPP, variable a priori prédictive de l'ultime puisque les gestionnaires sinistres se reposent beaucoup sur cette information pour évaluer une provision dossier/dossier. Cependant, cette variable étant difficile à extraire, nous ne l'avons obtenue que pour une partie des sinistres seulement. Ainsi, pour pouvoir évaluer l'impact de l'ajout de cette variable sur les sorties des modèles individuels, nous avons constitué une nouvelle base, que nous appelons "base restreinte", constituée uniquement des sinistres pour lesquels nous disposons du taux d'AIPP. Nous implémentons le modèle stochastique à états, ainsi que le Random Forest, sur cette base restreinte, avec le taux d'AIPP d'une part, et sans le taux d'AIPP d'autre part.

6.1 Prise en compte du dépassement de seuil

Le portefeuille de données est constitué de sinistres graves : nous rappelons que les sinistres sont considérés comme graves dès qu'un dépassement de seuil de 500k € de charge est observé. Autrement dit, **"sinistre grave un jour, sinistre grave toujours"**. Cependant, certains sinistres clos présentent une correction de charge sur les dernières années de développement, c'est-à-dire que la charge redescend en dessous du seuil de 500k €, comme nous pouvons le voir sur la figure 6.1a. Pour autant, le sinistre est tout de même considéré comme grave.



(a) Trajectoire d'un sinistre clôturé en dessous du seuil de 500k € de charge

(b) Trajectoire d'un sinistre clôturé au-dessus du seuil de 500k € de charge

FIGURE 6.1 – Exemple de trajectoire de charge (en rouge) et du paiement cumulé (en vert)

La figure ci-dessus permet de constater que la trajectoire des paiements cumulés est différente selon que le sinistre soit clos en dessous ou au-dessus du seuil. Dans le premier cas (sinistre clos sous le seuil), le paiement cumulé reste sous le seuil de 500k € jusqu'à sa

clôture. Dans le second cas (sinistre clos au-dessus du seuil), le paiement cumulé dépasse 500k € et ne redescend plus en dessous, jusqu'à sa clôture. La seule possibilité pour que le paiement cumulé redescende sous le seuil, serait d'avoir un recours suffisamment important : nous n'en observons pas sur les sinistres clôturés de notre base.

Au vu du fait que 52% des sinistres clos de la base de données ont un ultime inférieur à 500k €, il serait intéressant de prendre cette particularité en compte dans le calibrage du modèle stochastique à états. En effet, l'intuition sous-jacente à ce raffinement du modèle stochastique à états est que les distributions de paiements seraient différentes selon que nous ayons déjà dépassé le seuil ou pas, avec potentiellement des paiements plus importants si nous nous situons au-dessus du seuil. Or, si près de la moitié des sinistres ne dépassent pas ce seuil (en paiement cumulé), il est possible que la provision estimée soit plus faible qu'avec le modèle global, car nous appliquerions une distribution plus faible à une grande partie des sinistres.

Ainsi, nous avons décidé de modéliser l'évènement de dépassement de seuil selon deux méthodes :

- La première approche repose sur la position du dernier paiement cumulé, c'est-à-dire de calibrer les évènements de paiement et de clôture suivant la position du dernier paiement cumulé associé à chacun des dossiers et ensuite calibrer deux lois de paiements en séparant les paiements associés aux dossiers pour lesquels le dernier paiement cumulé est supérieur à 500k € de ceux inférieur à 500k €.
- La seconde méthode quant à elle consiste à calibrer le modèle sur la charge, c'est-à-dire de calibrer les évènements de paiement et de clôture suivant la position de la charge à chaque date de vision et ensuite calibrer deux lois de paiements en séparant les paiements relatifs aux charges supérieur à 500k € de ceux inférieur à 500k €.

6.1.1 Calibrage du modèle en observant la position du dernier paiement cumulé

Tout d'abord, nous devons calibrer les fréquences de paiement et de clôture en tenant compte de la position du dernier paiement cumulé. Ensuite, nous calibrons une loi de paiement pour chaque type de dossiers : les dossiers pour lesquels le dernier paiement cumulé est supérieur à 500k € d'une part, et les dossiers pour lesquels le dernier paiement cumulé est inférieur à 500k € d'autre part. Enfin, une provision par formule fermée est calculée.

Commençons par étudier rapidement la composition de notre base de données sur ce critère de position à la clôture (pour les sinistres clos) ou à la dernière année de développement observée (pour les RBNS).

Nous observons 1365 sinistres clos dont 704 sont associés aux sinistres ayant un dernier paiement cumulé en dessous du seuil de 500k € et 661 relatifs aux sinistres ayant un dernier paiement cumulé au-dessus du seuil de 500k €. Les distributions des délais de clôture observés sont présentées dans la figure ci-dessous : les sinistres clôturés avec un ultime supérieur à 500k € ont globalement des délais de clôture plus longs.

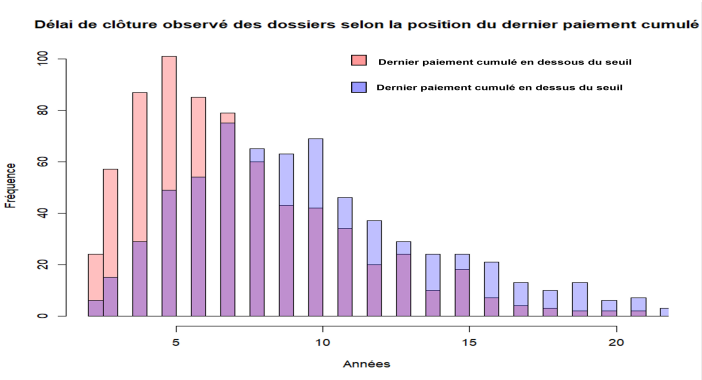


FIGURE 6.2 – Délai de clôture observé des dossiers selon la position du dernier paiement cumulé

Par ailleurs, nous observons 1095 RBNS, dont 832 sont associés aux sinistres ayant un dernier paiement cumulé en dessous du seuil de 500k € et 263 relatifs aux sinistres ayant un dernier paiement cumulé au-dessus du seuil de 500k €. Les distributions des expositions sont présentées dans la figure ci-dessous :

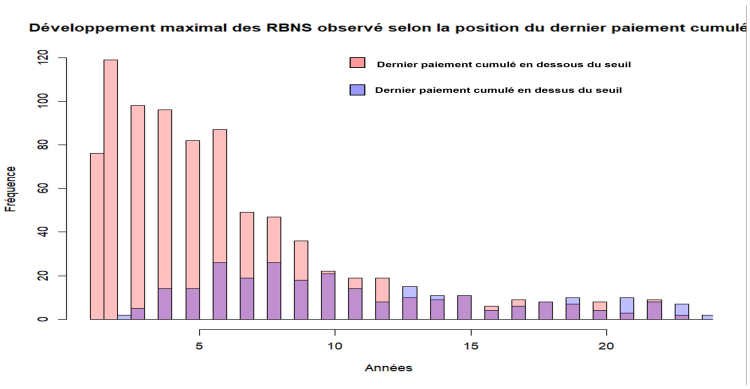


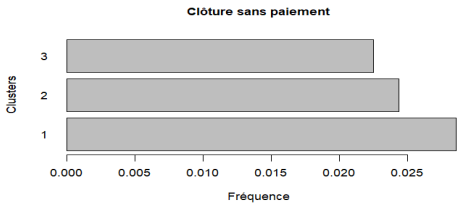
FIGURE 6.3 – Développement maximal des RBNS observés selon la position du dernier paiement cumulé

Calibrage des fréquences de paiement et de clôture tenant compte de la position du dernier paiement cumulé

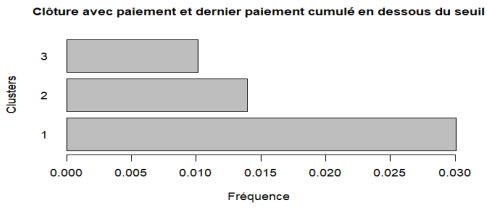
Comme pour le modèle de base, nous avons besoin d'estimer les fréquences associées aux événements de paiement et de clôture, pour calculer le montant de provision. Par conséquent, pour ce modèle, au lieu de travailler sur trois évènements nous en considérons cinq :

- 1 : Clôture sans paiement
- 2 : Paiement avec clôture et dernier paiement cumulé associé en dessous du seuil
- 3 : Paiement avec clôture et dernier paiement cumulé associé au-dessus du seuil
- 4 : Paiement sans clôture et dernier paiement cumulé associé en dessous du seuil
- 5 : Paiement sans clôture et dernier paiement cumulé associé au-dessus du seuil

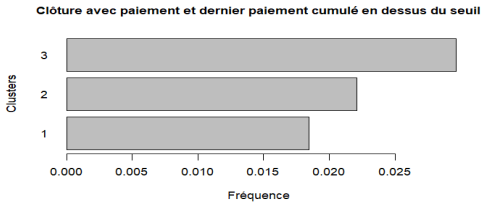
Ensuite, de la même façon que pour le modèle de base, nous estimons les fréquences associées à ces évènements. Les fréquences h_1, h_2, h_3, h_4 et h_5 calibrées sont les suivantes :



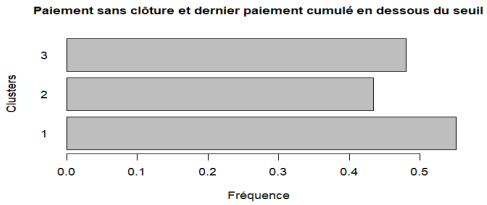
(a) Fréquence de clôture sans paiement (h_1)



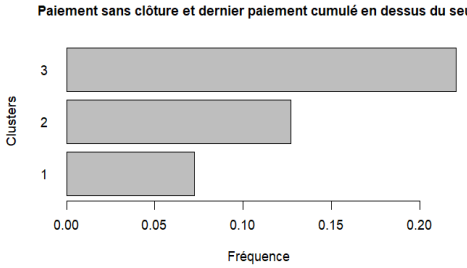
(b) Fréquence de clôture avec paiement et dernier paiement cumulé en dessous du seuil (h_2)



(c) Fréquence de clôture avec paiement et dernier paiement cumulé au-dessus du seuil (h_3)



(d) Fréquence de paiement sans clôture et dernier paiement cumulé en dessous du seuil (h_4)



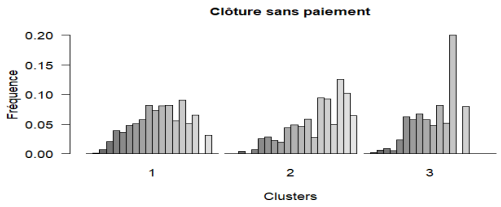
(e) Fréquence de paiement sans clôture et dernier paiement cumulé au-dessus du seuil (h_5)

FIGURE 6.4 – Fréquences des évènements calibrées par cluster et par position

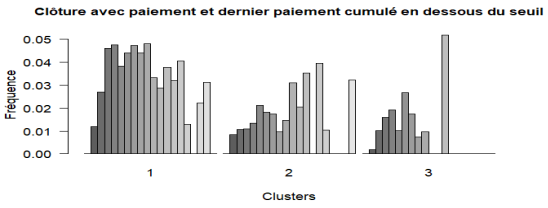
À partir des résultats obtenus, nous pouvons constater que :

- Les sinistres du cluster 1 présentent plus de clôtures sans paiement, comparés au deux autres clusters.
- Selon la position du dernier paiement cumulé par rapport au seuil, nous observons un comportement différent. Les sinistres ayant un dernier paiement cumulé en dessous du seuil se caractérisent par des paiements sans clôture plus fréquents que les paiements avec clôture, quel que soit le cluster. Pour les dossiers ayant un dernier paiement cumulé au-dessus du seuil, nous constatons que les sinistres du cluster 1 se caractérisent par des paiements avec clôture plus fréquents que les paiements sans clôture et les sinistres des clusters 2 et 3 présentent des paiements avec clôture moins fréquents que les paiements sans clôture.

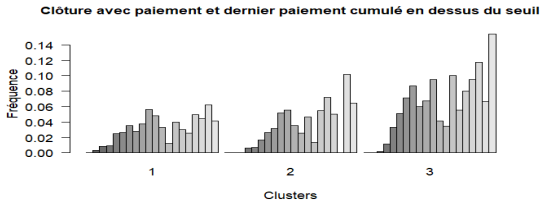
Passons maintenant aux fréquences calibrées : si nous supposons que les fréquences des évènements sont constantes par morceaux, elles dépendent du temps écoulé depuis la déclaration des sinistres. Nous précisons que le pas de temps est annuel.



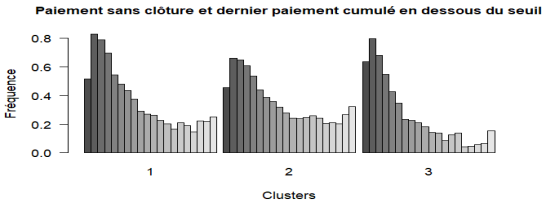
(a) Fréquence de clôture sans paiement (h_1)



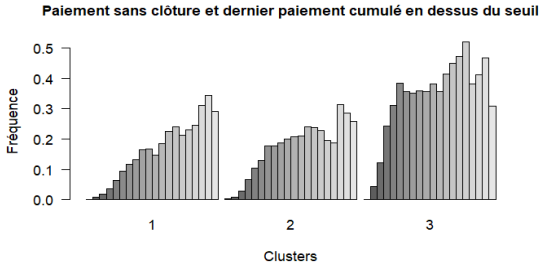
(b) Fréquence de clôture avec paiement et dernier paiement cumulé en dessous du seuil (h_2)



(c) Fréquence de clôture avec paiement et dernier paiement cumulé au-dessus du seuil (h_3)



(d) Fréquence de paiement sans clôture et dernier paiement cumulé en dessous du seuil (h_4)



(e) Fréquence de paiement sans clôture et dernier paiement cumulé au-dessus du seuil (h_5)

FIGURE 6.5 – Féquences des évènements calibrées par cluster et par position dans le cas où les fréquences sont dépendantes du temps écoulé depuis la déclaration des sinistres

- Les fréquences de clôture h_1 sont maximales pour les sinistres d'âge autour de 15 ans ; ces estimations permettent de caractériser les sinistres à fort potentiel de développement long : ceux-ci sont plutôt d'âge 18-19 ans (fréquences faibles) et appartiennent aux clusters 1 et 2.
- Les fréquences de clôture avec paiement présentent un profil très différent dépendamment de la position du dernier paiement cumulé et des clusters :
  - Les sinistres des clusters 1 et 2 présentent globalement les mêmes fréquences de clôture sans paiement quelque soit de la position du dernier paiement cumulé. Par ailleurs, les sinistres d'âge autour de 1 à 3 ans présentent des clôtures avec paiement plus faibles que ceux d'âge autour de 10 ans pour le cluster 1 et ceux d'âge autour de 13 ans pour le cluster 2.
  - Les sinistres du cluster 3 présentent des fréquences de clôture avec paiement plus élevées dans le cas où le dernier paiement cumulé est au-dessus du seuil. Les sinistres développés depuis 1 à 2 ans présentent des clôtures avec paiement moins fréquentes quelque soit la position du dernier paiement cumulé. Dans le cas où le dernier paiement cumulé est en dessous du seuil, les sinistres d'âge autour de 15 ans présentent un pic de clôture avec paiement, puis les fréquences

deviennent nulles , et dans le cas où le dernier paiement cumulé est au dessus du seuil, les clôtures avec paiement sont maximales pour les sinistres d'âge autour de 20 ans.

- Les fréquences de paiement sans clôture dépendent aussi de la position du dernier paiement cumulé. En effet, nous observons :
  - Une croissance des fréquences de paiements sans clôture des sinistres des différents clusters ayant un dernier paiement cumulé au-dessus du seuil et nous observons le scénario inverse sur les fréquences de paiements sans clôture des sinistres des différents clusters ayant un dernier paiement cumulé en dessous du seuil.
  - Les sinistres développés depuis 1 à 3 ans avec un dernier paiement cumulé en dessous du seuil présentent des paiements à venir potentiellement plus fréquents que les sinistres développés depuis 1 à 3 ans avec un dernier paiement cumulé au dessus du seuil, en revanche les sinistres développés depuis 18 à 20 ans présentent un potentiel accru quand le dernier paiement cumulé est au-dessus.

Calibrage des lois des paiements

Afin de calibrer les paiements, nous les séparons en deux types : ceux associés aux sinistres ayant un dernier paiement cumulé supérieur au seuil de 500k € et ceux associés aux sinistres ayant un dernier paiement cumulé inférieur au seuil de 500k €.

La figure suivante présente la distribution des paiements selon la position du dernier paiement cumulé associé. Les deux distributions sont bien distinctes : les paiements associés à un dernier paiement cumulé observé inférieur au seuil sont tous inférieurs à 500k €, tandis que les paiements associés à un dernier paiement cumulé observé supérieur à ce seuil peuvent être inférieurs à 500k € mais la distribution s'étale également au-delà de cette valeur.

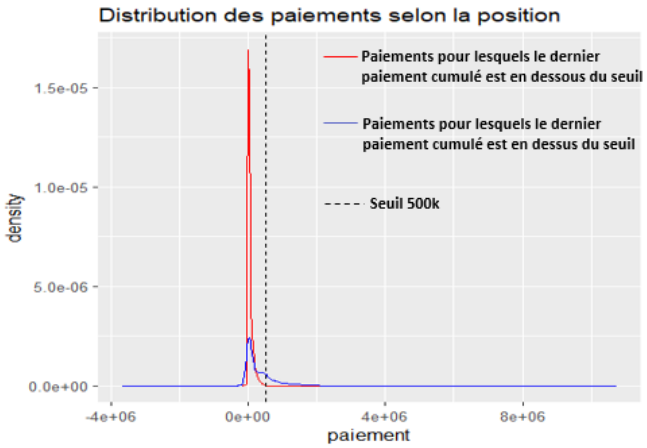


FIGURE 6.6 – Distribution des paiements selon la position du dernier paiement cumulé associé (en dessous ou au-dessus du seuil 500k €)

Comme précédemment, nous testons l'adéquation de deux types de lois : la loi Log-Normale et la loi mélange Log-Normale Exponentielle. La loi Log-Normale s'adapte bien aux paiements observés (voir figure C.1 et C.2 en annexe), en revanche les résultats sont très sensibles au facteur de shift de la distribution. La loi mélange, quant à elle, s'adapte

moins bien aux données. En effet, la loi Log-Normale de la loi mélange appliquée aux paiements non extrêmes ne peut être validée. Ainsi, nous gardons la loi Log-Normale décalée de 1.1 * la valeur minimale des recours.

Ainsi, deux lois de paiements sont calibrées. Pour ce faire, deux lois Log-Normale sont spécifiées pour caractériser la distribution des paiements, une pour les paiements associés aux sinistres ayant un dernier paiement cumulé en dessous du seuil et une deuxième pour ceux associés aux sinistres ayant un dernier paiement cumulé au dessus du seuil. Les paramètres calibrés sont présentés en annexe C.3 et C.4.

Estimation de la provision par formule fermée

A présent que les fréquences et les paramètres des lois de paiement sont calibrés, nous pouvons estimer le montant de la provision au titre des RBNS au 31/12/2019. En s'appuyant sur la formule fermée (3.5) le montant de la provision pour un seul RBNS est calculé comme la somme pondérée des réserves attendues associées aux évènements de dépassement de seuil (dernier paiement cumulé en dessous du seuil de 500k € et dernier paiement cumulé au-dessus du seuil de 500k €) par leurs probabilités. Les probabilités sont calculées comme le rapport entre le nombre de paiements relatifs à chaque évènement et le nombre total des paiements. La formule s'écrit comme :

$$\mu(s) = \mu(0) = p_1 * \frac{y_2 h_2 + y_4 h_4}{h_1 + h_2} + p_2 * \frac{y_3 h_3 + y_5 h_5}{h_1 + h_3} \tag{6.1}$$

Avec :

- p_1 : probabilité que le dernier paiement cumulé soit en dessous du seuil de 500k €.
- p_2 : probabilité que le dernier paiement cumulé soit au-dessus du seuil de 500k €.

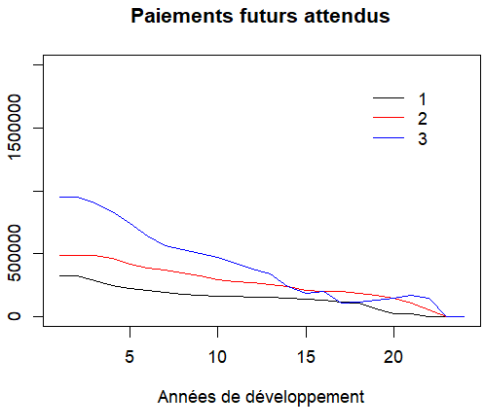


FIGURE 6.7 – Espérance des paiements futurs en fonction des clusters

Le montant de la provision, au titre des RBNS vus à fin 2019, que nous obtenons avec cette méthode s'élève à 1.00 Mds €.

6.1.2 Calibrage du modèle en observant la position de la charge à chaque année de développement

Nous présentons à présent le modèle implémenté suivant la position de la charge des dossiers, en suivant la même méthodologie que précédemment. Nous commençons par calibrer les fréquences des paiements et de clôture en tenant compte de la position de la charge par rapport au seuil. Ensuite, nous calibrons deux lois de paiements : une loi pour les paiements associés à une charge inférieure à 500k € et une loi les paiements associés à une charge supérieure à 500k €. Enfin, une provision est calculée par formule fermée.

Nous observons 1095 RBNS, dont 839 sont associés aux sinistres ayant une charge au-dessus du seuil de 500k € et 256 relatifs aux sinistres ayant une charge en dessous du seuil de 500k €. Les distributions des expositions sont présentées dans la figure ci-dessous :

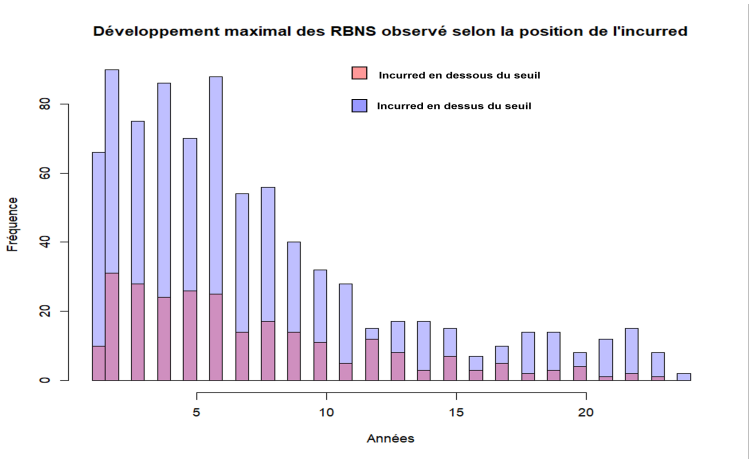


FIGURE 6.8 – Développement maximal des RBNS observé selon la position de la charge

Calibrage des fréquences de paiement et de clôture tenant compte de la position de la charge à chaque année de développement

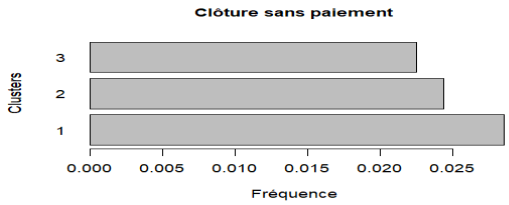
Comme précédemment, nous calibrons les fréquences de paiement et de clôture selon le type d'évènement : charge à date de vision en dessous ou au-dessus du seuil de 500k €. Les évènement en question sont les suivants :

- 1 : Clôture sans paiement
- 2 : Paiement avec clôture et charge en dessous du seuil
- 3 : Paiement avec clôture et charge au-dessus du seuil
- 4 : Paiement sans clôture et charge en dessous du seuil
- 5 : Paiement sans clôture et charge au-dessus du seuil

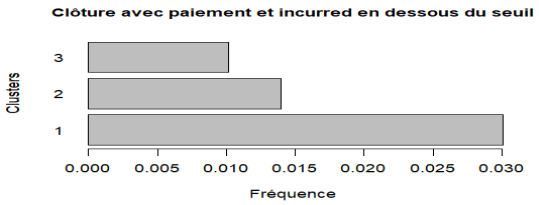
L'analyse de l'évènement de clôture sans paiement nous conduit à effectuer le même constat que dans l'étude précédente : les sinistres du cluster 1 présentent plus de clôtures sans paiement comparés à ceux des deux autres clusters.

Selon la position de la charge par rapport au seuil, nous observons un comportement différent. Pour les dossiers ayant une charge en dessous du seuil, nous constatons que les sinistres des clusters 1 et 2 se caractérisent par des paiements sans clôture plus fréquents que les paiements avec clôture. Les sinistres du clusters 3 quant à eux présentent

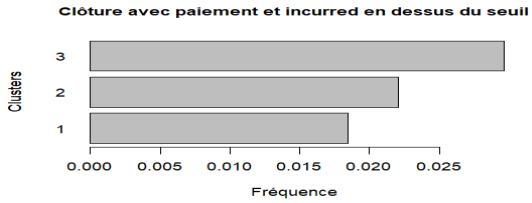
des clôtures avec paiement plus fréquentes que les paiements sans clôture. Concernant les dossiers ayant une charge au-dessus du seuil nous constatons que pour les différents clusters, la fréquence des paiements sans clôture est supérieure à celle des paiements avec clôture.



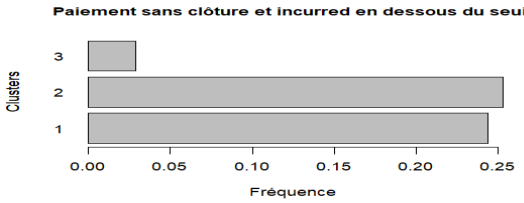
(a) Fréquence de clôture sans paiement (h_1)



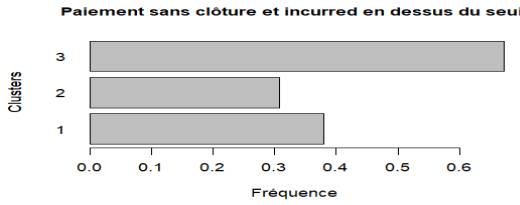
(b) Fréquence de clôture avec paiement et charge en dessous du seuil (h_2)



(c) Fréquence de clôture avec paiement et charge au-dessus du seuil (h_3)



(d) Fréquence de paiement sans clôture et charge en dessous du seuil (h_4)



(e) Fréquence de paiement sans clôture et charge au-dessus du seuil (h_5)

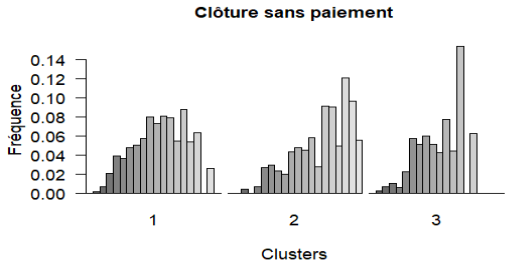
FIGURE 6.9 – Fréquences des évènements calibrées par cluster et par position (charge)

Passons maintenant aux fréquences calibrées : si nous supposons que les fréquences des évènements sont constantes par morceaux, elles dépendent du temps écoulé depuis la déclaration des sinistres. Nous précisons que le pas de temps est annuel.

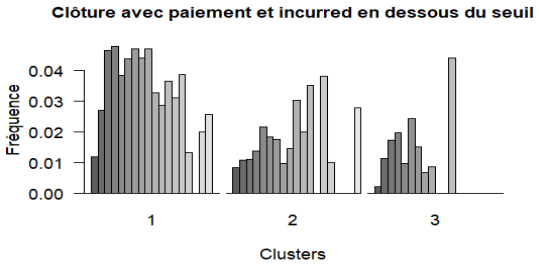
- Les fréquences de clôture h_1 sont maximales pour les sinistres d'âge autour de 14 et 15 ans dans les clusters 1 et 3 et autour de 19 ans dans le cluster 2 ; ces estimations permettent de caractériser les sinistres à fort potentiel de développement long : ceux-ci sont plutôt d'âge 19-20 ans (fréquences faibles) et appartiennent aux clusters 1 et 2.
- Les fréquences de clôture avec et sans paiement présentent un profil très différent dépendamment de la position de la charge et des clusters. En effet :
  - Les sinistres du cluster 1 ayant une charge en dessous du seuil et un âge autour de 2 ans sont moins clôturé avec paiement et présentent des paiements à venir potentiellement plus fréquents contrairement aux sinistres d'âge autour de de 10 ans qui présentent plus de clôture avec paiement et moins de paiement. Par ailleurs, les sinistres du cluster 1 ayant une charge au-dessus du seuil présentent des fréquences de paiement globalement stables sur la majorité des âges.

— Les sinistres du cluster 2 ayant une charge en dessous du seuil présentent des sinistres d'âge petit moins clôturés avec paiement et des fréquences de paiement plus élevées que ce que présente les sinistres à développement long. Pour les sinistres ayant une charge au-dessus du seuil le constat est le même que les sinistres du cluster 1 : les fréquences de paiement sont stables pour la majorité des âges.

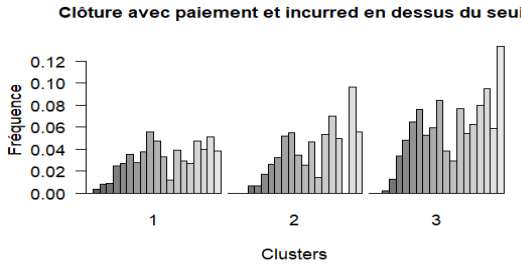
— Enfin, les sinistres du cluster 3 : les fréquences de paiements des sinistres ayant une charge en dessous du seuil sont quasiment nulles et celle de clôtures avec paiement sont maximales pour les sinistres d'âge 14 ans. Dans le cas où la charge est au-dessus du seuil les sinistres développés depuis 2 à 4 ans présentent un potentiel de paiement accru et des clôtures avec paiement moins fréquentes.



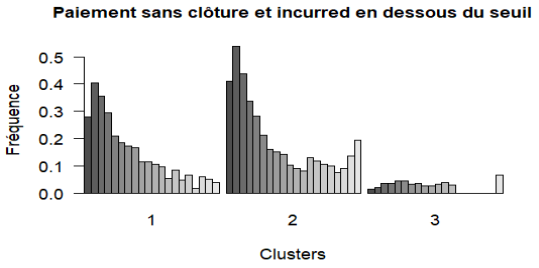
(a) Fréquence de clôture sans paiement (h_1)



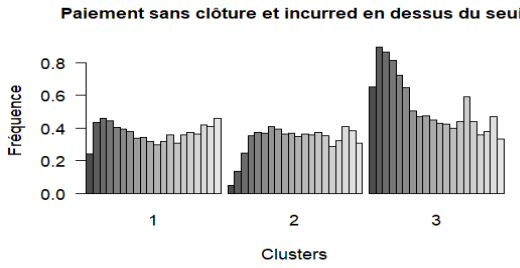
(b) Fréquence de clôture avec paiement et charge en dessous du seuil (h_2)



(c) Fréquence de clôture avec paiement et charge au-dessus du seuil (h_3)



(d) Fréquence de paiement sans clôture et charge en dessous du seuil (h_4)



(e) Fréquence de paiement sans clôture et charge au-dessus du seuil (h_5)

FIGURE 6.10 – Féquences des évènements calibrées par cluster et par position (charge) dans le cas où les fréquences sont dépendantes du temps écoulé depuis la déclaration des sinistres

Calibrage des paiements

Nous séparons les paiements en deux types : pour chaque date de vision, nous récupérons d'une part les paiements relatifs aux charges inférieures à 500k, et ceux relatif aux charges supérieures à 500k € d'autre part.

La figure suivante présente les distribution des paiements selon la position de la charge associé : elles sont assez similaires. Cela s'explique par le fait que la plupart des sinistres, y compris ceux qui ont des paiements faibles, ont une charge supérieure au seuil. Ainsi, le critère de position de la charge par rapport au seuil ne permet pas de séparer les paiements élevés des paiements plus faibles, aussi nettement que le critère de position du paiement cumulé par rapport au seuil.

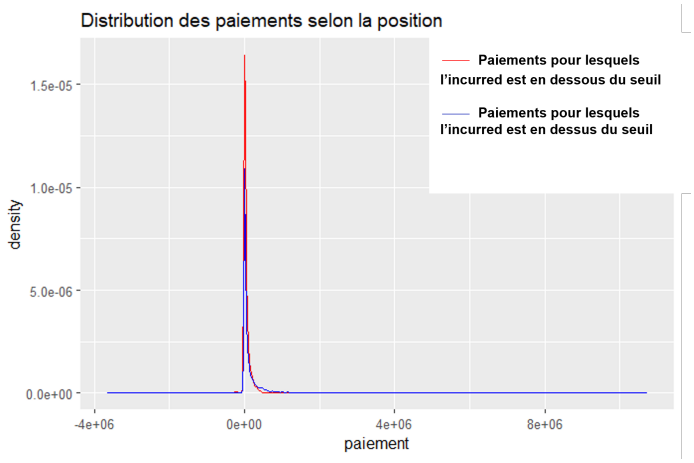


FIGURE 6.11 – Distribution des paiements selon la position de la charge (en dessous ou au-dessus du seuil 500k €)

Ici également nous testons deux lois : la loi Log-Normale et la loi mélange Log-Normale Exponentielle. La loi Log-Normale fitte mieux les paiements observés (voir C.5 et C.6 en annexe). Ainsi, nous gardons cette loi. Les paramètres calibrés sont présentés en annexe C.8 et C.7.

Estimation de la provision par formule fermée

Nous estimons à présent le montant de la provision au titre des RBNS au 31/12/2019. De la même façon que précédemment la provision pour un seul RBNS est calculée comme la somme pondérée des réserves attendues associées aux évènements de dépassement de seuil (charge en dessous du seuil de 500k € et charge au-dessus du seuil de 500k €), par leurs probabilités.

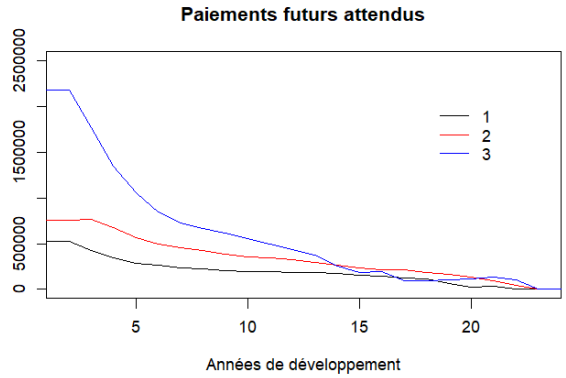


FIGURE 6.12 – Espérance des paiements futurs en fonction des clusters

Le montant de la provision, au titre des RBNS vus à fin 2019, que nous obtenons avec cette méthode s'élève à 1.60 Mds €.

6.1.3 Comparaison des deux approches - conclusion

Selon l'approche considérée nous obtenons des résultats différents, le montant de la provision obtenu avec le modèle à états avec dépassement de seuil modélisé sur la charge est légèrement inférieur comparé à celui obtenu avec le modèle à états de base qui s'élève à 1.67 Mds €. En revanche, le modèle à états avec dépassement de seuil modélisé sur le dernier paiement cumulé nous donne une provision nettement inférieure qui s'élève à 1.00 Mds €.

Si l'approche la plus "naturelle" aurait été de travailler sur la position de la charge, qui comporte en général plus d'information sur les sinistres graves aux développements longs, plutôt que sur le dernier paiement cumulé, nous constatons que la charge ne nous permet pas, dans cet exemple précis, de capter l'impact des différents types de trajectoires de développement sur l'ultime. Ainsi, le modèle avec dépassement de seuil modélisé sur la charge conduit à une provision similaire à celle obtenue par le modèle à états sans dépassement de seuil.

Par contre, il est intéressant de constater que la provision évaluée par le modèle avec dépassement de seuil modélisé sur le dernier paiement cumulé observé est beaucoup plus faible, ce qui est en ligne avec le fait qu'une part importante des sinistres graves sont en réalité clôturés à un coût inférieur au seuil de gravité.

Remarque : ces résultats ne donnent qu'une idée des tendances obtenues par les modèles. Pour avoir des valeurs plus fiables et pouvoir sélectionner le meilleur modèle, il faudrait étudier la variabilité des modèles et calculer les erreurs de process et d'estimation.

6.2 Etude de l'impact de l'ajout de la variable "Taux d'AIPP"

Nous étudions à présent l'impact de la présence d'une variable fortement corrélée avec la sévérité de sinistre sur le montant de la provision, sur les deux types de modèles

individuels implémentés (paramétrique et non-paramétrique). Ainsi, nous implémentons le modèle stochastique à états d'une part, et le modèle global Random Forest entraîné sur les sinistres clos et RBNS complétés d'autre part, sur les dossiers pour lesquels nous disposons du taux d'AIPP.

La base de données restreinte aux dossiers pour lesquels nous disposons du taux d'AIPP est constituée de 1376 dossiers dont 496 sont clos. Ainsi, sur 2460 dossiers, il nous reste 1084 dossiers pour lesquels l'information taux d'AIPP est indisponible. Notons que les taux d'AIPP récupérés sont les taux les plus récents enregistrés par le gestionnaire, et différents de 0. Si un taux d'AIPP remonte à 0, c'est qu'on n'a jamais eu l'information concernant le taux. Par ailleurs, l'extraction des taux d'AIPP s'est effectuée sur un périmètre plus restreint que celui de la base initiale.

	Base globale	Base restreinte
Nombre total de dossiers	2460	1376 (56% des dossiers de la base globale)
Nombre de dossiers clos	1365	496 (37% des clos de la base globale)
Nombre de RBNS	1095	880 (80% des RBNS de la base globale)
Ultime clos	1.3 Mds €	0.56 Mds € (43% de l'ultime au titre des sinistres clos de la base globale)
Provisions RBNS gestionnaires	2.0 Mds €	1.94 Mds € (96% des provisions au titre des RBNS 2019)

TABLE 6.1 – Analyse de la base globale et de la base restreinte

Les sinistres clos de la base restreinte, soit 37% des sinistres clos de la base globale représentent 43% de l'ultime des sinistres clos de la base globale, ce qui est cohérent et montre que les sinistres clos de la base restreinte sont en moyenne un peu plus graves en termes de sinistralité que les sinistres pour lesquels nous ne disposons pas du taux d'AIPP.
Par ailleurs, les RBNS de la base restreinte, qui représentent 80% des RBNS de la base totale, représentent également 96% de la provision RBNS évaluée par les gestionnaires sinistres.

6.2.1 Analyse des données

Afin d'avoir une idée de la répartition de l'ultime en fonction du taux d'AIPP, cinq tranches d'AIPP sont construites de façon intuitive : les victimes ayant un taux d'AIPP de 0% d'une part (aucune atteinte de l'intégrité physique et psychique), les victimes ayant un taux d'AIPP particulièrement élevé (supérieur ou égal à 60%), d'autre part. Entre ces deux classes, nous répartissons les taux d'AIPP en 3 tranches intermédiaires : $[1\% - 5\%]$, $[6\% - 29\%]$, $[30\% - 59\%]$. La répartition des dossiers en fonction de ces tranches est présentée dans la figure suivante : les sinistres comprenant un taux d'AIPP compris entre 6%

et 59% sont majoritaires. La distribution des taux d'AIPP sur cette base est cohérente avec le fait qu'elle est composée de sinistres graves.

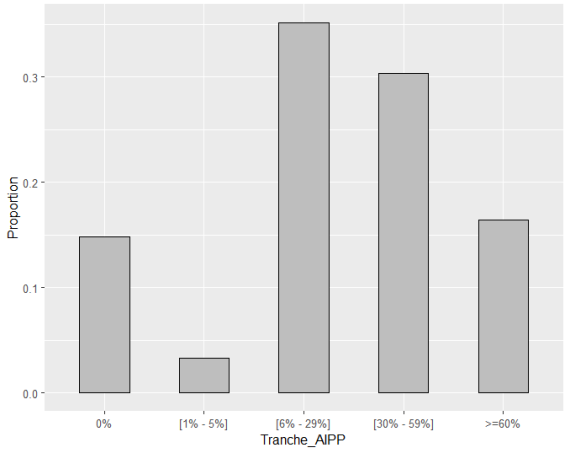


FIGURE 6.13 – Proportions de dossiers par tranche de taux d'AIPP

Ensuite, nous nous intéressons à la répartition de l'ultime suivant les tranches de taux d'AIPP pour les sinistres clos. La répartition de l'ultime en fonction de ces tranches est présentée dans la figure et tableau suivants :

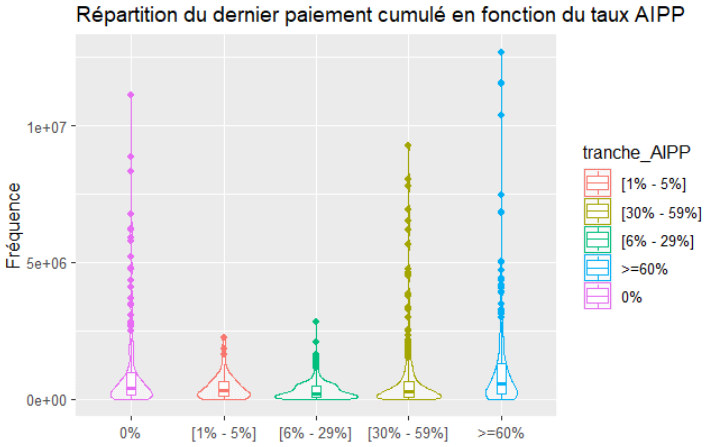


FIGURE 6.14 – Répartition de l'ultime en fonction des tranches du taux AIPP

Tranche d'AIPP	Moyenne(Ultime)	Sd(Ultime)	CV(Ultime)
0%	1 280 224	2 104 435	1.6
$[1\% - 5\%]$	547 610	605 128	1.1
$[6\% - 29\%]$	398 368	414 110	1.0
$[30\% - 59\%]$	771 278	1 420 690	1.8
$>= 60\%$	1 441 042	2 301 587	1.6

TABLE 6.2 – Statistiques sur l'ultime en fonction des tranches du taux d'AIPP

Nous observons une répartition d'ultime différente selon chaque tranche de taux d'AIPP. En effet, les tranches "0%" et ">= 60%" sont les plus onéreuses, et les plus volatiles avec la tranche $[30\% - 59\%]$.

Clustering supervisé sur les sinistres clos

Afin d'obtenir des clusters homogènes en termes de risque, nous utilisons des techniques de clustering supervisé : nous avons choisi d'utiliser l'algorithme CART, appliqué sur l'ensemble des sinistres clos. La variable cible est l'ultime payé.

L'arbre de décision nous donne quatre clusters qui se distinguent les uns des autres par le montant de la provision, le taux d'AIPP et l'usage du véhicule.

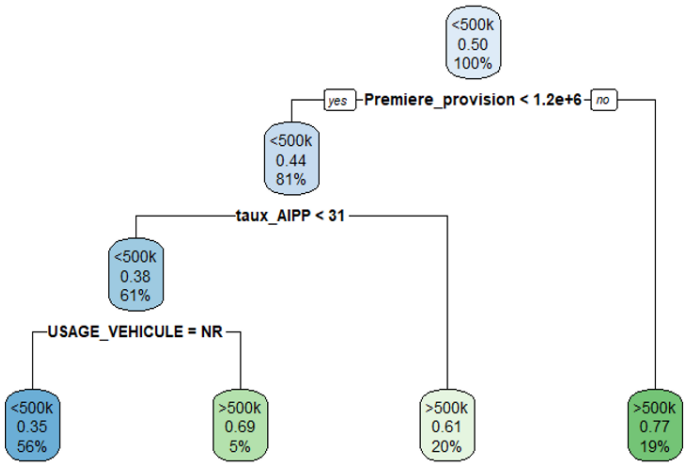


FIGURE 6.15 – Arbre CART sur la base d'AIPP

L'arbre simplifié ainsi construit (en figure ci-dessus) se lit de la façon suivante : chaque noeud est représenté par une classe d'ultime «Ultime inférieur à 500k €» et «Ultime supérieur à 500k €». La racine est ainsi scindée en deux noeuds suivant que la provision est inférieure ou non à 1.2 M €. Puis le noeud pour lequel nous avons une provision inférieure à 1.2 M € est à son tour scindé en deux suivant que le taux d'AIPP est inférieur ou non à 31%. Enfin, le noeud ayant un taux d'AIPP inférieur à 31% est aussi scindé en deux selon l'usage de véhicule.

La répartition de l'ultime en fonction des clusters est présentée dans la figure suivante :

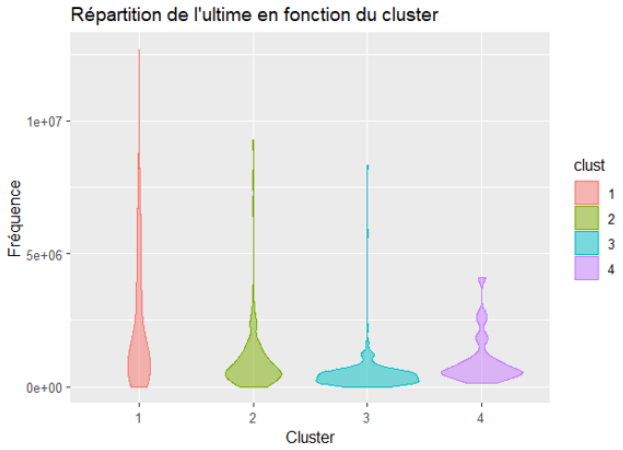


FIGURE 6.16 – Répartitions de l'ultime en fonction des clusters CART

Cluster	Moyenne(Ultime)	Sd(Ultime)	CV(Ultime)
1	1 612 342	2 502 877	1.5
2	733 075	1 208 136	1.6
3	453 662	740 994	1.6
4	1 296 973	1 444 655	1.1

TABLE 6.3 – Statistiques sur l'ultime en fonction des clusters CART

Nous constatons que les clusters 1 et 4 rassemblent des sinistres qui ont en moyenne une sévérité plus élevée que dans le cluster 2 et 3. Cependant, nous avons une forte variabilité des ultimes dans les classes 1, 2 et 3 et une variabilité faible dans la classe 4.

La position des dossiers clos est différente selon les clusters : dans le cluster 1, 2 et 4 nous retrouvons majoritairement des dossiers qui restent en dessous du seuil 500k € jusqu'à la clôture, contrairement au cluster 3 où nous enregistrons plus de dossiers qui restent au-dessus du seuil jusqu'à la clôture.

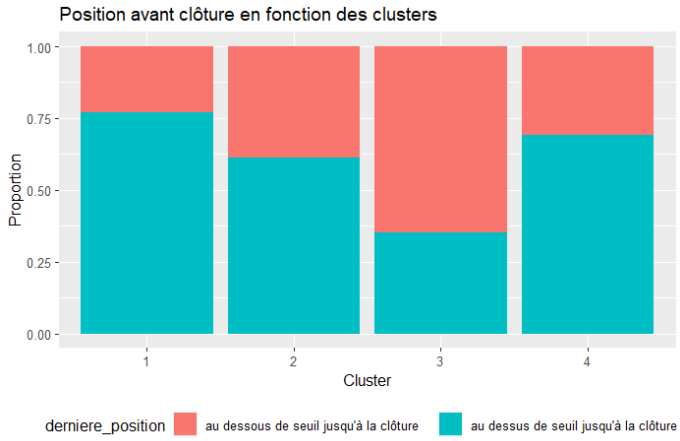


FIGURE 6.17 – Dernière position avant clôture en fonction des clusters - Base AIPP

La répartition des tranches d'AIPP est différente d'un cluster à un autre. En effet, dans le cluster 1 nous retrouvons les cinq différentes tranches d'AIPP avec une proportion majoritaire de sinistres graves associés à un taux d'AIPP compris entre 30% et 59% et ceux associés à un taux d'AIPP supérieur à 60% : cela rejoint bien ce à quoi l'on s'attendait intuitivement, puisque ce cluster a l'ultime moyen le plus élevé, donc contient les sinistres les plus graves en termes d'AIPP et d'ultime. Le cluster 2 quant à lui est composé seulement de sinistres graves en termes d'AIPP, mais un peu moins que le cluster 1 en termes d'ultime. Le cluster 4 quant à lui est assez particulier, puisqu'il ne compte aucun sinistre ayant un taux d'AIPP élevé : ils ont tous entre 0% et 29% d'AIPP ; cependant, ces sinistres sont particulièrement graves en termes d'ultime (les sinistres en majorité clôturés au dessus du seuil, et l'ultime moyen du cluster dépasse le million). Ainsi, le cluster 4 regroupe les sinistres pour lesquels on n'a pas d'atteinte à l'intégrité physique ou psychique mais qui ont nécessité une indemnisation importante (exemple : chirurgie lourde et rééducation qui ne laisse ensuite pas de séquelles). Enfin, le cluster 3 contient les sinistres les moins graves au sens de l'ultime et au sens du taux d'AIPP.

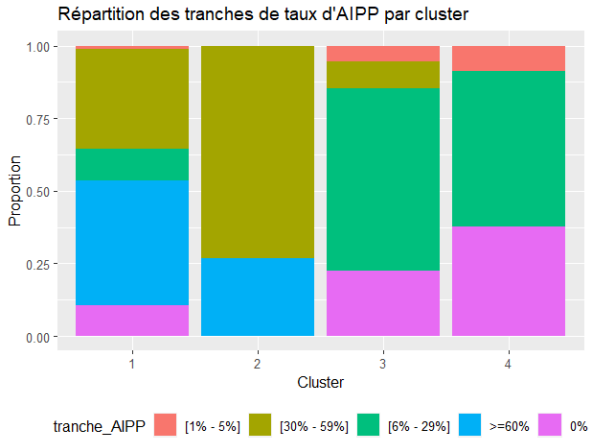


FIGURE 6.18 – Repartition des tranches de taux d'AIPP en fonction des clusters - Base AIPP

6.2.2 Résultats du modèle à états

Dans cette partie nous calibrons le modèle à états, présenté dans la section 3.1 du chapitre 3, sur la base de donnée restreinte avec l'information taux d'AIPP.

Calcul du montant de la provision avec la formule fermée

Nous observons 496 sinistres clos, avec une distribution des délais comme dans la figure suivante :

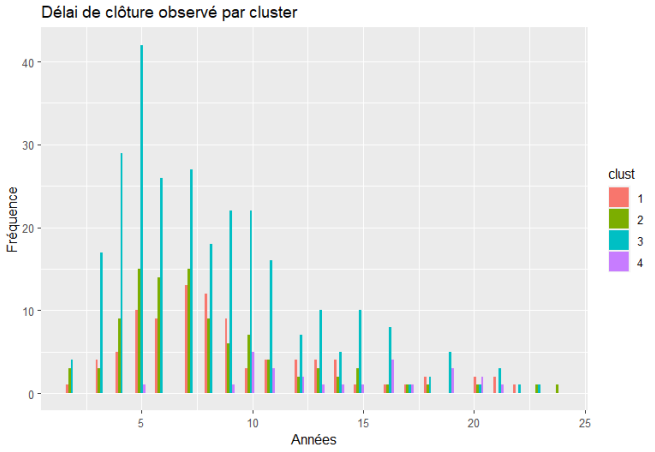


FIGURE 6.19 – Délais de clôture observés - Base AIPP

Nous observons aussi 888 RBNS, avec une distribution de l'exposition comme le montre la figure ci-dessous :

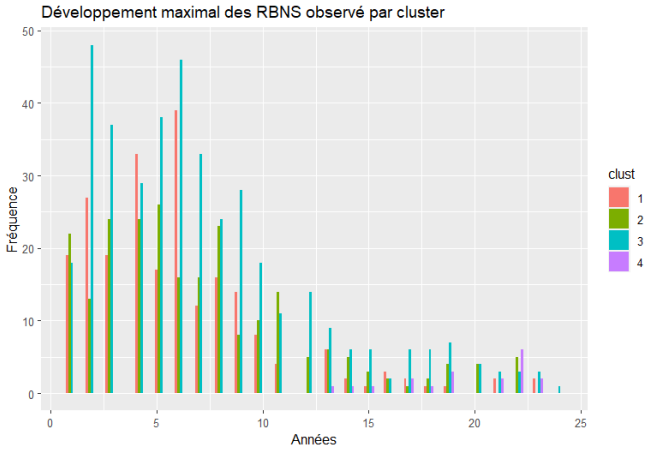


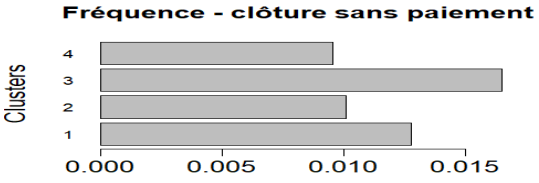
FIGURE 6.20 – Développement maximal des RBNS observé - Base AIPP

Calibrage des fréquences de paiement et de clôture

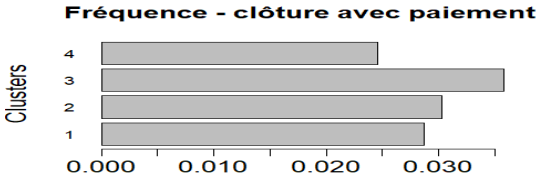
Afin de calculer le montant de la provision nous avons besoin d'estimer les fréquences associées aux événements de paiement et de clôture.
Nous rappelons que, dans un premier temps, nous calibrons les fréquences associées aux évènements de paiement et de clôture. Les évènements en question sont :

- 1 : Clôture sans paiement
- 2 : Paiement avec clôture
- 3 : Paiement sans clôture

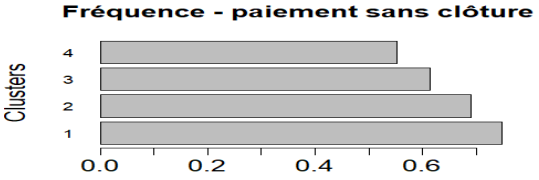
Les fréquences calibrées dans le cas où les fréquences associées aux évènements de clôture et de paiement sont constantes sont présentées dans la figure suivante :



(a) Fréquence de clôture sans paiement (h_1)



(b) Fréquence de clôture avec paiement (h_2)



(c) Fréquence de paiement sans clôture (h_3)

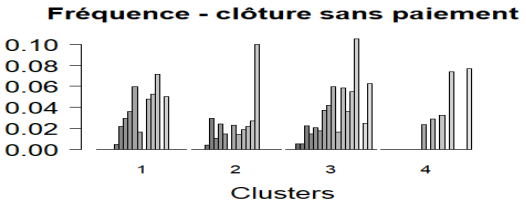
FIGURE 6.21 – Fréquences des évènements calibrées par cluster - Base AIPP

Nous constatons que les sinistres des clusters 1 et 3 comportent plus de paiements et présentent des clôtures plus fréquentes, contrairement aux sinistres du cluster 4 qui

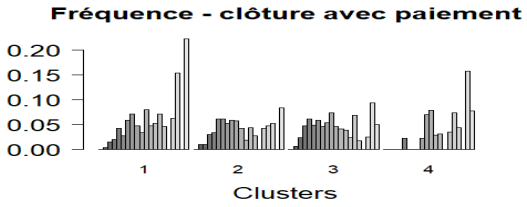
présentent des clôtures moins fréquentes, combinées à des paiements moins fréquents. Les sinistres du cluster 2 se caractérisent par des paiements fréquents et des clôtures moins fréquentes.

Dans le cas où les fréquences des évènements sont constantes par morceaux, nous obtenons les résultats suivants :

- La fréquence de clôture sans paiement h_1, sont maximales pour la majorité des clusters pour les sinistres d'âge autour de 12-14 ans ; ces estimations permettent de caractériser les sinistres à fort potentiel de développement long : ceux-ci sont plutôt d'âge 3-4 ans et appartiennent aux clusters 1 à 3 (fréquences faibles).
- La fréquence de clôture avec paiement sont maximales pour les sinistres d'âge 19 - 20 ans et sont très faibles (voir nulle) pour les sinistres d'âge 1-3 ans.
- Les sinistres d'âge autour de 2-3 ans présentent des paiements à venir potentiellement plus fréquents que les sinistres développés depuis 10 à 12 ans.



(a) Fréquence de clôture sans paiement (h_1)



(b) Fréquence de clôture avec paiement (h_2)

(c) Fréquence de paiement sans clôture (h_3)

FIGURE 6.22 – Fréquences des évènements calibrées par cluster dans le cas où les fréquences sont dépendantes du temps écoulé depuis la déclaration des sinistres - Base AIPP

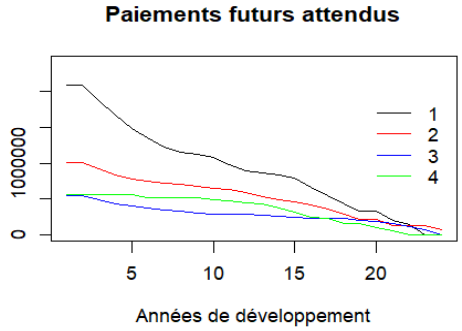
Calibrage des paiements

Ici également nous testons deux lois : la loi Log-Normale et la loi mélange Log-Normale Exponentielle. La loi Log-Normale fitte mieux les paiements observés (voir C.9 en annexe). Ainsi, une loi Log-Normale shiftée de 1.01 * la valeur minimale des recours est calibrée sur les données.

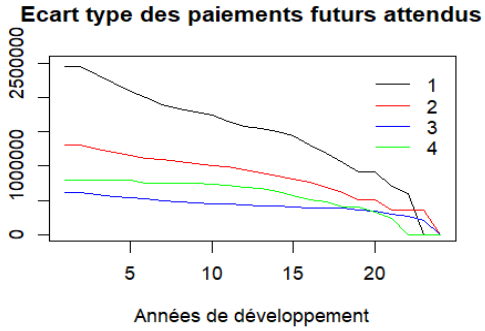
Les résultats obtenus par formule fermée sont les suivants :

- En fonction de la durée de développement, les paiements futurs espérés restent stables sur les deux premières années, dû à la croissance de la fréquence de paiement et de la fréquence de clôture.
- Les paiements espérés baissent ensuite, dû à la combinaison d'une décroissance de la fréquence de paiement et d'une croissance de fréquence de clôture.

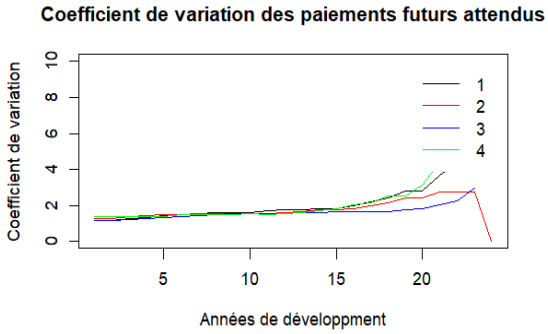
- Les paiements futurs espérés, pour les sinistres du cluster 2 et 3, sont quasiment pareils à partir de la 18ème année de développement.
- En se basant sur les coefficients de variation, les clusters se caractérisent par un niveau d'incertitude légèrement élevé sur les dernières années de développement (à partir de la 18-ème année), concernant le cluster 2, nous observons une baisse de cette incertitude sur la dernière année de développement.



(a) Paiements futurs attendus



(b) Ecart type des paiement futurs



(c) Coefficient de variation des paiements futurs

FIGURE 6.23 – Espérance, écart type et coefficient de variation des paiements futurs en fonction des clusters - Base AIPP

Le montant de la provision, au titre des RBNS vus à fin 2019, que nous obtenons avec cette méthode s'élève à 2.21 Mds €auquel nous ajoutons le montant des paiements à date pour obtenir une charge ultime de 2.77 Mds €.

Les résultats de la simulation sont présentés dans la figure suivante :

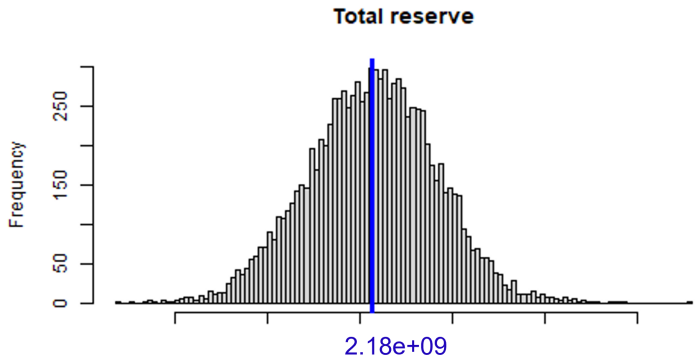


FIGURE 6.24 – Distribution de la provision totale - Base AIPP

Le même calibrage est appliqué à la base de données restreinte excluant la variable taux d'AIPP. Le montant de la provision, au titre des RBNS vus à fin 2019 obtenu avec ce calibrage s'élève à 2.28 Mds € donc une charge ultime de 2.84 Mds €, et une provision moyenne obtenue par simulation estimée à 2.24 Mds € à laquelle est ajoutée les paiements à date pour obtenir une charge ultime qui s'élève à 2.80 Mds €.

Afin de pouvoir comparer les deux résultats, une analyse des erreurs d'estimation et de process est nécessaire. Les résultats obtenus sont les suivants :

	RMSE	Erreur de process	Erreur d'estimation
Y compris le taux d'AIPP	494 206 144	85 371 712	408 834 433
Sans taux d'AIPP	633 113 859	88 954 958	544 158 902

TABLE 6.4 – Erreurs de prédiction du modèle à états calibré sur la base restreinte

À partir de ces résultats nous pouvons voir qu'en termes d'erreur le modèle calibré sur la base restreinte en tenant compte de la variable taux d'AIPP est légèrement meilleur. En effet, la prise en compte de variables ayant un lien direct avec la sévérité a un impact positif sur la performance du modèle à états.

La charge ultime obtenue sur la base restreinte est supérieure à celle obtenue sur la base totale, ceci peut s'explique par la sensibilité du modèle à états à la loi des paiements et aux fréquences calibrées.

6.2.3 Résultats du modèle B - global Random Forest

Afin de savoir comment un modèle non paramétrique réagit à la prise en compte de la variable taux d'AIPP, nous calibrons le modèle global Random Forest entraîné sur les sinistres clos et RBNS complétés par les facteurs de Mack Chain-Ladder sur la base de données restreinte tenant compte du taux d'AIPP et sur la même base excluant le taux d'AIPP.

Comparaison des ultimes réels et des pseudo ultimes pour les années de backtesting

Dans un premier temps, nous nous intéressons aux écarts entre les ultimes réels et les « pseudo-ultimes » calculés par Mack Chain-Ladder.

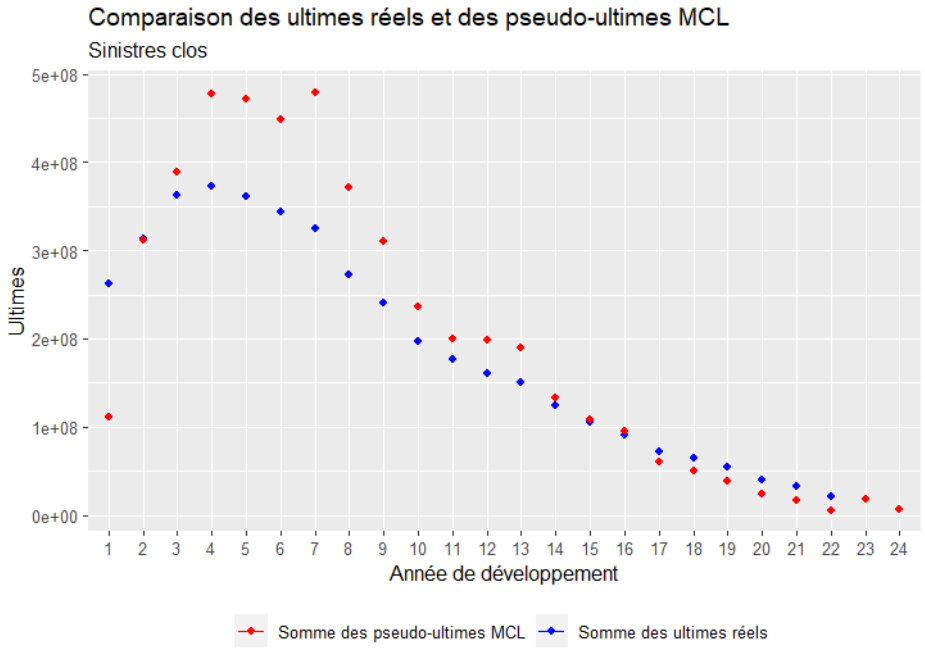


FIGURE 6.25 – Comparaison des ultimes réels et pseudo ultimes par années de développement - Base AIPP

Nous constatons que, par année de développement, les pseudos ultimes calculés surestiment l'ultime réel sur les années de développement de 3 à 14 et le sous-estiment sur le reste des années.

Nous comparons aussi, pour chaque année de backtesting, la valeur des « pseudo-ultimes » pour les RBNS qui sont ensuite clos avant le 31/12/2019, avec leur valeur réelle.

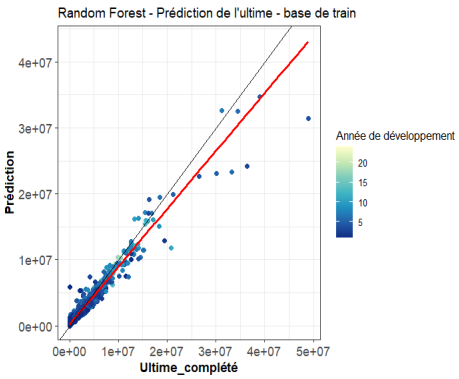
Année N	Ultime réel	Pseudo-ultime	Erreur
2010	298 075 870	116 562 894	-61%
2011	381 348 252	155 952 779	-59%
2012	432 400 028	217 403 377	-50%
2013	461 989 849	269 572 770	-42%
2014	459 410 081	314 156 917	-32%
2015	443 668 728	378 312 532	-15%
2016	354 572 819	429 468 230	21%
2017	239 387 969	273 626 720	14%
2018	121 777 211	205 396 690	69%

TABLE 6.5 – Données sur les RBNS vus à fin N, clos entre le 31/12/N et le 31/12/2019

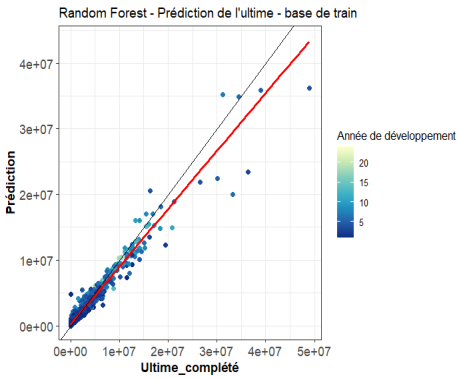
Selon les années de backtesting les « pseudo-ultimes » sous-estiment fortement l'ultime

réel sur les trois premières années, mais ces écarts diminuent par la suite et augmentent sur les trois dernières années avec une forte surestimation sur l'année 2018.

Validation du modèle avec les bases test



(a) Prédiction de l'ultime sur la base train (yc taux d'AIPP)

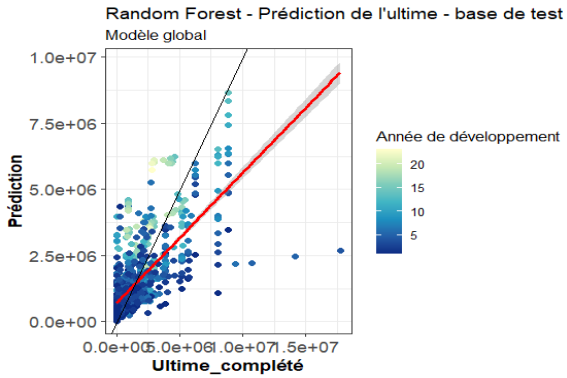


(b) Prédiction de l'ultime sur la base train (sans le taux d'AIPP)

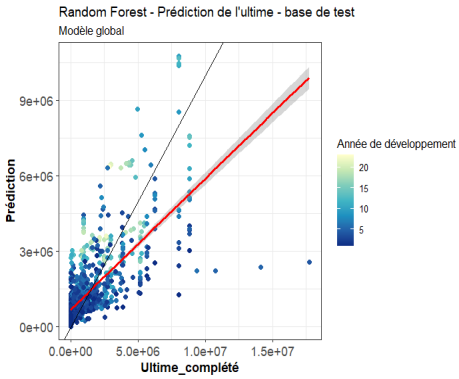
FIGURE 6.26 – Prédictions de l'ultime par le modèle global entraîné sur les clos et RBNS-Random Forest sur les bases d'entraînement

%Variance expliquée (yc AIPP)	%Variance expliquée (sans AIPP)	RMSE base d'apprentissage (yc AIPP)	RMSE base d'apprentissage (sans AIPP)
85.13%	83.88%	444 057	462 204

TABLE 6.6 – Performances du modèle global entraîné sur les clos et RBNS-Random Forest (base train)



(a) Prédiction de l'ultime sur la base test (yc taux d'AIPP)

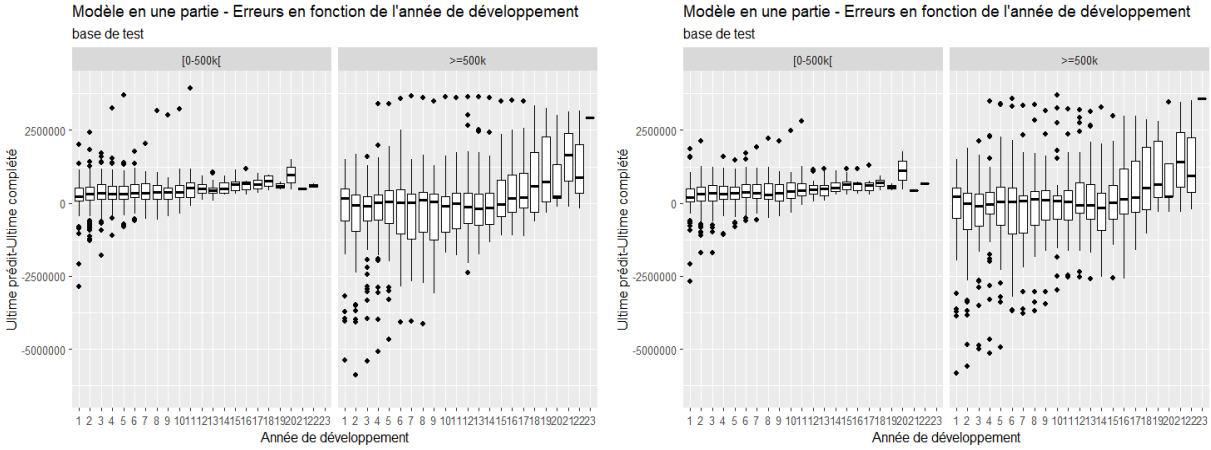


(b) Prédiction de l'ultime sur la base test (sans le taux d'AIPP)

FIGURE 6.27 – Prédictions de l'ultime par le modèle global entraîné sur les clos et RBNS-Random Forest sur les bases de test

RMSE base de test (yc AIPP)	RMSE base de test (sans AIPP)	Erreur globale (yc AIPP)	Erreur globale (sans AIPP)
1 161 320	1 190 490	-11 169 440	-28 143 111

TABLE 6.7 – Performances du modèle global entraîné sur les clos et RBNS- Random Forest (base test)



(a) Avec le taux d'AIPP

(b) Sans le taux d'AIPP

FIGURE 6.28 – Erreurs de prédiction du modèle global entraîné sur les clos et RBNS sur les deux bases test en fonction des années de développement-Random Forest

Le modèle surapprend la base d'apprentissage incluant le taux d'AIPP et celle qui l'exclut, en revanche nous observons un pourcentage de variance expliquée légèrement plus élevé en présence de la variable taux d'AIPP que ce que nous observons en son absence. Également en termes de RMSE et d'erreur globale le modèle est performant en présence de la variable taux d'AIPP qu'en son absence.

Comparaison des prédictions des modèles sur les sinistres clos – exercice de backtesting

	Erreurs de prédiction entre l'ultime réel et les différentes prédictions		
Année N	Gestionnaire	Avec le taux d'AIPP	Sans le taux d'AIPP
2010	22%	-12%	-10%
2011	30%	-1%	1%
2012	42%	-8%	4%
2013	38%	1%	3%
2014	37%	8%	2%
2015	28%	19%	20%
2016	32%	79%	73%
2017	26%	59%	51%
2018	22%	103%	92%

TABLE 6.8 – Erreurs de prédiction entre l'ultime réel et les prédictions

Les résultats de l'exercice de backtesting ne sont pas tout à fait pertinents pour valider la robustesse des modèles et la taille de la base de données y est pour beaucoup.

Comparaison des ultimes des RBNS estimés au 31/12/2019

	Prédictions du modèle sur les deux bases	
Année N	Avec le taux d'AIPP	sans le taux d'AIPP
2019	2 259 073 252	2 257 515 618

TABLE 6.9 – Prédiction de l'ultime par le modèle sur les deux bases

Les prédictions du modèle en présence de la variable taux d'AIPP et en son absence sont équivalentes.

	Moyenne	Coefficient de variation
Base restreinte y compris taux d'AIPP	2.26 Mds €	5.2%
Base restreinte sans taux d'AIPP	2.39 Mds €	5.0%

TABLE 6.10 – Performances des modèles

Ces résultats donnent une idée sur le comportement du modèle non paramétrique en présence de la variable taux d'AIPP et en son absence. Pour avoir des valeurs plus fiables en termes de robustesse de modèle et de variabilité, il nous faudrait une base de données plus grande.

À ce stade nous pouvons constater que la présence de la variable taux d'AIPP n'améliore pas la performance du modèle, du fait de la petite taille de la base de données et la plus grande partie de l'erreur des modèles implémentés est portée par les pseudo-ultimes et non pas par l'erreur de prédiction du modèle en soi. Ainsi, nous espérons obtenir des résultats plus pertinents sur une base de données volumineuse.

Conclusion

En assurance non-vie, l'estimation du montant des provisions est un enjeu important. Suite à la révolution numérique que connaît le secteur de l'assurance, un tournant s'effectue dans la recherche actuarielle sur l'exploitation de toutes les données disponibles en provisionnement. Différents travaux sur des méthodes ligne à ligne ont été menés : dans le cadre de modélisation de ces modèles, chaque sinistre est considéré individuellement et les caractéristiques du sinistre sont prises en compte pour prédire l'ultime.

L'objectif de ce mémoire était de donner un aperçu global des méthodes de provisionnement ligne à ligne existantes puis d'en sélectionner certaines et de les challenger sur l'estimation de la provision des RBNS d'un portefeuille de sinistres graves de la branche RCC Automobile afin d'en évaluer les avantages et les inconvénients. Ensuite, nous souhaitions les comparer avec les méthodes classiques Chain-Ladder et Mack. Enfin, nous voulions voir dans quelle mesure les modèles individuels implémentés pouvaient être modulables pour s'adapter au mieux aux spécificités des sinistres de la branche étudiée.

En termes de résultats, les prédictions obtenues au titre des RBNS observés au 31/12/2019 avec les différents modèles individuels retenus, et avec les méthodes agrégées, sont assez similaires.
Nous disposons d'une valeur "de référence" qui nous semble être un ultime juste et prudent, de 2.0 Mds € - 2.1 Mds € : il s'agit de la provision constituée par les gestionnaires sinistres (de 2.7 Mds €), retraitée de 20%, l'exercice de backtesting nous ayant permis de savoir qu'ils surestiment l'ultime d'au moins 20%. Les modèles agrégés classiques s'alignent sur cette valeur, puisque nous obtenons une estimation de l'ultime de 2.1 Mds €.
Le modèle à états quant à lui donne une estimation de 2.3 Mds €, légèrement supérieure à la valeur de référence. Enfin, l'étude des modèles de machine learning entrainés sur les sinistres clos uniquement nous a permis de confirmer les hypothèses théoriques selon lesquelles le biais d'observation des sinistres clos uniquement conduirait à une sous-estimation de l'ultime. Les modèles de machine learning entrainés sur les sinistres clos et RBNS complété donnent une estimation plus juste, qui s'aligne avec les valeurs de référence et les méthodes agrégées : en particulier, nous retenons le modèle basé sur l'algorithme de Random Forest, qui donne un ultime de 2.0 Mds €.

Si les modèles individuels, plus complexes que les méthodes agrégées, donnent les mêmes résultats que celles-ci, nous sommes alors en droit de nous demander : quel est l'apport du micro-provisionnement par rapport au provisionnement classique ?

Tout d'abord, les modèles individuels de provisionnement ajoutent de la compréhension sur les sinistres. Ainsi le clustering supervisé en amont du calibrage du modèle à états nous a permis de regrouper les dossiers par groupes de caractéristiques homogènes et de distinguer les sinistres en fonction de leur sévérité : ainsi, nous avons pu observer

des fréquences associées aux évènement de paiement et de clôture différentes selon les groupes. Cette segmentation nous a permis de calculer, via les formules fermées, une espérance et une variance des paiements futurs par groupe et par année de développement. En revanche, l'implémentation modèle à états nous a permis de constater qu'il est sensible à la loi des paiements calibrées ainsi qu'à la composition de la base en termes de nombre de dossiers clos et de RBNS.

Les modèles de machine learning, quant à eux, permettent d'identifier les principales variables disponibles influant sur le montant de charges (dans notre étude nous avons préféré conserver toutes les variables). Aussi, grâce à ces modèles nous avons pu distinguer les erreurs de prédictions sur la base de test selon la classe d'ultime (ultime inférieur à 500k € et ultime supérieur à 500k €) : ce cadre de modélisation donne plus de possibilités pour corriger les erreurs du modèle, car nous sommes capable de repérer où le modèle se trompe.

Enfin, nous avons pu apprécier le caractère modulable des modèles individuels, qui permettent de prendre en compte des spécificités des sinistres : ici, il s'agissait du dépassement du seuil de 500k € de charge, mais le modèle à états peut également prendre en compte d'autres types d'événements, comme les réouvertures, par exemple. L'intégration de cette particularité des sinistres graves nous a permis de remettre en question le principe "grave un jour, grave toujours" et d'affiner les prédictions en distinguant les paiements associés aux sinistres qui s'avèrent être non graves à la clôture.

L'implémentation des modèles individuels de type machine learning nous a également permis de souligner deux "faiblesses" globales de ce type de modèles.

- Les covariables prédictives, le nerf de la guerre : pour être performant, un modèle de machine learning doit contenir des variables prédictives de la variable cible. Dans le cas de notre étude, il nous fallait donc des variables en lien direct avec la sévérité des sinistres, comme le taux d'AIPP. Or, ce type de variables ne sont pas forcément aisées à récupérer.(Cela n'a pas pu être vérifié directement sur tous les modèles implémentés, par manque de données disponnibles.)
- Une base de données conséquente en termes d'observations : il est difficile d'apprendre des patterns sur des données restreintes. Dans notre étude, les résultats des modèles de machine learning peuvent être mitigés, notamment sur l'apport de la variable "taux d'AIPP", mais cela est surement uniquement dû au volume très faible d'observations dont nous disposons.
- Par ailleurs, nous avons pu remarquer également un inconvénient du modèle individuel à états : la difficulté à trouver une loi qui s'adapte à la distribution souvent complexe des paiements, ainsi que la grande sensibilité du modèle à la loi calibrée. Ce point peut altérer les résultats obtenus.

Pour conclure, malgré leurs limites, les modèles individuels donnent plus de sens à la provision estimée, par rapport aux méthodes agrégées. Par ailleurs, une étude approfondie - non menée dans le cadre de ce mémoire - de la provision constituée avec les gestionnaires sinistres, en fonction des covariables, est en soi déjà très utile pour les assureurs, qui pourraient alors repérer les variables qui expliqueraient les écarts entre les provisions d/d ainsi que les révisions de provisions, et l'ultime, à chaque année de développement des sinistres. De cette façon, un meilleur pilotage, plus objectif, des règles d'attribution des provisions pourrait être obtenu.

La modélisation individuelle en provisionnement non-vie est un domaine dans lequel beaucoup de travaux restent à faire, notamment : une étude supplémentaire sur les IBNyR afin de pouvoir réaliser une comparaison complète entre les résultats des modèles ligne à ligne et ceux de la méthode classique Mack ; l'application des modèles sur une base de données volumineuse incluant des variables en lien direct avec la sévérité des sinistres ; une étude approfondie de la variabilité des modèles, et surtout de la comparaison des différents types de modèles entre eux en termes d'erreur globale, d'estimation et de process.

Bibliographie

[1] G. Andrea. “An individual claims reserving model for reported claims .” In : European Actuarial Journal (2021).

[2] K. Antonio, E. Godecharle et R. V. Oirbeek. “A Multi-State Approach and Flexible Payment Distributions for Micro-Level Reserving in General Insurance.” In : (2016).

[3] K. Antonio et R. Plat. “Micro-level stochastic loss reserving for general insurance”. In : Scandinavian Actuarial Journal 2014 (2014).

[4] E. Arjas. “The claims reserving problem in non-life insurance : Some structuralideas”. In : Astin Bulletin 19(2) (1989).

[5] M. Ayuso et M. Santolino. “Prediction of individual automobile RBNS claim reserves in the context of Solvency II.” In : (2008).

[6] A. L. Badescu, L. X. Sheldon et T. Dameng. “A marked Cox model for the number of IBNR claims : estimation and application”. In : (2016).

[7] A. L. Badescu, L. X. Sheldon et T. Dameng. “A marked Cox model for the number of IBNR claims : Theory”. In : Insurance : Mathematics and Economics 69 (2016).

[8] M. BAUDRY et C. Y. ROBERT. “Non parametric individual claim reserving in insurance.” In : (2019).

[9] A. Boumezoued et L. Devineau. “Individual claims reserving : a survey”. In : (2017).

[10] I. Chaoubi et al. “Micro-level Reserving for General Insurance Claims using a Long Short-Term Memory Network.” In : (2022).

[11] A. Charpentier et M. Pigeon. “Macro vs. micro methods in non-life claims reserving (an econometric perspective).” In : (2016).

[12] T. Chen et C. Guestrin. “XGBoost : A Scalable Tree Boosting System”. In : (2016).

[13] L. Delong, M. Lindholm et M. V. Wüthrich. “Collective reserving using individual claims data.” In : Scandinavian Actuarial Journal (2021).

[14] L. Delong et M. V. Wüthrich. “Neural networks for the joint development of individual payments and claim incurred Risks.” In : (2020).

[15] C. Dutang. “Cours ENSAE - Actuariat de l’sssurance Non-Vie.” In : (2022).

[16] F. Duval et M. Pigeon. “Individual Loss Reserving Using a Gradient Boosting-Based Approach.” In : (2019).

[17] A. Gabrielli, R. Richman et M. V. Wüthrich. "Neural network embedding of the over-dispersed Poisson reserving model." In : Scandinavian Actuarial Journal (2020).

[18] F. Guiahi. "A probabilistic model for IBNR claims". In : CAS Proceedings (1986).

[19] S. Haastrup et E. Arjas. "Claims reserving in continuous time ; a nonparametric bayesian approach." In : ASTIN Bulletin (1996).

[20] C. Hachemeister. "A stochastic model for loss reserving." In : (1980).

[21] S. Happ et M. V. Wüthrich. "Paid–incurred chain reserving method with dependence modelling." In : ASTIN Bulletin (2013).

[22] O. Hesselager. "A Markov model for loss reserving". In : Astin Bulletin 24(2) (1994).

[23] S. Hochreiter et J. Schmidhuber. "Long Short-Term Memory Neural computation." In : (1997).

[24] J. J. Hopfield. "Neural networks and physical systems with emergent collective computational abilities Proceedings of the national academy of sciences." In : (1982).

[25] W. S. Jewell. "Predicting IBNYR events and delays : I. continuous time". In : Astin Bulletin 19(2) (1989).

[26] K. Kuo. "Individual claims forecasting with Bayesian mixture density networks ." In : arXiv preprint (2020).

[27] C. R. Larsen. "An individual claims reserving model". In : Astin Bulletin 37(1) (2007).

[28] *Le taux AIPP, l'assurance auto*. URL : `https://www.ornikar.com/assurance-auto/sinistre/assurance-accident/taux-aipp`. Site consulté le 17/10/2022.

[29] O. Lopez. "A censored copula model for micro-level claim reserving." In : (2018).

[30] O. Lopez, X. Milhaud et P.-E. Thérond. "Tree-based censored regression with applications in insurance." In : (2016).

[31] T. Mack. "Distribution-free calculation of the standard error of chain ladder reserve estimates". In : Astin Bulletin 23(2) (1993).

[32] M. Merz et M. V. Wüthrich. "Paid–incurred chain claims reserving method." In : Insurance : Mathematics and Economics (2010).

[33] R. Norberg. "Prediction of outstanding liabilities II. model variations and extensions". In : Astin Bulletin 29(01) (1999).

[34] R. Norberg. "Prediction of outstanding liabilities in non-life insurance". In : Astin Bulletin 23(1) (1993).

[35] M. Pigeon. "Individual Models for Loss Reserving and Reinsurance." In : (2014).

[36] *Seuil d'AIPP*. URL : `https://www.index-assurance.fr/dictionnaire/seuil-aipp/`. Site consulté le 17/10/2022.

[37] J. Sill et al. "Feature-Weighted Linear Stacking." In : (2009).

[38] G. Taylor. "Loss reserving models : Granular and machine learning forms Risks." In : (2019).

[39] R. J. Verrall et M. V. Wüthrich. "Understanding reporting delay in general insurance". In : Risks 4 (2016).

[40] M. V. Wüthrich. "Machine learning in individual claims reserving." In : Scandinavian Actuarial Journal (2018).

[41] X. Zhao et X. Zhou. "Applying copula models to individual claim loss reserving methods". In : Insurance : Mathematics and Economics 46 (1986).

Annexes

Annexe A

Zoom sur les modèles ligne à ligne implémentés

La fonction de vraisemblance

Pour chaque sinistre observé n et survenu à la date T_n, nous notons :

- $V_k^{(n)}$: le temps depuis la déclaration du k-ème événement dans le développement des paiements,
- $S^{(n)}$: délai de clôture pour un sinistre n,
- $\tau^{(n)}$: le temps pendant lequel le développemnt des sinistres est observé, qui s'écrit comme suit : $\tau^{(n)} = \min(S^{(n)}, \tau - T_n - U_n)$,
- $E_k^{(n)}$: type d'évenement associé,
- $\sigma_k^{(n)}(i)$: indique que le k-ème évenement associé au sinistre n est de type i $\in \{1, 2, 3\}$,
- $X_k^{(n)}$: le paiement associé au k-ème évenement,
- $P(.)$: fonction de densité associée à la distribution des paiements.

La fonction de vraisemblance associée au processus de développement des paiements de sinistre s'écrit :

$$\begin{aligned}\prod_{n\geq 1} \exp\left(-\int_0^{\tau^{(n)}} (h_1+h_2+h_3)(u)\mathrm{d}u\right) \prod_{k\geq 1} h_1(V_k^{(n)})^{\sigma_k^n(1)} h_2(V_k^{(n)})^{\sigma_k^n(2)} h_3(V_k^{(n)})^{\sigma_k^n(3)} \\ \times \prod_{n\geq 1}\prod_{k\geq 1}\left\{\sigma_k^{(n)}(1) + P(X_k^{(n)})\left(\sigma_k^{(n)}(2)+\sigma_k^{(n)}(3)\right)\right\}\end{aligned} \tag{A.1}$$

Proposition 3. *Pour tout $j \in \mathbb{N}^*$, l'éspérance et la variance des paiements relatif à un sinistre dans l'état j après un certain temps u peuvent s'écrire comme suit :*

$$\mathbb{E}[X(u,\infty)|S(u)=j] = \sum_{m\in\mathbb{N}^*}\sum_{n\in\mathbb{N}^*\setminus\{m\}} \int_u^\infty p_{jm}(u,v)\lambda_{mn}(v)y_{mn}(v)\mathrm{d}v \tag{A.2}$$

$$Var(X(u,\infty)|S(u)=j) = \sum_{m\in\mathbb{N}^*}\sum_{n\in\mathbb{N}^*\setminus\{m\}}\int_u^\infty p_{jm}(u,v)\lambda_{mn}(v)\left\{\sigma^2_{mn}(v)+r^2_{mn}(v)\right\}\mathrm{d}v \tag{A.3}$$

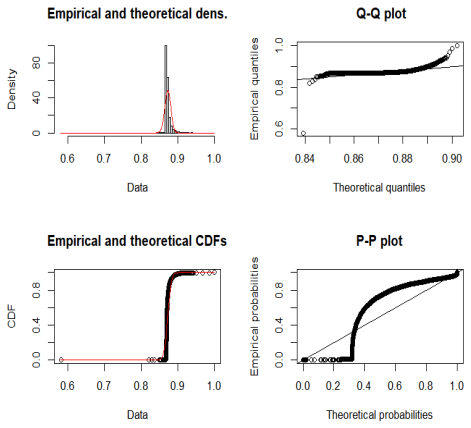
$$avec\ r^2_{mn}(v) = y^2_{mn}(v) + \mathbb{E}[X(v,\infty)|S(v)=n] - \mathbb{E}[X(v,\infty)|S(v)=m]$$

Annexe B

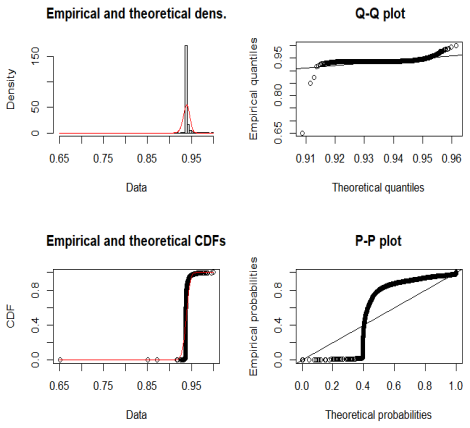
Implémentation et résultats

B.1 Calibrage du modèle à états

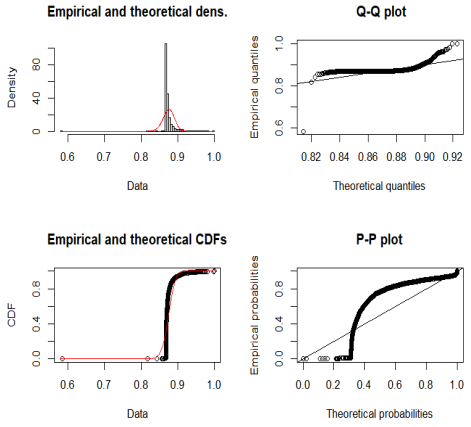
B.1.1 Calibrage des lois de paiement



(a) Cluster 1

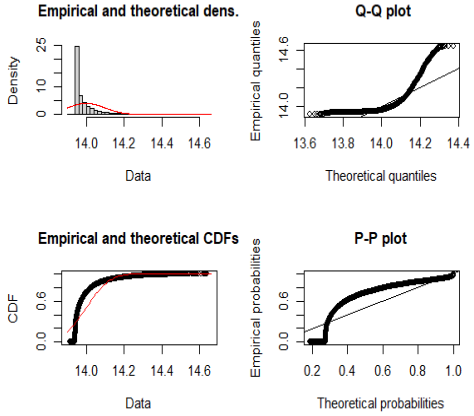


(b) Cluster 2

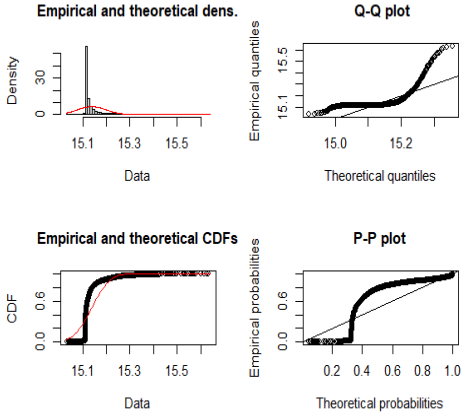


(c) Cluster 3

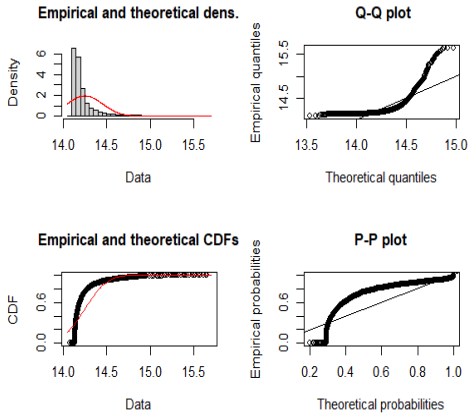
FIGURE B.1 – Loi Log-Bêta calibrée sur les paiements observés par cluster



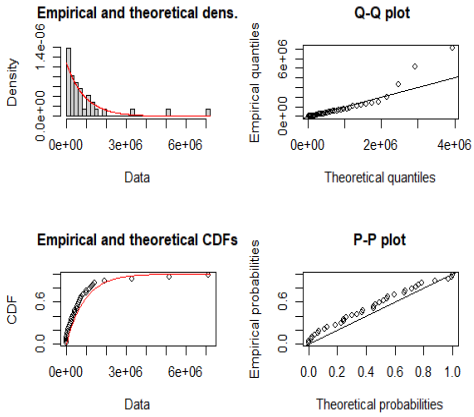
(a) Log-Normale (Cluster 1)



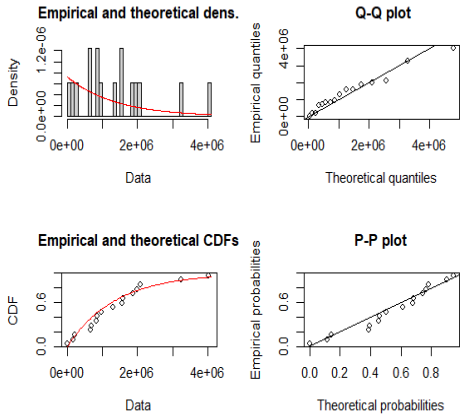
(b) Log-Normale (Cluster 2)



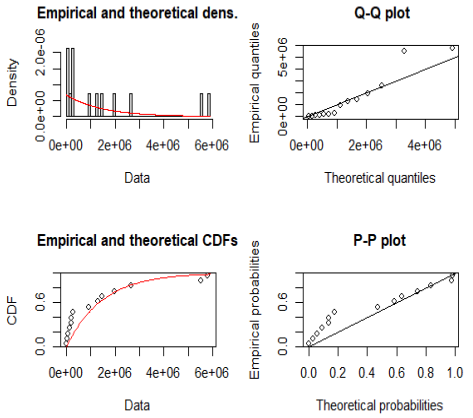
(c) Log-Normale (Cluster 3)



(d) Exponentielle (Cluster 1)



(e) Exponentielle (Cluster 2)



(f) Exponentielle (Cluster 3)

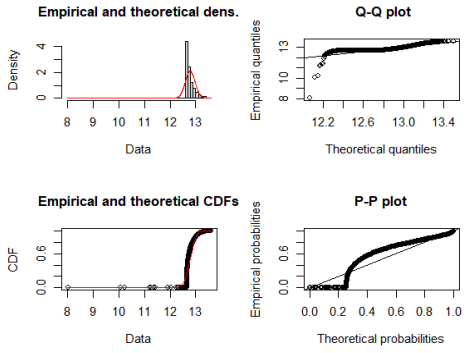
FIGURE B.2 – Loi mélange Log-Normale Exponentielle calibrée sur les paiements observés par cluster

Annexe C

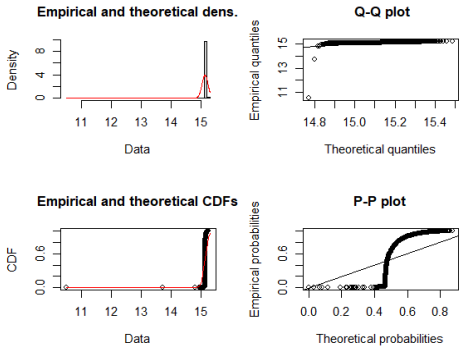
Extension des modèles de provisionnement ligne à ligne implémentés

C.0.1 Calibrage du modèle en observant la position du dernier paiement cumulé

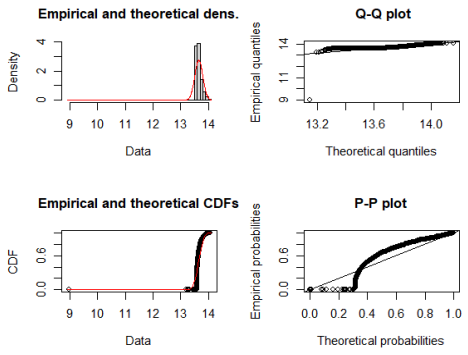
Calibrage des paiements



(a) Cluster 1

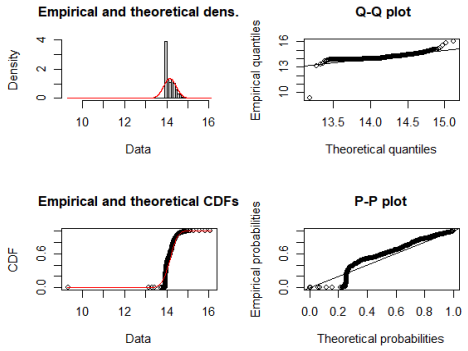


(b) Cluster 2

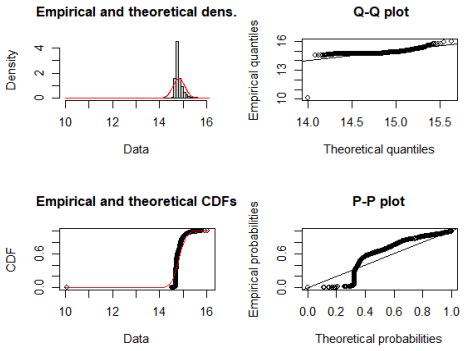


(c) Cluster 3

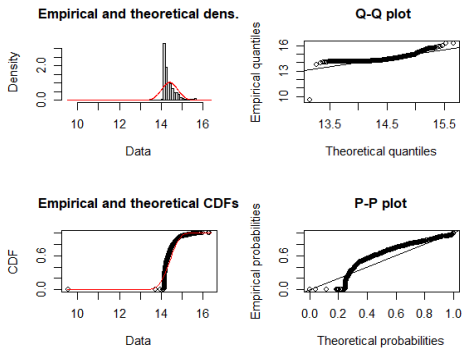
FIGURE C.1 – Loi Log-Normale calibrée sur les paiements observés par cluster (Dernier paiement cumulé associé est en dessous du seuil)



(a) Cluster 1

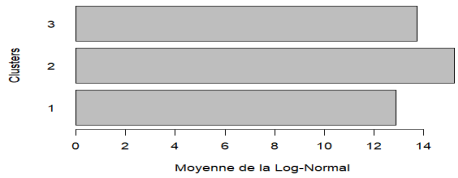


(b) Cluster 2

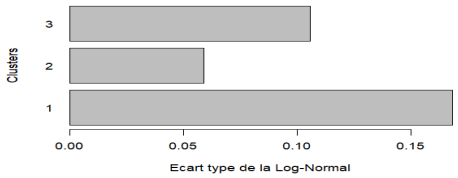


(c) Cluster 3

FIGURE C.2 – Loi Log-Normale calibrée sur les paiements observés par cluster (Dernier paiement cumulé associé est au-dessus du seuil)

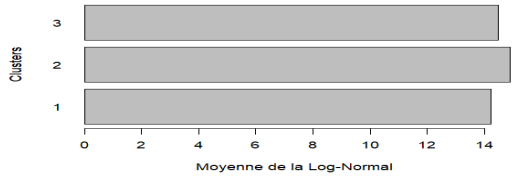


(a) Moyenne (μ)

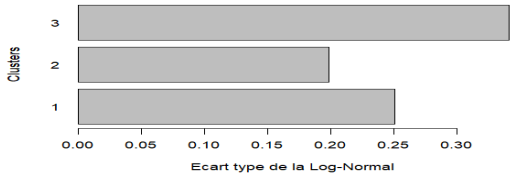


(b) Ecart type (σ)

FIGURE C.3 – Paramètres de la loi des paiements estimés pour lesquels le dernier paiement cumulé est en dessous du seuil



(a) Moyenne (μ)

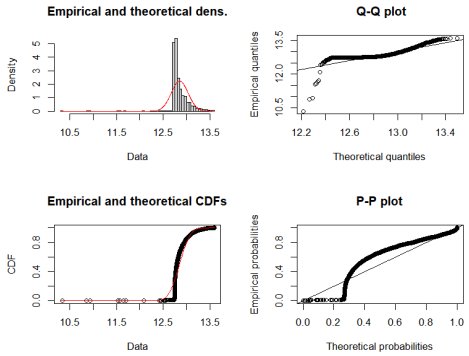


(b) Ecart type (σ)

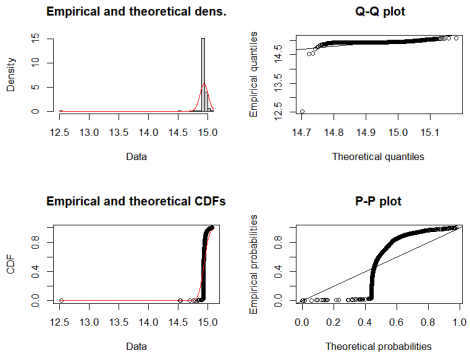
FIGURE C.4 – Paramètres de la loi des paiements estimés pour lesquels le dernier paiement cumulé est au-dessus du seuil

C.0.2 Calibrage du modèle en observant la position de la charge à chaque année de développement

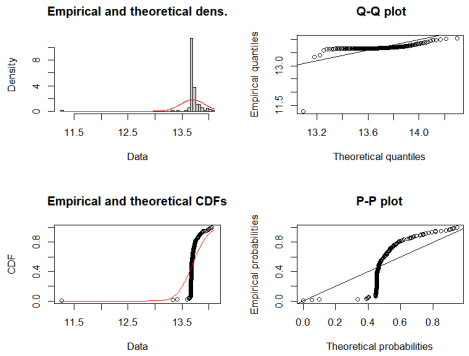
Calibrage des paiements



(a) Cluster 1

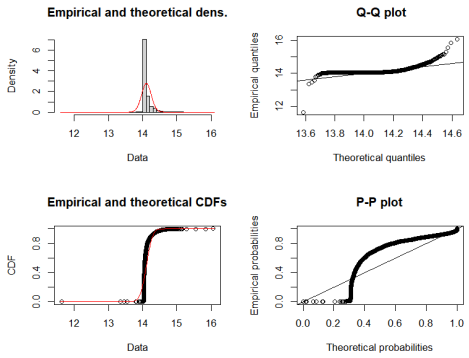


(b) Cluster 2

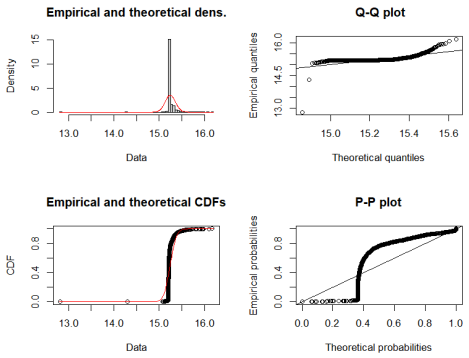


(c) Cluster 3

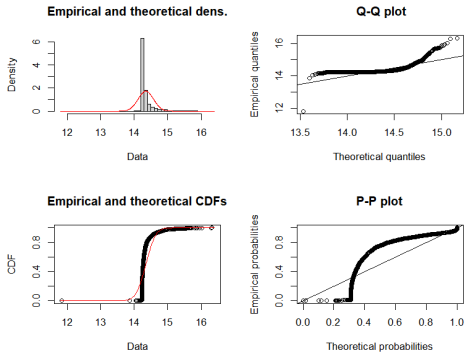
FIGURE C.5 – Loi Log-Normale calibrée sur les paiements observés par cluster (Charge associée est en dessous du seuil)



(a) Cluster 1



(b) Cluster 2



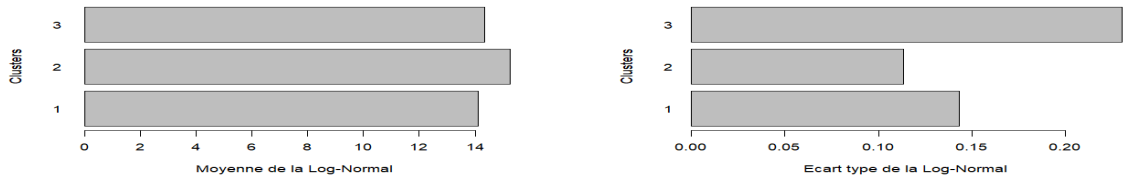
(c) Cluster 3

FIGURE C.6 – Loi Log-Normale calibrée sur les paiements observés par cluster (Charge associée est au-dessus du seuil)

(a) Moyenne (μ)

(b) Ecart type (σ)

FIGURE C.7 – Paramètres de la loi des paiements estimés pour lesquels la charge est en dessous du seuil



(a) Moyenne (μ)

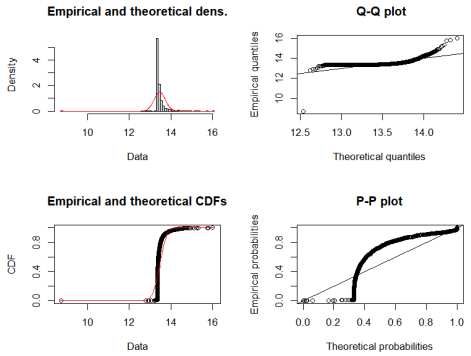
(b) Ecart type (σ)

FIGURE C.8 – Paramètres de la loi des paiements estimés pour lesquels la charge est au-dessus du seuil

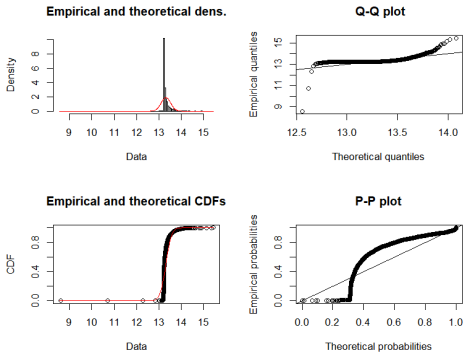
C.1 Etude de l'impact de l'ajout de la variable "Taux d'AIPP"

C.1.1 Résultats du modèle à états

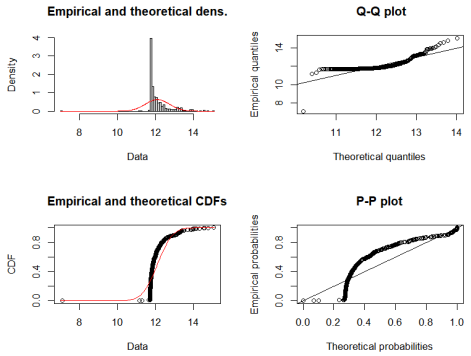
Calibrage des paiements



(a) Cluster 1

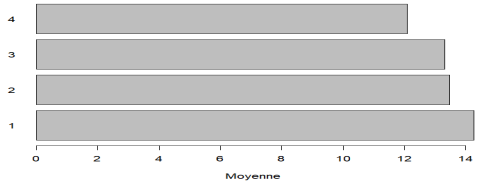


(b) Cluster 2

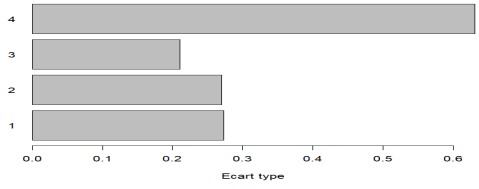


(c) Cluster 3

FIGURE C.9 – Loi Log-Normale calibrée sur les paiements observés par cluster - Base AIPP



(a) Moyenne (μ)



(b) Ecart type (σ)

FIGURE C.10 – Paramètres de la loi des paiements estimés - Base AIPP

Table des figures

Liste des tableaux